UNIVERSITÉ PARIS 7 - DENIS DIDEROT

UFR DE PHYSIQUE

THÈSE DE DOCTORAT

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PARIS 7

Spécialité : Interfaces Physique–Biologie

École doctorale : Constituants Élémentaires et Systèmes Complexes

Préparée dans l'URA 2171 – Génomes et Génétique
Unités Génétique in Silico et Génomique des Microorganismes Pathogènes,
à l'Institut Pasteur

Présentée par

Marc BAILLY-BECHET

Biais de codons et régulation de la traduction chez les bactéries et leurs phages.

Dirigée par Massimo VERGASSOLA

Soutenue le 29 juin 2007 devant le jury composé de :

Jean-Marc	Di Meglio	(président),
Michele	Caselle	(rapporteur),
Jean	Lobry	(rapporteur),
Antoine	Danchin,	
Eduardo	Rocha,	
Massimo	Vergassola.	

Remerciements

Ces mots sont classiques, mais cette thèse n'aurait pas pu se passer aussi bien sans que de nombreuses personnes n'y aient été impliquées, aussi bien professeurs que famille et amis. Dans une tentative utopique de n'oublier qu'un minimum de personnes, je tiens à remercier tout particulièrement :

En premier lieu Massimo, bien sûr, qui, très pédagogiquement, m'a appris le métier de chercheur, et a su faire s'opérer en moi la transition entre un étudiant désordonné et un investigateur raisonnablement méthodique. Sa remarquable présence, toujours là quand j'avais une question mais jamais trop pressant, a été un soutien constant et un exemple que j'aspire à suivre.

Ensuite les membres de mon jury, pour l'intérêt qu'ils ont porté à ma thèse, et pour toutes les suggestions qu'ils ont pu faire avant et pendant la soutenance. En particulier je pense aux deux rapporteurs, Michele Caselle et Jean Lobry, pour leur grande célérité dans l'écriture des rapports, et à Jean-Marc Di Meglio pour avoir accepté de présider ma soutenance.

Eduardo Rocha et Antoine Danchin, également membes du jury, ont un rôle à part : ils ont été à la fois juges, collaborateurs, et surtout exemples de ce qui se fait de mieux en biologie et en bioinformatique. Leurs nombreux conseils m'ont été d'un très grand secours durant ces trois années, et m'ont permis de voir comment rigueur et ouverture d'esprit pouvaient être complémentaires l'une de l'autre.

Parmi les personnes qui ont permis à cette thèse de se dérouler dans d'aussi bonnes conditions, je dois remercier en particulier Frank Kunst, pour avoir hébergé mon travail dans l'unité "Génomique des Microorganismes Pathogènes" pendant près d'un an et demi. Je n'oublierai pas non plus tous les membres de l'équipe, et toutes les discussions que nous avons eues : Carmen, Philippe, et bien sûr toute la "GMP Team", en particulier Christel, Jana, Matthieu, Christophe et Mathieu...

Je dois également dire merci à Yves Charon et l'équipe de l'ED 382, qui se sont battus pour que moi et mes camarades de DEA puissions tous obtenir une bourse en début de thèse et choisir un sujet qui nous passionne.

Une pensée spéciale va également à Michel Kerszberg, Florence Pradillon et Françoise Gaill. Sans eux jamais mon travail de DEA n'aurait pu se développer en un projet de plus grande ampleur, pour finalement conduire à une publication.

Merci à mon premier collaborateur sur un projet entièrement personnel, Guillaume. Ce fut un plaisir de travailler avec toi et toutes tes idées.

Une mention spéciale pour mes entraîneurs pendant ces trois ans, Gilles, Vincent, Manu, Freddy et Zank, qui ont eu, quoique indirectement, une grande influence sur mon travail.

Finalement, il reste encore beaucoup de gens que je ne saurais oublier, parce qu'ils ont été là, à un moment où à un autre. Merci à toute ma famille, mes parents et mes sœurs ; et merci à tous les copains : Nico, Cro, Claire L. et Claire G., Sophie, Erwan, Yiming, Kev, Seb S. et Seb K., Cyrille, Tristan, Aurélien, Henry, Laurent, Bruno, Tonio, Silvère... Et tous ceux auxquels je penserai après que cette page soit imprimée.

Et Héloïse, qui dira qu'elle n'y est pour rien, et à qui je répondrai qu'elle y est pour beaucoup.

Avant-propos

Le séquençage du premier génome a été terminé en 1995 (Fleischmann et al., 1995). Depuis, les données biologiques n'ont cessé de s'accumuler, et d'ici quelques années plus de 1000 génomes seront disponibles. Toutes ces données sont autant de promesses pour le futur, avec tout ce qu'elles peuvent nous apporter pour mieux comprendre les problèmes d'environnement, de santé, mais également l'histoire qu'elles relatent, celle des êtres vivants, et de son moteur, l'évolution.

Mais une telle accumulation de données ne résulte pas aisément en une augmentation parallèle des connaissances. En effet, le savoir acquis grâce à elles nous a surtout permis de comprendre que les principes biologiques que l'on croyait généraux ne représentent qu'une facette du monde vivant. Les nombreuses découvertes permises par l'avènement de l'ère génomique, sur la régulation génétique, le développement, l'organisation cellulaire et bien sûr l'évolution, ne font en effet que renforcer nos impressions sur la complexité des systèmes biologiques et leur diversité.

Les systèmes biologiques sont donc des systèmes complexes, comprenant des milliers de molécules aux interactions encore plus nombreuses. Chacune de ces molécules est formée de centaines d'acides aminés ou de bases nucléotidiques, ayant tous des propriétés différentes. Toute l'information – ou du moins une grande partie – nécessaire au bon fonctionnement de ce système est présente dans le génome. Un tel système se prête naturellement à l'application des méthodes de la physique statistique, même s'il n'en est pas l'objet d'intérêt traditionnel. La physique statistique a en effet développé des techniques pour l'étude des systèmes composés d'un grand nombre de particules, recouvrant une gamme d'interactions très large. Leur application nécessite des précautions : comme je l'ai mentionné plus haut, la recherche de principes généraux doit être entreprise avec discernement quand on parle d'organismes vivants, chaque espèce – voire chaque individu – étant particulier. Mais si la compréhension fine du fonctionnement cellulaire à l'échelle d'un faible nombre de gènes et de protéines peut être abordée expérimentalement, la compréhension globale du vivant ne peut que difficilement l'être à l'heure actuelle. Et c'est là que l'apport de la physique statistique est important : elle peut permettre de placer des cadres, des règles qui, bien que grossières, peuvent servir de guide à des expériences plus ciblées. Accéder à une connaissance globale n'est peut-être pas directement possible ; mais l'emploi des méthodes de physique statistique en génomique peut permettre de s'orienter, et de trouver les directions qui permettront d'y parvenir.

Cette thèse se situe donc au croisement de deux disciplines : la physique statistique, apportant les méthodes mathématiques et numériques, et la génomique, avec son immense jeu de données et toutes les questions d'actualité qu'elle pose. Cette interdisciplinarité

permet de poser un œil nouveau sur des problèmes anciens, et d'y apporter des éléments de réponse que seules les analyses à grande échelle permettent d'obtenir. Je me suis en particulier intéressé, durant ces trois années, au problème du biais d'usage de codons chez les organismes bactériens et leurs virus, les bactériophages. Ce problème date de l'époque des premières analyses comparatives de gènes entre organismes, et de nombreuses causes lui ont été associées, sans qu'aucune ne parvienne à expliquer toutes les observations faites à ce jour.

J'ai essayé de rendre cette thèse aussi accessible et intéressante que possible aux différents publics qui pourraient l'avoir en main. Les physiciens trouveront, je l'espère, suffisamment de détails sur les organismes et leur appareillage cellulaire dans les trois premiers chapitres – en parallèle avec une revue partielle de découvertes plus récentes –, pour pouvoir apprécier les détails des analyses ultérieures. Les biologistes de formation trouveront dans les chapitres 4 et 5 l'essentiel de ce qu'il est nécessaire de savoir pour bien appréhender l'intérêt des méthodes que j'ai employées durant mes travaux. Finalement, j'ai décrit mes travaux en mettant l'accent sur les résultats obtenus plus que sur les techniques utilisées, qui sont décrites dans les publications incluses dans cette thèse, et ma seule concession à la technicité est la description de la méthode de classification que nous avons employée, dans l'appendice B. L'appendice A, quant à lui, contient la description de mes travaux sur les écosystèmes hydrothermaux et la publication associée, également réalisés pendant cette thèse.

Table des matières

Table des figures

Première partie

Bactéries, phages et traduction

Chapitre 1

Organismes étudiés

Les deux types d'organismes étudiés lors de cette thèse sont les bactéries et leurs virus, les bactériophages. Ces organismes, dont l'étude a commencé il y a plus d'un siècle, n'ont toujours pas livré tous leurs secrets ; je ne prétendrai donc pas faire plus qu'une brève présentation de leur organisation dans cette première partie. La diversité observée dans le règne bactérien rend toute tentative de systématisation difficile. La recherche de lois générales qui dirigent le métabolisme ou la génétique des bactéries est un défi, car chaque espèce, voir chaque organisme, est unique. Néanmoins, je vais énoncer ici certains principes, qui semblent généraux. Avant de détailler l'appareillage moléculaire du système de traduction, je vais donc commencer mon étude par une description générale des bactéries et des bactériophages, du point de vue du microbiologiste et de celui du généticien. Je m'inspirerai pour cela de références classiques, à savoir Lewin (2004) and Brown (2007) pour la génétique, Prescott et al. (2002) pour la microbiologie et Berg et al. (2002) pour la description des processus cellulaires.

1.1 Bactéries

Il semble nécessaire d'expliquer les raisons qui peuvent pousser au choix de l'étude des bactéries plutôt que d'organismes comme l'homme ou d'autres eucaryotes supérieurs. Le choix des bactéries et des phages en tant qu'organismes modèles n'est pas anodin : du point de vue du physicien, ce sont parmi les organismes les plus simples, possédant le moins de structures internes, et ayant un mode de vie relativement aisé à appréhender. Les bactéries sont unicellulaires, ce qui facilite encore leur étude, bien que les difficultés rencontrées lors de l'étude d'organismes pluricellulaires resurgissent quand on se penche sur des problèmes à l'échelle des populations bactériennes. De plus, la facilité à cultiver des bactéries, et à les manipuler génétiquement, en font un système d'étude de choix pour réaliser des expériences, et donc pour avoir accès à des jeux de données variés et précis. Les bactéries sont étudiées depuis longtemps à cause de leur rôle prépondérant au niveau de l'environnement et de la santé : il y a en chaque être humain plus de bactéries que de cellules humaines et de nombreuses maladies connues (allant de la peste à la tuberculose) sont provoquées par des bactéries pathogènes. Ceci a permis l'accumulation d'un savoir relativement étendu à leur sujet, qui constitue une base solide pour commencer un travail. Finalement, le dernier argument est numérique : peu de systèmes biologiques sont aussi présents sur Terre que les bactéries, au nombre de 5.10^{30}, un nombre que l'on a plus l'habitude d'observer en physique qu'en biologie (Whitman et al., 1998).

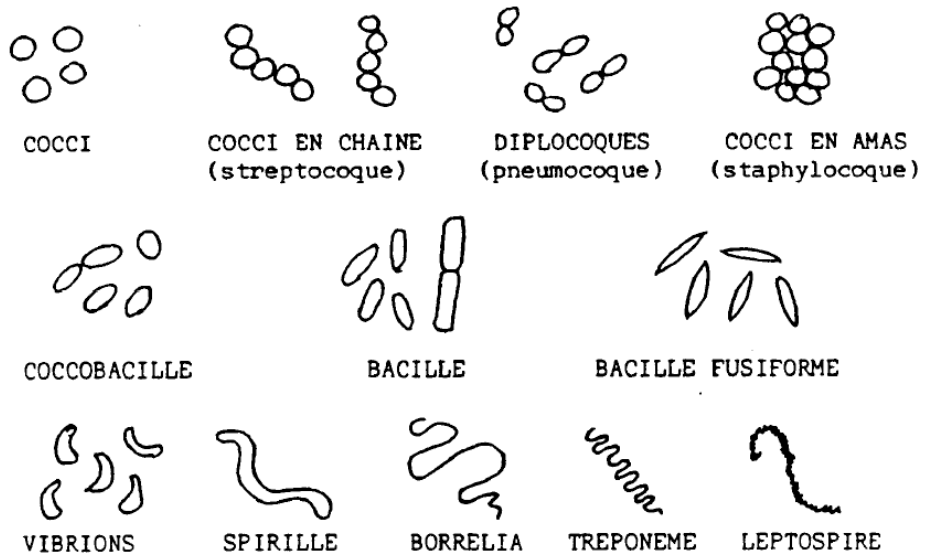


FIG. 1.1 – Exemples de morphologies bactériennes.

1.1.1 Carte d'identité d'une bactérie

Les bactéries sont des organismes unicellulaires de petite taille, d'une longueur pouvant aller de moins d'un μm pour le genre *Mycoplasma* à quelques centaines de μm pour les plus grandes, comme *Thiomargarita namibiensis* qui mesure 750 μm et est visible à l'œil nu. Leur forme est variable selon les espèces (Fig. 1.1) : sphérique – on parle de coque –, allongée – un bacille –, hélicoïdale – un spirochète –, ou même carrées. À la surface de certaines se trouvent des protubérances, qui peuvent être des flagelles servant à la motilité ou des fimbriae.

Les bactéries sont des organismes procaryotes, sans noyau. Des considérations physiologiques simples ont longtemps fait croire que tous les procaryotes formaient un seul grand groupe d'espèces monophylétique. En réalité, les eubactéries sont un des trois grands domaines du vivant, avec les eucaryotes et les Archaea. Ces dernières, au vu des ressemblances physiologiques qu'elles offrent avec les eubactéries, ont été confondues avec elles jusqu'aux travaux de C. Woese, qui ont montré que leur contenu génétique pouvait être très différent, en se basant sur des phylogénies d'ARN 16S (Fox et al., 1977). Le terme d'Archaea, historique, ne doit pas être trompeur : ces organismes ne sont pas plus vieux que les bactéries. Les Archaea partagent certaines caractéristiques physiologiques avec les bactéries, mais d'autres avec les eucaryotes. On les trouve souvent dans des environnements extrêmes, même si certaines eubactéries[1] vivent également dans de tels environnements.

[1]On emploiera à partir de maintenant le terme bactérie plutôt que eubactérie, par souci de simplicité.

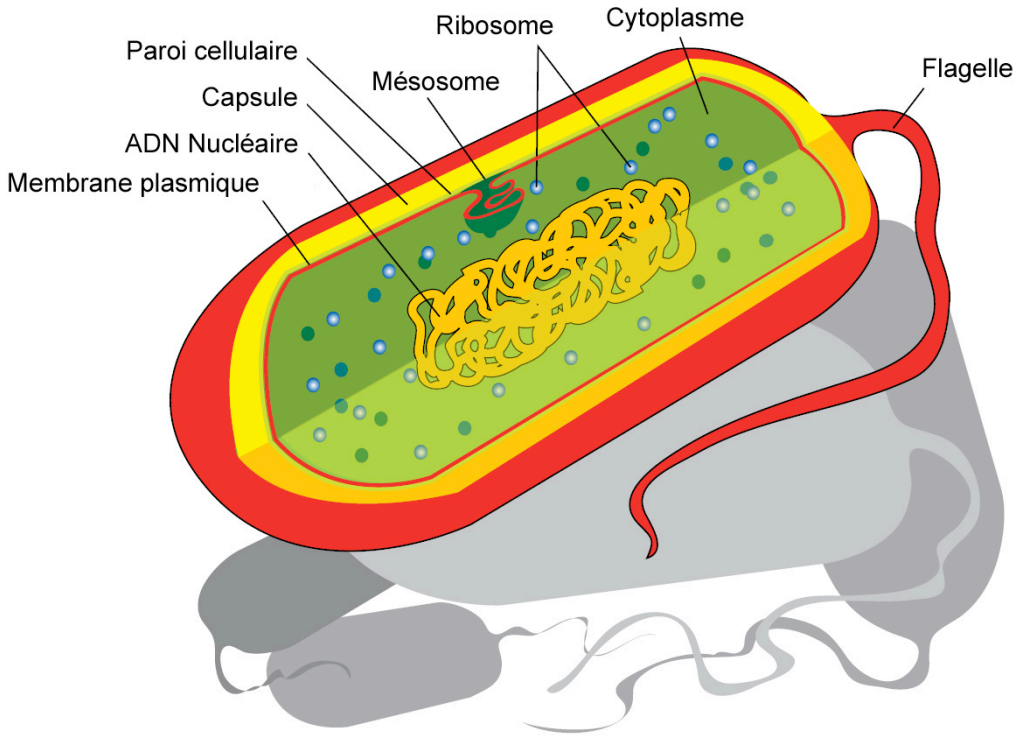


FIG. 1.2 – Diagramme d'une cellule procaryote.

La structure cellulaire des bactéries est relativement simple. On trouve, en partant du centre de la cellule et en allant vers l'extérieur (Fig. 1.2) :

- Un nucléoïde, région contenant essentiellement l'ADN bactérien. Contrairement au noyau eucaryote, le nucléoïde bactérien est dépourvu de membrane : il s'agit simplement d'une région dans le cytoplasme, dans laquelle la densité d'ADN est élevée. L'ADN y est présent sous une forme surenroulée très compacte, formant entre 40 et 50 branches radiant du centre du nucléoïde. Le surenroulement est favorisé par des protéines semblables aux histones eucaryotes, les protéines HU.
- Le cytoplasme, dans lequel on trouve les protéines impliquées dans tous les processus de la cellule, métabolisme, réplication, transcription ou traduction, ainsi que les corps d'inclusions, vésicules aux contenus variés. Les corps d'inclusion les plus courants contiennent du glycogène, qui sert de réserve d'énergie et de carbone à la cellule. La vacuole gazeuse, qui contient des gaz atmosphériques, permet par exemple aux bactéries marines de flotter à la surface, en augmentant leur volume.
- La membrane plasmique, qui isole le cytoplasme et le nucléoïde du milieu extérieur. Cette membrane est composée d'une bicouche de phospholipides, dans laquelle sont insérées des protéines membranaires. C'est l'interface de la cellule avec le milieu extérieur : certaines protéines membranaires permettent des échanges d'ions (canaux ioniques) et d'eau (aquaporines) entre le milieu extérieur et le cytoplasme. De plus, il a souvent été observé que la membrane est rattachée au nucléoïde ; on pense que cela permet le partage précis du matériel génétique lors de la division cellulaire.
- Presque toutes les bactéries, à l'exception des mycoplasmes, disposent en plus d'une

paroi, qui procure à la bactérie une meilleure isolation du milieu extérieur. Il existe deux grands types de parois, nommés en fonction de la couleur obtenue lors du test à la coloration de Gram. On distingue :

i) Les bactéries Gram-positives, dont la paroi est formée d'une épaisse couche de peptidoglycanes, d'environ 70 nm d'épaisseur.

ii) Les bactéries Gram-négatives, dont la paroi est formée tout d'abord d'une mince couche de peptidoglycanes (1 à 3 nm d'épaisseur), suivie d'un espace périplasmique puis d'une nouvelle membrane, dite externe, formée d'une nouvelle bicouche de phospholipides.

– Autour de la paroi, on peut trouver chez certaines bactéries des structures très résistantes, qui servent à protéger l'organisme de dangers plus importants. Dans cette catégorie on trouve par exemple la capsule bactérienne, composée de polysaccharides, qui permet à *Streptococcus pneumoniae* de ne pas être phagocyté par les macrophages, ou les endospores fabriquées par *Bacillus subtilis* pour survivre longtemps dans des conditions extrêmes sous une forme dormante.
– À la surface de la paroi de certaines bactéries on observe des structures dédiées à la motilité ou à l'adhésion, respectivement des flagelles (15 à 20 μm de long) ou les plus petits fimbriae (quelques μm de long). Ces structures ont en fait leurs racines au niveau de la membrane plasmique, et traversent les différentes couches jusqu'au milieu extérieur. L'adhésion et la motilité bactérienne sont des problèmes d'importance en particulier dans l'étude des biofilms, des communautés bactériennes qui se développent sur une surface en sécrétant une matrice extracellulaire. Le travaux consacrés à ces populations prennent de plus en plus de place dans la microbiologie moderne, à cause de leur intérêt relatif à des problèmes environnementaux et de santé (Webb et al., 2003).

Les bactéries vivent dans des milieux très variés. Certaines sont associées à un hôte, soit de manière intracellulaire, soit, dans le cas d'un hôte pluricellulaire, par adhésion à la surface de ses cellules. Parmi elles, plusieurs sont pathogènes, et c'est ce qui a très tôt motivé leur étude. Nombre de maladies sont causées par des bactéries, de la gastroentérite, qui peut être provoquée par *Campylobacter jejeuni*, au redoutable choléra causé par *Vibrio cholerae*, ou à la lèpre (*Mycobacterium leprae*). La pathogénicité des bactéries est un des sujets les plus étudiés en bactériologie, que ce soit du point de vue moléculaire ou génétique (Finlay and Falkow, 1997). D'autres bactéries qui vivent dans un hôte ont des effets plus neutres pour lui : la plus connue de celles-ci est *Escherichia coli*, qui vit naturellement à l'intérieur de l'intestin de nombreux mammifères, dont l'homme, et qui se nourrit des aliments ingérés. On parle alors de commensalisme. Finalement des relations réciproquement profitables existent aussi, comme les bactérie du genre *Rhizobium*, qui vivent en relation symbiotique dans des plantes, et fixent l'azote en échange de la nourriture fournie par l'hôte. On parle alors de symbiose ou de mutualisme. D'autres modes de vie existent pour les bactéries, que l'on retrouve dans tous les milieux : le sol, les milieux aquatiques, mais aussi dans la croûte terrestre (*Bacillus infernus* vit à 2700 m sous la surface du sol (Boone et al., 1995)) ou dans des cheminées hydrothermales (*Pyrococcus fumarii* (Blochl et al., 1997)), ou encore dans les glaces de l'Antarctique (*Pseudoalteromonas haloplanktis TAC125*, (Médigue et al., 2005)). Ces bactéries sont extrémophiles, c'est à dire qu'elles vivent dans des milieux *a priori* difficiles. D'autres types d'extrémophiles vivent dans des milieux très salés, comme la Mer Morte (on parle d'halophiles), ou très

acides (acidophiles). Ce spectre très large d'habitats est un bon exemple de la diversité des organismes eux-mêmes.

Le cycle de vie bactérien est, comme sa structure, relativement simple. En effet la majorité des microorganismes croissent simplement en taille, puis répliquent leur matériel génétique et se divisent en deux cellules filles identiques, se partageant le contenu cellulaire. Le temps nécessaire à une cellule pour se diviser varie en fonction des conditions de température, d'oxygénation, et de façon plus générale en fonction de la qualité de l'environnement. Ce temps peut être très court : les cellules de *Pseudomonas natriegens* ont un temps de génération de moins de 10 minutes (Eagon, 1962), tandis que la plus classique *E. coli* est capable de se diviser toutes les 20 minutes – ce qui est déjà une performance, ce temps étant plus court que celui nécessaire à la réplication du génome –, d'où son utilité en tant qu'organisme de laboratoire. D'autres organismes ont des cycles de vie plus longs, comme *Mycobacterium tuberculosis* qui se réplique en 16 heures. On peut cependant remarquer que les organismes qui se divisent le plus vite, quand ils sont dans les bonnes conditions, n'ont matériellement pas le temps pour d'autre activité que de doubler leur contenu cellulaire, répliquer leur ADN et se diviser. À ce niveau, les contraintes sur les vitesses de réplication et de synthèse protéique deviennent très importantes, et ces organismes sont déjà à la limite de ce qui est réalisable d'un point de vue biochimique.

1.1.2 Données génomiques

Au jour de l'écriture de ce paragraphe[1], on recense 479 génomes bactériens complètement séquencés (dont 37 d'Archaea) et 707 en cours de séquençage. Dans peu de temps, plus de 1000 génomes seront disponibles. Cette profusion de données génomiques permet de donner une idée de la composition moyenne d'un génome bactérien, bien qu'il faille se garder de généraliser trop vite : l'échantillon des génomes disponibles est surtout représentatif des organismes facilement cultivables en laboratoire, et ayant un intérêt médical ou économique. Cet échantillon est peut-être très biaisé, quand on sait que le nombre des espèces réellement existantes a été estimé à 10^6 (Whitman et al., 1998). De plus, même parmi les espèces connues, il existe une grande variabilité entre individus au niveau génétique, et les études réalisées sur une lignée particulière ne sont pas facilement généralisables. Notre présentation des génomes bactériens et de leurs caractéristiques ne sera donc qu'indicative, quoique fidèle à l'état des connaissances actuelles.

L'ADN bactérien est soit circulaire, soit linéaire. Il se décompose en un ou deux chromosomes, accompagnés de plasmides plus petits qui peuvent le complémenter, et parfois en former une grande partie. Deux exemples d'organisation très différents sont *E. coli* K12 et *Borrelia burgdorferi* B31 : *E. coli* a un unique chromosome circulaire, et est encore aujourd'hui utilisée comme un modèle d'organisation génomique. *B. burgdorferi* voit son génome divisé en un chromosome linéaire, et 21 plasmides, dont 10 sont linéaires. Dans ce cas particulier les plasmides composent environ 36% du génome, et contiennent de nombreux gènes essentiels, comme ceux codant pour la biosynthèse de la membrane.

La taille d'un génome bactérien peut varier sur deux ordres de grandeur, allant de 160 kb pour *Candidatus Carsonella ruddii* PV à 10 Mb pour *Solibacter usitatus* Ellin6076, le plus grand génome séquencé à ce jour (voir Fig. 1.3). Le nombre de gènes contenus dans un génome est environ de 1 pour 1 kb de longueur ; l'exemple typique en est *E. coli* K12,

[1] Le 7 avril 2007.

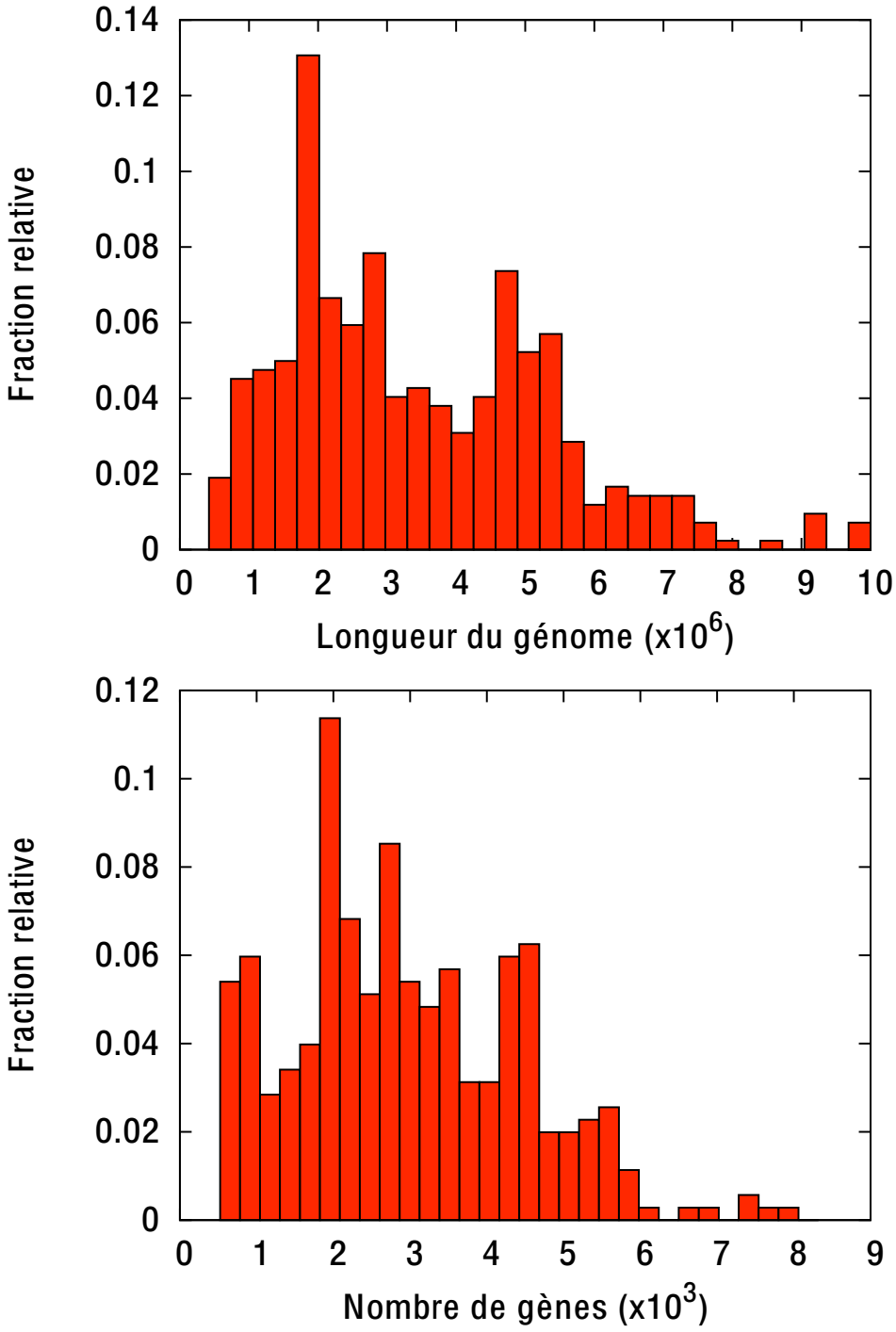


FIG. 1.3 – Distribution des longueurs des génomes bactériens séquencés (en haut) et de leur nombre de gènes (en bas).

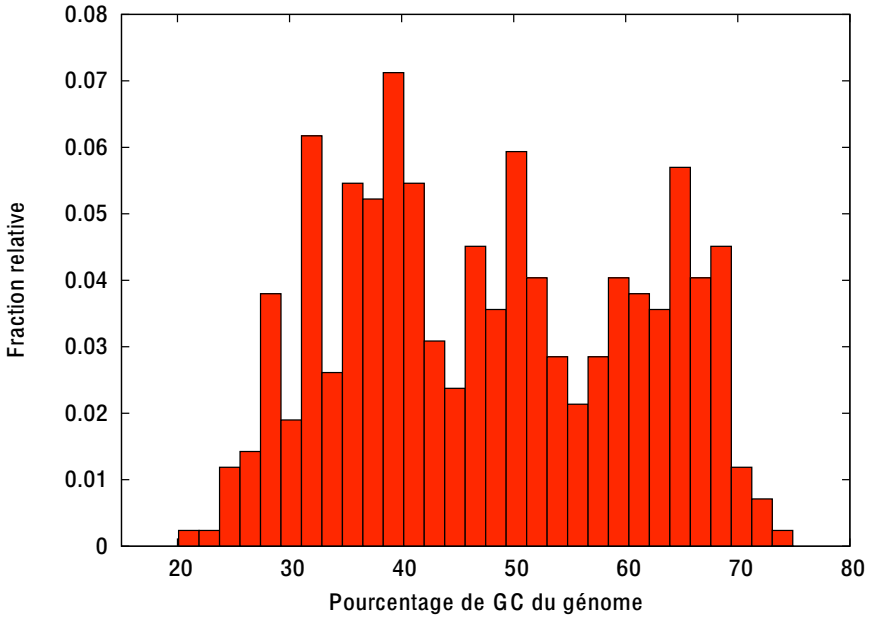


FIG. 1.4 – Distribution du pourcentage de GC dans les génomes bactériens.

avec 4400 gènes pour une longueur de 4.6 Mb. Ce nombre relativement élevé de gènes par rapport à la longueur du génome est possible grâce à une densité en gènes beaucoup plus élevée que celle des génomes eucaryotes, avec 89% de la séquence complète codant pour des protéines chez *E. coli* par exemple.

Finalement, la dernière caractéristique montrant la diversité qui peut régner dans les génomes bactériens est le pourcentage en GC, défini comme le rapport du nombre de bases G et C sur le chromosome au nombre de bases total. Il peut varier énormément selon les espèces, allant de 16.6% chez *C. Carsonella ruddii PV* à près de 75% chez *Anaeromyxobacter dehalogenans 2CP-C* (voir Fig. 1.4). Les capacités de codage de ces génomes très biaisés sont affectées, un pourcentage en GC très différent de 50% impliquant d'énormes contraintes sur les protéines qui sont synthétisées par l'organisme (voir la section "Code génétique", chapitre 3, page 63). Le fait que les organismes ayant des génomes courts semblent avoir un faible pourcentage en GC a posé la question des liens entre pourcentage en GC, longueur des génomes et température optimale de croissance. Ces liens ont été étudiés récemment par de nombreuses équipes (Galtier and Lobry, 1997; Hurst and Merchant, 2001; Musto et al., 2006; Wang et al., 2006). Les résultats des analyses sont contradictoires, trouvant ou non des corrélations significatives entre ces variables, et jetant un doute sur l'hypothèse simple selon laquelle le taux de GC devrait croître avec la température, car il stabilise thermiquement le génome. Pour d'autres traits la situation est plus claire : par exemple le pourcentage en GC est corrélé au mode de vie de la bactérie, les organismes au mode de vie parasitique ayant un taux de GC plus faible que les génomes de leurs hôtes (Rocha and Danchin, 2002). Au contraire, le contenu en GC des organismes bactériens est plus élevé chez les organismes aérobies, mais les causes de

ce biais sont floues : l'explication de ce phénomène par la stabilisation thermodynamique de l'ADN face aux dommage oxidatifs ne semble pas être le facteur dominant de ce biais, et la composition en acides aminés semble jouer un rôle prépondérant sur cette variation (Naya et al., 2002).

1.1.3 Organisation chromosomique des génomes bactériens

Les génomes bactériens sont hautement organisés. Les gènes ne sont pas placés aléatoirement sur le chromosome, mais regroupés d'une façon bien particulière, généralement sous la forme d'opérons. Un opéron est un ensemble de gènes très proches sur le chromosome, transcrits dans le même sens. Ils sont souvent régulés par un mécanisme commun en amont du premier gène. Cela implique que les gènes appartenant à un opéron sont transcrits simultanément, et que leurs produits vont être présents au même moment et au même emplacement dans la cellule. Mais d'autres mécanismes plus fins existent, par exemple la régulation de polarité de l'opéron *gal* par un ARN non codant, Spot42, chez *E. coli*. Celui-ci réprime spécifiquement l'expression du premier gène de l'opéron en s'appariant à sa séquence d'ADN, sans influencer sur les autres (Gottesman, 2004). Chez les procaryotes, une grande fraction du génome est organisée de la sorte ; par exemple, chez *B. subtilis* (Kunst et al., 1997), il existe au total 1049 opérons identifiés, et ils contiennent 3177 de ses 4225 gènes, soit plus de 75% de son génome. Chez *E. coli* K12, les chiffres sont un peu moins élevés, mais on trouve quand même 600 opérons (Blattner et al., 1997) d'une longueur moyenne d'environ 3 gènes. Ces quelques valeurs montrent que l'organisation en opérons est la règle plutôt que l'exception chez les bactéries.

Les gènes regroupés ont très souvent des fonctions liées ; ils peuvent participer à la même voie métabolique, ou produire des protéines qui doivent interagir pour jouer leur rôle. Deux exemples classiques sont l'opéron lactose (Jacob and Monod, 1961), et l'opéron de la voie de biosynthèse du tryptophane, tous deux présents chez *E. coli*. L'opéron lactose contient 3 gènes, *lacZ*, *lacY* et *lacA*. Le promoteur de cet opéron est sensible à la présence de lactose, un sucre complexe qui ne peut pas être assimilé directement par l'organisme. Quand celui-ci est présent dans l'environnement, la répression de l'opéron cesse et les trois gènes sont exprimés simultanément (voir Fig. 1.5) :

- *lacZ* code pour la β-galactosidase, une enzyme qui peut cliver le lactose en glucose et galactose plus facilement utilisable par l'organisme.
- *lacY* code pour une perméase, qui peut importer le lactose de l'extérieur de la cellule à l'intérieur.
- *lacA* acétyle les β-galactosides, ce qui permet l'emploi d'un autre mécanisme pour les expulser de la cellule. En effet, la perméase codée par *lacY* est tellement efficace que, sans régulation du type imposé par l'acétylase, la cellule courrait le risque d'exploser...

Ces trois produits protéiques peuvent donc avoir une action coordonnée, permettant à la cellule d'importer le lactose et de le dégrader en sucres plus petits directement utilisables. L'opéron lactose est étudié depuis longtemps (Miller and Reznikoff, 1978) et son promoteur est maintenant utilisé de manière courante dans les techniques de manipulation génétique, pour contrôler l'expression de gènes.

Un autre opéron est celui de la voie de biosynthèse du tryptophane. Le tryptophane est un acide aminé aromatique contenant un hétérocycle indole. Il est essentiel à la synthèse protéique, mais n'est pas présent dans tous les milieux, ce qui implique que les organismes

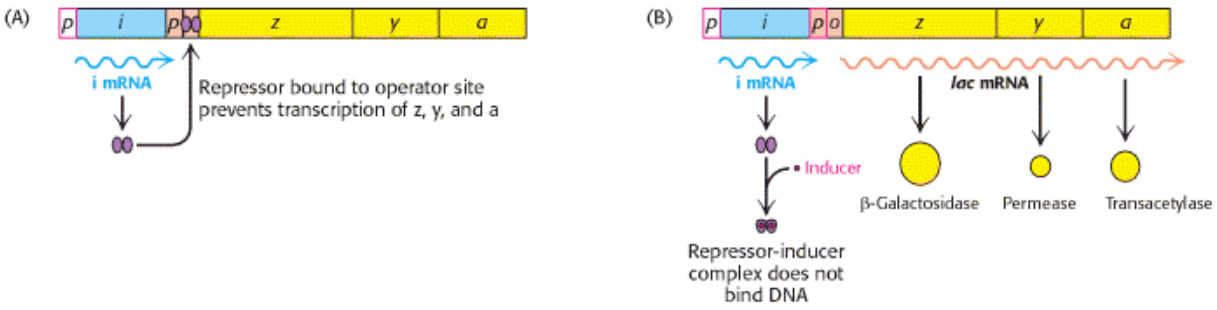


FIG. 1.5 – Schéma de l'opéron lactose. Le gène *laci* est un répresseur qui empêche l'expression de l'opéron en absence de lactose, mais n'appartient pas au même opéron.

comme *E. coli* doivent être capables de le synthétiser à partir d'autres substrats, dans ce cas l'acide chorismique. L'opéron contient 5 gènes, codant pour 5 enzymes qui servent toutes à catalyser une transformation intermédiaire de la voie de biosynthèse. Cet opéron est régulé par atténuation transcriptionnelle[1] (Gollnick and Babitzke, 2002), c'est à dire qu'il est réprimé en présence de tryptophane dans le milieu extérieur. Ceci permet, ici aussi, de n'exprimer un groupe de gènes que grâce à un seul signal activateur, et d'exprimer tous les gènes nécessaires au fonctionnement de la voie métabolique de façon coordonnée.

Un autre facteur d'organisation des génomes procaryotes est l'organisation relative des gènes sur les deux brins de l'ADN. En effet les deux brins ne sont pas équivalents, et sont caractérisés par leur orientation et leur placement par rapport à l'origine de réplication. Tout d'abord, l'ADN simple brin est une molécule orientée : les bases qui le composent, plus précisément les désoxyriboses dont elles sont formées, sont alternativement liés par leur position 5' et leur position 3'. Sous sa forme double brin, l'ADN voit ses deux brins complémentaires orientés dans des directions opposées : on dit qu'un brin est orienté de 5' vers 3', tandis que l'autre va de 3' vers 5'[2]. Or, sur un chromosome circulaire, la réplication se produit à partir d'un point bien défini, nommé origine de réplication, et caractérisée par des sites d'accrochage pour la protéine *dnaA*, qui va initier la réplication. On observe ensuite la création de deux fourches de réplication qui partent dans des directions opposées, et s'arrêtent à l'arrivée à un terminus de réplication, le plus souvent diamétralement opposé à l'origine sur le génome. Cette réplication a donc un sens de propagation défini par rapport à l'orientation de l'ADN si on connaît l'origine et le terminus de réplication. On définit localement le brin précoce, sur une portion de l'ADN, comme celui qui est synthétisé de 5' vers 3' dans le sens d'avancée de la fourche de réplication. Le brin tardif, lui, est dans le sens contraire, donc celui que la polymérase synthétise "apparemment" de 3' vers 5' (voir Fig. 1.6).

Au niveau moléculaire, les ADN polymérases, qui synthétisent les nouveaux brins d'ADN au niveau de la fourche lors de la réplication, ne fonctionnent que sur un substrat orienté de 3' vers 5' (et donc synthétisent un brin de 5' vers 3'). Le brin précoce est donc assemblé de manière continue lors de la réplication, mais ce n'est pas le cas pour le brin tardif. Celui-ci est synthétisé de manière discontinue par une ADN polymérase, sous la forme de petits fragments d'ADN nommés fragments d'Okazaki, du nom de leur

[1] Voir la section "ARN de transfert" du chapitre suivant, page 43.
[2] Par convention, les séquences d'acides nucléiques sont lues de 5' vers 3'.

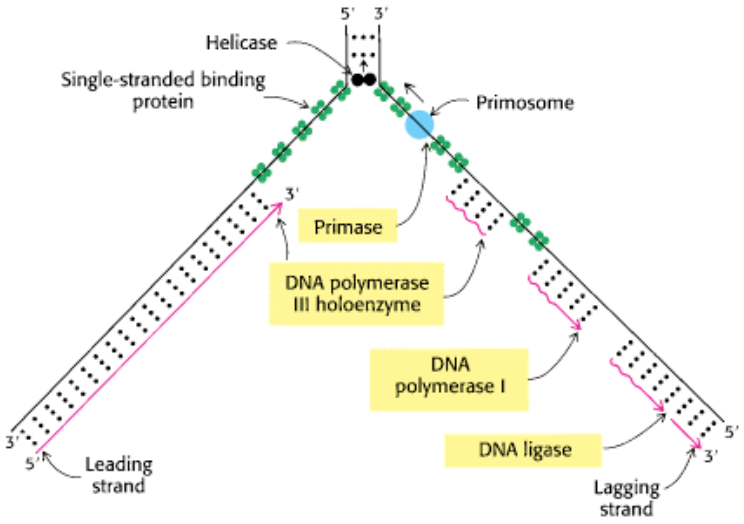


FIG. 1.6 – Schéma de la réplication chez les procaryotes. On voit que le nouveau brin tardif, complémentaire de l'ancien brin précoce, est formé de petits fragments d'ADN, tandis que le nouveau brin précoce est formé continûment.

découvreur. Ces fragments sont ensuite reliés entre eux par une ADN ligase, pour obtenir la copie conforme du brin tardif. Ce mécanisme a pour conséquence un temps d'exposition plus long des bases du complémentaire du brin tardif sous une forme simple brin, ce qui change les probabilités de mutation entre les deux brins lors de la réplication. Ce point sera étudié plus en détails au chapitre 3 (voir page 75). Les différences entre les deux brins ne viennent pas que de la manière dont ils sont répliqués. Au niveau génomique, les gènes peuvent se répartir sur les deux brins d'ADN. En pratique on observe que beaucoup d'organismes ont un biais pour avoir leurs gènes situés sur le brin précoce. Par exemple, *B. subtilis* voit 75% de ses gènes sur son brin précoce. Pour d'autres organismes le biais n'est visible qu'au niveau des gènes essentiels[1] : les 7 opérons codant pour les ARN ribosomaux d'*E. coli* sont tous sur le brin précoce (Guy and Roten, 2004). Une explication de ce phénomène vient des interactions entre l'ADN polymérase de la fourche de réplication et l'ARN polymérase qui effectue la transcription. Cette dernière étant environ 20 fois plus lente que la réplication – avec une vitesse de 50 nucléotides par seconde contre 1000 pour la réplication –, des collisions entre les deux appareils moléculaires peuvent fréquemment se produire dans l'organisme. Mais l'ARN polymérase va toujours effectuer la transcription dans le sens 3' vers 5' du brin antisens. Lors de la transcription d'un gène situé sur le brin précoce, la polymérase va donc être localisée sur son complémentaire, le brin tardif, sur lequel la fourche de réplication avance dans le même sens que l'ARN polymérase. Donc les collisions au niveau de gènes codés sur le brin précoce se font alors que les deux complexes fonctionnent dans le même sens, tandis qu'au niveau des gènes du brin tardif ils se percutent de plein fouet. Il a été démontré expérimentalement que les chocs frontaux des deux systèmes étaient plus délétères pour l'organisme que les pauses générées par les

[1]Un gène essentiel est un gène dont l'expression est nécessaire à la survie de l'organisme.

C	Production et conversion d'énergie
D	Mitose et contrôle du cycle cellulaire
E	Métabolisme et transport des acides aminés
F	Métabolisme et transport des nucléotides
G	Métabolisme et transport des carbohydrates
H	Métabolisme des coenzymes
I	Métabolisme des lipides
J	Traduction
K	Transcription
L	Réplication et réparation de l'ADN
M	Biogénèse de la membrane et de la paroi cellulaire
N	Motilité cellulaire
O	Chaperons, dégradation des protéines et modifications post-traductionnelles
P	Métabolisme et transport des ions inorganiques
Q	Biosynthèse, transport et catabolisme des métabolites secondaires
R	Activité biochimique non associée à une fonction
S	Fonction inconnue
T	Transduction du signal
U	Trafic intracellulaire et sécrétion

TAB. 1.a – Liste des 19 catégories de fonctions dans la classification COG.

rencontres des deux machineries évoluant dans le même sens (French, 1992; Mirkin and Mirkin, 2005). Ceci crée une sélection qui favorise les organismes ayant leurs gènes plutôt sur le brin précoce. Au départ, cette sélection était supposée avoir lieu en particulier chez les organismes à croissance rapide, et sur les gènes fortement exprimés (Brewer, 1988). Des résultats ultérieurs ont montré que la sélection, en réalité, s'exerçait préférentiellement sur les gènes essentiels de l'organisme (Rocha and Danchin, 2003a,b), et dépendent de l'ADN polymérase effectuant la réplication (Rocha, 2002). Ces résultats montrent comment le génome peut être structuré et optimisé à grande échelle, relativement à des contraintes complexes, faisant intervenir l'interaction de plusieurs systèmes moléculaires.

1.1.4 Les gènes bactériens et leurs fonctions

La très grande diversité observée au niveau génomique chez les bactéries contraste avec l'homogénéité des fonctions réalisées par le métabolisme bactérien. Bien que certains gènes présents dans des organismes particuliers n'aient pas d'homologues ou d'équivalents dans d'autres microbes, la majorité des gènes identifiés se partagent un ensemble de fonctions assez restreintes (Riley, 1993). Ces fonctions ont été rassemblées en catégories de gènes ayant des séquences homologues et ayant des fonctions similaires dans la classification COG (Tatusov et al., 2003, 2000). Cette classification décompose l'activité cellulaire des organismes microbiens en 19 catégories, chacune représentant une gamme de fonctions particulières. Ces fonctions donnent une idée du paysage des fonctions des organismes microbiens, et nous les donnons pour présenter brièvement le fonctionnement cellulaire (Table 1.a).

On voit que les fonctions bactériennes se répartissent en 3 grands types, quelque peu arbitraires :

- Les catégories "Métabolisme et transport" représentent les gènes codant pour des protéines qui, soit importent du milieu extérieur certaines molécules utiles, soit les construisent à partir d'autres substrats (E, F, G, H, I, P, Q).
- Un sous-ensemble des fonctions bactériennes est la synthèse de protéines (J, K, O) et d'ADN (L).
- Un autre est la régulation des autres fonctions, via la communication avec l'extérieur et les liens avec l'environnement (C, M, N, T, U).

Une partie non négligeable du matériel génétique bactérien est présent dans le génome sous la forme de prophages, ou plus généralement de séquences insérées composées d'ADN mobile, par exemple les "séquences d'insertion" ou IS (Lawrence et al., 1992) similaires à des transposons ou des îlots de pathogénicité (Dobrindt et al., 2002; Oelschlaeger et al., 2002). Ces séquences sont caractérisées par la présence d'une enzyme de recombinaison, intégrase, transposase ou invertase, qui permet l'insertion et parfois l'excision de la séquence dans un génome. Elles sont parfois mutées au point de ne plus être actives. En effet, certaines séquences de ce type, à l'origine mobile, sont désormais fixées dans le génome hôte. Dans le cas d'une séquence de prophage qui ne peut plus être induite, on parle de conversion lysogénique. Certains cas de conversion lysogénique permettent à la bactérie de gagner de nouvelles fonctions, et de nombreux cas de pathogénicité bactérienne sont dus à l'expression d'un gène situé sur un prophage incorporé au génome bactérien. Une des plus tristement célèbre de ces bactéries est *Yersinia pestis* CO92, qui a causé – entre autres – la peste noire au XIV$^{\text{ème}}$ siècle et tué un tiers de la population européenne (Parkhill et al., 2001). Son génome contient de nombreuses séquences d'insertion, sources de sa pathogénicité. Un autre exemple moins dramatique pour l'homme sont les bactériocines sécrétées par *Pseudomonas aeruginosa* PAO1, similaires à des protéines de phages (Casjens, 2003). Elles sont probablement d'origine virale, et leur rôle est de de détruire les bactéries aux alentours. Ces séquences d'insertion, d'une façon générale, ont été beaucoup étudiées, et sont maintenant acceptées comme une des principales sources de variabilité entre souches : quasiment toutes les différences au niveau génétique, entre *E. coli* K12 et *E. coli* O157H7, sont dues à des séquences d'insertion (Ohnishi et al., 1999).

1.1.5 Évolution des génomes bactériens

Les génomes bactériens évoluent beaucoup, et grâce à des processus variés (Woese, 1987). Le faible temps nécessaire aux bactéries pour se reproduire permet une évolution rapide, combiné à leur ancienneté, qui a conduit à des divergences énormes entre espèces. Par exemple, l'ancêtre commun de *E. coli* et *B. subtilis* vivait il y a plus d'un milliard et demi d'années, soit bien avant celui de tous les vertébrés, ou même celui de l'homme et de la levure...

Les processus évolutifs qui ont lieu dans les génomes bactériens sont les mêmes que dans toute forme de vie. Mais grâce à leur taux de reproduction élevé, et à leur grande population efficace[1], les résultats de la sélection sont très visibles chez les bactéries. Nous allons ici présenter brièvement les principaux processus qui dirigent l'évolution bactérienne. En premier viennent les mutations dans les séquences génétiques. Ces mutations, dont

[1]En génétique des populations, la taille efficace d'une population est le nombre d'individus participant au processus reproductif.

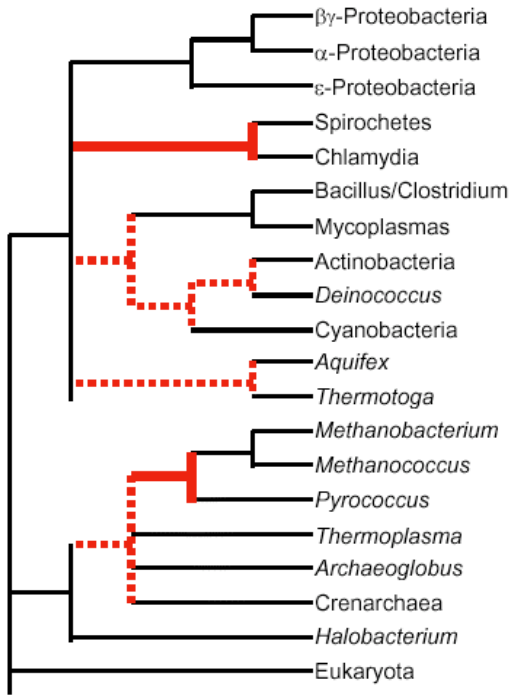


FIG. 1.7 – Phylogénie des bactéries et des Archaea. De bas en haut, on observe sur cet arbre les eucaryotes, le branchement des Archaea et celui des bactéries. Les liens pointillés sont incertains.

on a estimé le taux à 10^{-9} par base et par génération (voir Denamur and Matic (2006) pour une discussion récente de ces chiffres), ont depuis Darwin été considérées comme le moteur dominant de l'évolution des génomes. Cette hypothèse a d'ailleurs permis les premières analyses phylogénétiques sur des séquences génétiques, qui ont conduit à un certain degré d'unification des connaissances taxinomiques (voir Fig. 1.7).

Cependant, les taux de mutation bénéfiques sont suffisamment faibles, pour qu'un autre processus, qu'on supposait anecdotique jusqu'à ces dernières années, puisse les concurrencer. Il s'agit du transfert horizontal de gènes, où l'incorporation de gènes ou de portion de gènes d'une espèce dans une autre. Grâce à des analyses récentes, ce mécanisme évolutif est maintenant reconnu comme un des processus majeurs de l'évolution des génomes bactériens (Jain et al., 1999; Médigue et al., 1991). Plusieurs mécanismes d'acquisition de nouvelles séquences d'ADN étaient pourtant connus depuis longtemps, mais trop mal caractérisés pour qu'on soupçonne leur importance. L'acquisition d'ADN localisé à l'extérieur de leur membrane par des bactéries compétentes, nommée transformation, a été longuement étudiée et caractérisée, mais on suppose qu'elle ne se produit que dans 1% des cas à l'état naturel, permettant de récupérer des séquences d'ADN libérées par exemple par la lyse d'une bactérie voisine dans une population (Dubnau, 1991). Une autre méthode connue de transfert génique qui n'a lieu que dans des conditions spéciales est la conjugaison bactérienne, par laquelle deux cellules en contact peuvent échanger de l'ADN (Narra and Ochman, 2006). Par contre, il est connu que des phages tempérés – ou même de simples capsides codées par un prophage dégénéré (Humphrey et al., 1997) – peuvent transporter des séquences d'ADN d'un hôte à l'autre, et que ces séquences ont des chances de se voir incorporées dans le génome du nouvel hôte. Dans ce cas on parle

de transduction. De plus, les phages tempérés élargissent les possibilités évolutives des bactéries, par leur insertion et leur potentielle conversion lysogénique (Brussow et al., 2004). En effet, des sections prophagiques très variées, ayant une structure en mosaïque, ont été identifiées dans un grand nombre de génomes bactériens (Welch et al., 2002). On sait d'ailleurs maintenant qu'elles peuvent provoquer une évolution très rapide, et on suppose même qu'elles sont à l'origine de nouveaux gènes (Daubin and Ochman, 2004). Ceci montre qu'en plus de diversifier les bactéries, les séquences prophagiques, en évoluant à mi-chemin des séquences bactériennes et phagiques, permettent un brassage très rapide des gènes et un gain rapide de fonction par leurs hôtes. C'est également l'entrée massive de ces processus évolutifs dans la balance qui ont brouillé les frontières entre espèces et introduit la notion de réseau phylogénétique, à comparer à celle d'arbre phylogénétique, qui met l'accent sur l'évolution verticale (Doolittle, 1999).

Un mécanisme complémentaire du transfert horizontal sous toutes ses formes est la dégradation génétique. En effet, si le transfert horizontal avait lieu seul, on observerait une accumulation de nouveaux gènes ou du moins de nouvelles séquences dans les génomes bactériens (Lawrence et al., 2001). Or il n'en est rien, et ces génomes ne montrent que peu de séquences non fonctionnelles. Ceci implique l'existence d'une pression très forte de dégradation à l'intérieur des génomes bactériens, à savoir une tendance à systématiquement perdre des séquences d'ADN. Les causes sélectives de ce biais ne sont pas dévoilées, mais pourrait être liées au fait que la réduction de la taille du génome permettrait d'accélérer la réplication cellulaire (Selosse et al., 2001). Cette hypothèse semble cependant fragile au vu de l'absence de corrélations notables entre longueur des génomes et taux de croissance chez les bactéries, et en conséquence il a été supposé que cette perte systématique de séquences génétiques pouvait être un moyen de défense contre les séquences parasites invasives, comme les séquences de phages (Mira et al., 2001). Quelle qu'en soit la raison, le biais de dégradation existe bien, et a été observé chez des organismes ayant subi une énorme réduction génomique, comme par exemple le genre *Buchnera*, un symbiote intracellulaire obligatoire, ou le pathogène *Rickettsia*. Le fonctionnement du mécanisme de dégradation n'est pas évident ; il est possible que de grandes séquences d'ADN disparaissent avec une faible fréquence, ou que les séquences soient tout d'abord soumises à une dégradation fonctionnelle par des mutations avant d'être sélectivement supprimées du génome. Un cas supportant cette hypothèse est le génome de *M. leprae*, qui contient de très nombreux pseudogènes, et serait peut-être une étape évolutive intermédiaire (Cole et al., 2001).

1.2 Phages

1.2.1 Carte d'identité

Les bactériophages – souvent abrégé en phages – sont des virus infectant spécifiquement les bactéries. Le même nom est donné à ceux qui infectent les Archaea, même si leurs propriétés sont différentes (Prangishvili et al., 2006). Ils ont été identifiés en 1917 par Félix d'Herelle (d'Herelle, 1922), qui a immédiatement vu leur potentiel médical pour lutter contre les bactéries (la tumultueuse vie de d'Herelle, ainsi qu'une description de ses expériences sur la thérapie phagique, sont résumés dans Häusler (2006)). Ils sont formés d'une capside protéique qui contient leur matériel génétique. La nature de la

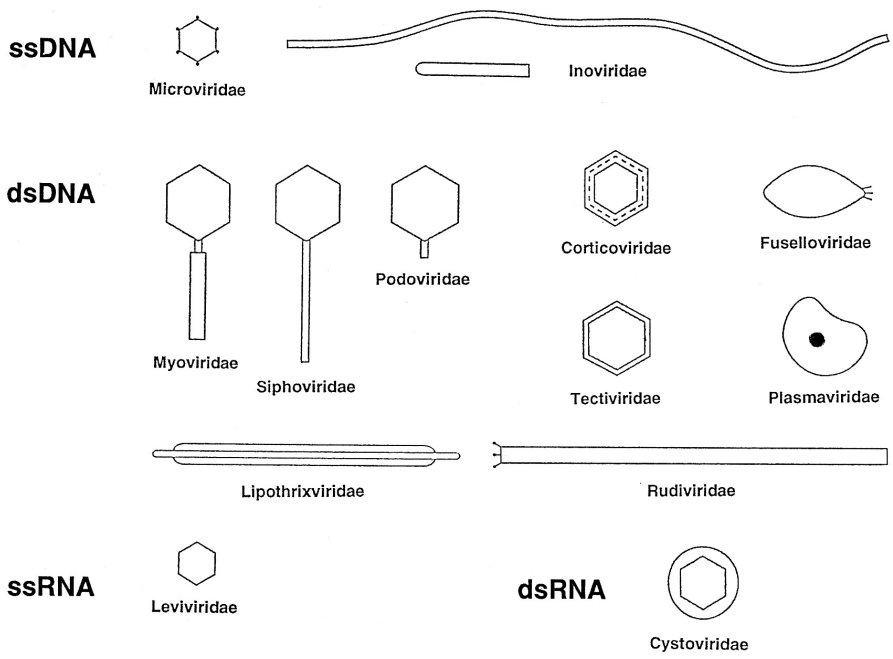


FIG. 1.8 – Principales morphologies des bactériophages.

chaîne d'acides nucléiques peut varier : on trouve des phages à simple brin ou double brin, d'ADN ou d'ARN. La capside peut être de diverses formes : on connaît des phages de formes cubiques, icosaédriques, isométriques, allongées (Weinbauer, 2004). Dans 96% des cas connus (Brussow and Hendrix, 2002), la capside comprend une queue, qui va être utilisée lors de l'attachement du phage à la bactérie et l'injection de son matériel génétique dans la cellule hôte. La taxonomie des phages se base d'ailleurs pour les classer sur des critères morphologiques (Fig. 1.8). La taille des phages est variable selon les espèces, avec des capside de moins de 30 nm et d'autres de plus de 100 nm. La taille de la capside est très corrélée à la quantité de matériel génomique (De Paepe and Taddei, 2006), et une hypothèse est que cette taille est ajustée pour être la plus petite possible, tout en étant capable de contenir le matériel génétique du phage et de résister à la pression créée à l'intérieur par la répulsion des bases chargées de l'ADN.

Le matériel génétique des phages est extrêmement condensé, le plus petit génome de phage mesurant moins de 3500 bases (Groeneveld et al., 1996). De fait, la très grande majorité des phages ne possède pas de gènes codant pour les protéines du système de traduction, ni de celui de réplication. Les phages ne peuvent donc pas se répliquer seuls : ils doivent pénétrer une cellule bactérienne et détourner sa machinerie de son usage normal pour être transcrits, traduits et se répliquer. Il n'existe pas de règle absolue concernant les hôtes de phages : certains ont un seul hôte très spécifique, d'autres peuvent infecter un large spectre d'hôtes. Ce fonctionnement obligatoirement parasitique, ainsi que la faible taille des virus les plus courts posent le problème de savoir si les phages correspondent ou non à la définition du vivant : peut-on considérer quelques protéines et si peu d'ADN comme "vivants" ? Cette question continue à être posée actuellement, donnant parfois lieu à d'intéressantes précautions verbales lors des séminaires et dans les articles...

Pour achever cette présentation succincte quelques valeurs numériques permettront

d'apprécier l'abondance des phages, qui n'a été que récemment reconnue. On estime à près de 10^{31} le nombre de phages sur la planète, la majorité étant localisée dans les mers et les océans (Wommack and Colwell, 2000). Une estimation amusante de ce chiffre est l'analogie suivante : tous les phages des océans mis bout à bout formeraient une chaîne de 10^7 années-lumière, et pèseraient environ autant que 75.10^6 baleines (Suttle, 2005). Ces chiffres font des phages les plus nombreux organismes sur la planète, et la biomasse la plus importante après celle des procaryotes.

1.2.2 Génomique

Les génomes de phages sont hautement optimisés. Composés d'un chromosome linéaire ou circulaire, ils contiennent peu de matériel génétique qui ne soit pas utilisé. Leur densité de gènes est au moins aussi élevée que celle des bactéries, avec plus de 90% de séquences codantes. La longueur du génome et le nombre de gènes peuvent varier : l'organisme modèle des phages lysogéniques, le phage λ de *E. coli*, contient 90 gènes pour un génome de 49 kb, tandis que le modèle des phages lytiques, T4, en contient 288, avec un génome de 170 kb. Mais ces chiffres peuvent énormément varier, allant de seulement 10 gènes pour certains coliphages à plus de 1200 chez *Mimivirus*, le dernier séquencé des virus géants d'eucaryotes (Raoult et al., 2004).

Les génomes de phages ont tendance à avoir un fort pourcentage en bases A et T, également caractéristique des espèces parasites (Rocha and Danchin, 2002). Les raisons de ce biais sont mal comprises, mais une hypothèse a été avancée : les bases A et T étant plus faciles à synthétiser par les bactéries, les phages pourraient les utiliser plus fréquemment dans leurs génomes. Une autre hypothèse pourrait être une sélection pour une plus grande facilité d'ouverture des boucles dans l'ADN lors du passage de l'ARN polymérase, durant la transcription : en effet les liaisons G-C sont énergétiquement plus coûteuses à briser que les liaisons A-T, car elles sont composées de 3 liaisons hydrogènes au lieu de 2. Cependant, cette hypothèse est mise en défaut par l'existence d'une exception notable, le cyanophage SL-2, dont le génome est composé de 2-aminoadénine au lieu d'adénine, une base qui forme trois liaisons hydrogène avec son complémentaire (Kirnos et al., 1977).

Une caractéristique importante des génomes de phages est la présence de mécanismes de régulation simples mais parfaitement adaptés. Pour certains phages, comme le phage λ, on connaît la fonction de quasiment tous les gènes présents sur le génome, ce qui a permis d'étudier en détails leur réseau transcriptionnel, ainsi que les interactions moléculaires de la régulation. Le faible nombre de gènes d'un phage ne permet pas l'emploi de beaucoup de régulateurs ; les mécanismes de régulation sont donc très épurés, ce qui en fait un excellent système modèle pour l'étude de l'organisation chromosomique.

Les gènes de phages sont généralement regroupés dans trois catégories, qui représentent l'ordre temporel dans lequel ils sont transcrits. On trouve d'abord les gènes précoces, exprimés dès l'entrée de l'ADN phagique dans son hôte, qui ont pour rôle principal de neutraliser le métabolisme cellulaire de l'hôte afin de libérer des ressources pour le phage. Ces gènes sont sous le contrôle de promoteurs strictement identiques à ceux de l'hôte, voire plus forts, permettant leur transcription par l'ARN polymérase de l'hôte dès leur arrivée dans le cytoplasme de l'hôte. Ensuite viennent les gènes intermédiaires, puis ceux dits tardifs. Ils sont exprimés de plus en plus tard par rapport au moment de l'infection. Ils sont placés sur le génome après les gènes précoces, dans l'ordre dans lequel leur

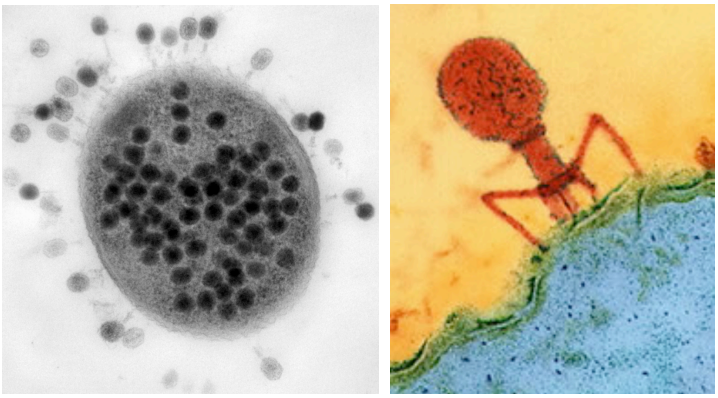


FIG. 1.9 – À gauche, plusieurs phages T4 entourant une cellule d'*E. coli*. Dans une telle situation, les phages sont très nombreux, et on observe une augmentation du temps de latence précédant la lyse après pénétration de l'hôte par l'ADN du phage : on parle alors d'inhibition de la lyse (voir texte). À droite, un phage T4 à la surface d'une cellule d'*E. coli*. Les appendices situés des deux côtés permettent la reconnaissance de récepteurs spécifiques à la surface de la cellule.

fonction est nécessaire : si un gène est régulateur d'un autre, il sera placé avant lui sur le chromosome du phage, et le fait d'être transcrit plus tôt que sa cible lui permettra d'exercer sa fonction de régulateur. Plusieurs mécanismes de régulation existent chez les phages. Par exemple la transcription de certains gènes lytiques chez λ est régulée par un mécanisme d'antiterminaison, qui permet à l'ARN polymérase d'outrepasser un terminateur de transcription situé en amont et de transcrire ces gènes uniquement quand une protéine codée par un gène précoce est exprimée. Un autre système de régulation qui a été étudié intensément est celui du choix entre cycle lytique et lysogénique, et de son maintien ; dans ce cas c'est l'équilibre précis entre un répresseur, produit par le phage, et des protéases de l'hôte, dont la production varie en fonction des conditions extérieures, qui permet le maintien de la phase lysogénique.

1.2.3 Cycle de vie

Les phages ont un cycle de vie plus ou moins complexe, selon les espèces. Certains agissent toujours de la même façon, tandis que d'autres enchaînent une phase lysogénique et une phase lytique lors de l'infection. Nous allons détailler ici les cycles de vie possibles pour les phages à ADN.

Le cycle de vie d'un phage commence toujours par l'infection de l'organisme hôte. Tout d'abord, il y a reconnaissance de récepteurs situés sur la paroi de la cellule hôte par le phage (Fig. 1.9 à droite). Divers mécanismes sont ensuite utilisés par le phage pour faire passer son matériel génétique dans la cellule hôte, le plus courant étant la perforation de la membrane de l'hôte par la queue du phage et l'extraction de l'ADN,

depuis la capside, à l'intérieur du cytoplasme hôte, phénomène sur lequel on a peu de connaissances mécaniques. Une fois le matériel génétique du phage dans le cytoplasme de l'hôte, différentes continuations ont été observées. Certains phages, n'ayant pas la possibilité de s'insérer dans le chromosome bactérien, rentrent automatiquement en cycle de vie lytique, ou en infection chronique. Pour les autres, c'est l'état de l'environnement, à la fois la physiologie de la cellule hôte, mais aussi l'environnement extérieur, qui vont déterminer le choix entre les différents cycles de vie. Ce choix entre l'entrée en phase lytique ou en phase lysogénique est régulé par un mécanisme génétique très sensible. Typiquement, l'entrée en phase lysogénique sera déclenchée si les conditions environnementales perçues par le phage ne sont pas très bonnes. Par exemple, sur un hôte isolé, ou en carence d'acides aminés, un phage n'entrera pas en cycle lytique : ceci a été contre-sélectionné par l'évolution, les virions produits dans cette situation ayant peu de possibilités d'infection par la suite. Chez le phage T4, un autre phénomène de régulation du cycle de vie a été observé, durant lequel la lyse de l'hôte est retardée quand de nombreux phages sont présents autour (Fig. 1.9 à gauche). Cette régulation donne un avantage au phage infectant la cellule, en lui permettant de profiter au mieux des ressources de son hôte (Paddison et al., 1998).

Nous allons maintenant présenter les principales étapes des différents cycles de vie des bactériophages (Fig. 1.10). L'étape d'infection chronique est peu connue : il s'agit peut-être d'un intermédiaire évolutif entre les cycles lytiques et lysogéniques. Cet intermédiaire aurait pu se développer à partir du cycle lytique, et aurait été favorisé car laissant en vie l'hôte ; ceci aurait plus tard donné naissance au cycle de vie lysogénique (Bouma and Lenski, 1988).

Cycle de vie lytique Ce cycle de vie est celui qui conduit à la lyse de la cellule hôte, qui n'est utilisée que comme ressource par le phage. Les stratégies sont multiples, mais ont toutes le même effet : dans un premier temps, les proteínes du phage vont forcer la transcription préférentielle de ses gènes, par rapport à ceux de la cellule hôte. Pour cela, le phage peut neutraliser la machinerie cellulaire de son hôte en dégradant son ADN ; produire ses propres ARN polymérases ; ou encore produire des enzymes qui vont permettre à l'ARN polymérase de l'hôte de reconnaître préférentiellement ses séquences génétiques et de les transcrire. Cette période est très rapide, et on observe un arrêt de toute production cellulaire de l'hôte en moins d'une minute (Miller et al., 2003). C'est la phase de transcription des gènes précoces. Ensuite, grâce à sa propre ADN polymérase – certains phages codent leur propre polymérase (Knopf, 1998) – ou à celle de son hôte, le génome viral va être répliqué de nombreuses fois. En parallèle, les gènes dits tardifs vont s'exprimer, synthétisant les protéines de la capside du phage. Ces protéines ont la propriété de pouvoir s'assembler naturellement en une structure de capside, parfois avec l'aide d'enzymes sécrétées par le phage. Les ADN viraux vont être ensuite aspirés par un moteur moléculaire dans les capsides (Fuller et al., 2007; Hendrix, 1998), donnant naissance à de nouveaux virions semblables à celui qui avait infecté la cellule au départ. Finalement, beaucoup de phages possèdent des gènes codant pour des enzymes qui vont lyser partiellement la cellule hôte, permettant une sortie plus aisée des virions vers l'extérieur. En l'absence de ce mécanisme, quand la concentration de virions dans la cellule devient trop élevée, celle-ci va éclater et les libérer, leur permettant d'infecter à leur tour d'autres cellules, terminant le cycle.

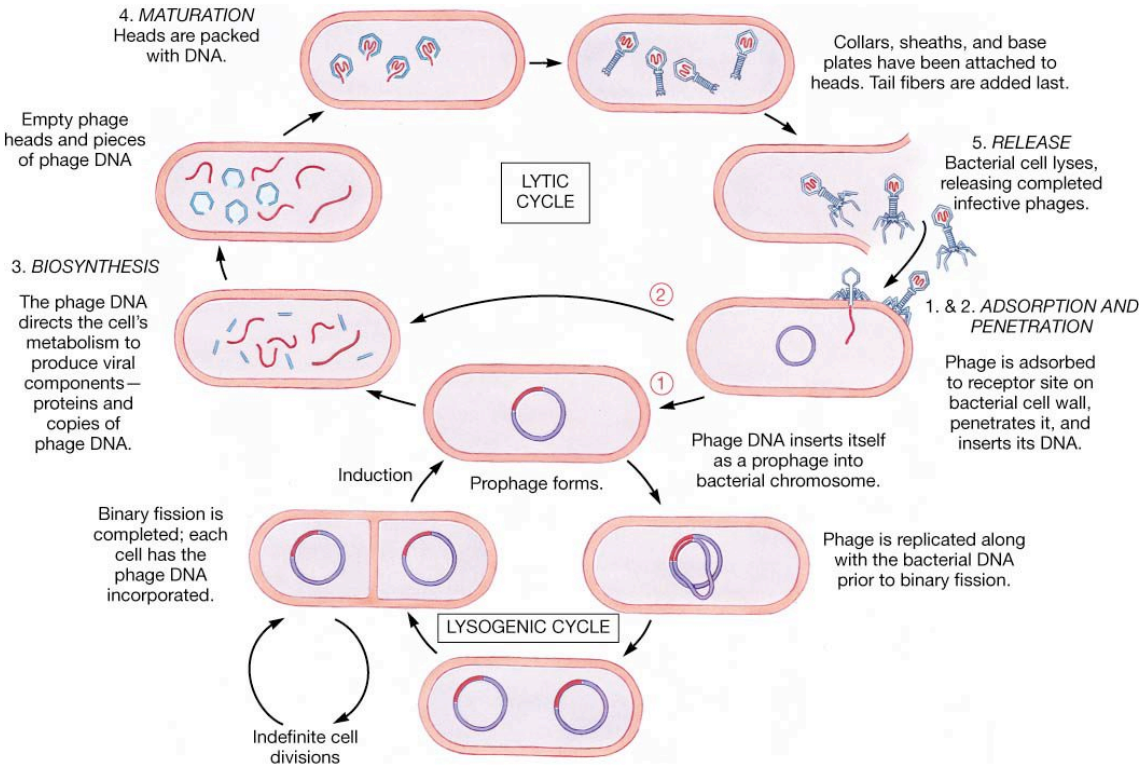


FIG. 1.10 – Cycles de vie lytique et lysogénique d'un phage.

Cycle lysogénique Cet état est accessible aux phages à ADN double brin possédant une intégrase, et aux rétrovirus à ARN capables de rétro-transcrire leur matériel génétique sous la forme d'ADN et de s'insérer dans le geńome hôte[1]. Dans le cas d'un phage à ADN, le matériel génétique du phage est inséré dans le génome hôte grâce à l'intégrase, une enzyme qui coupe l'ADN et insère au niveau de la coupure la séquence d'ADN du phage. L'insertion de la séquence phagique a lieu à certaines séquences palindromiques[2] bien particulières, reconnues par l'intégrase, et a souvent lieu à l'intérieur d'une séquence d'ARN de transfert (Campbell, 1992, 2002), à cause des structures particulières de ces gènes. De plus, les phages lysogéniques, que l'on nomme aussi tempérés, ont la propriété de compléter la séquence dans laquelle ils s'insèrent : la partie codante du gène de l'hôte située au-delà du point d'insertion est souvent similaire à l'extrémité 5' du phage, qui contient le motif permettant la recombinaison (Canchaya et al., 2002). Ainsi, même s'ils s'insèrent dans un gène essentiel, les phages lysogéniques ne vont pas immédiatement tuer leur hôte, puisqu'une une partie de la séquence dans laquelle ils s'insèrent est dupliquée. Ceci permet de conserver intacte la séquence dans laquelle le phage s'est inséré, la portion de séquence déplacée créant un pseudogène en 3'.

La lysogénie ne conduit pas directement à la destruction de l'hôte. Au contraire, aussi longtemps que le phage est inséré dans le génome – un état nommé prophage –, seule une

[1] À l'heure actuelle, les seuls rétrovirus connus sont des virus d'eucaryotes ; mais la découverte de rétrophages est tout à fait envisageable.

[2] Une séquence palindromique d'ADN est une séquence qui peut être lue dans les deux sens, à la complémentarité près. Donc ATG|GTA et ATG|CAT sont deux séquences palindromiques, puisque dans le deuxième cas, si on change de brin au niveau du |, on lit bien ATG—GTA.

fraction de son matériel génétique va s'exprimer, codant pour des récepteurs de surface qui vont immuniser l'hôte contre d'autres attaques par des phages de la même espèce. L'intérêt de cet état pour le phage réside dans le fait qu'il va être transporté et répliqué par l'hôte comme son propre ADN, et va échapper à une situation environnementale potentiellement difficile au moment de l'infection. Les bénéfices pour l'hôte sont l'immunité aux autres phages, parfois beaucoup plus : comme vu précédemment, des bactéries pathogènes s'avèrent parfois être des souches contenant un prophage possédant des gènes de virulence, exprimés de façon à ne pas attaquer leur hôte (Brussow et al., 2004). Deux exemples particuliers sont la toxine du choléra, exprimée par *V. cholerae*, mais dont le gène est en réalité porté par un prophage inséré dans son génome, et l'expression des toxines "Shiga-like" par *E. coli* O157, toxines elles aussi codées par un gène provenant d'une section prophagique du génome. On peut donc observer temporairement une forme de mutualisme entre hôte et phage. Cependant, à l'échelle de la population, les bénéfices dus à la présence d'une séquence prophagique n'ont d'intérêt que si la séquence insérée subit une conversion lysogénique et ne peut plus être induite, c'est à dire repasser en phase lytique (Mira et al., 2001). En effet, l'affaiblissement de l'hôte peut induire le passage en phase lytique : l'ADN prophagique est excisé du chromosome bactérien, et le phage rentre en cycle de vie lytique. Dans ce cas le bénéfice pour l'hôte dû à la présence du prophage est largement annulé suite au coût évolutif de la perte d'une partie de sa descendance. Les conditions nécessaires à cette induction sont partiellement connues, et impliquent la rupture de l'équilibre chimique entre la forme dimère et la forme monomère du répresseur qui maintient la phase lysogénique. Un des facteurs pouvant provoquer cette rupture d'équilibre est l'exposition aux rayons UV, qui va favoriser la fracture des formes dimères.

Pseudolysogénie L'état de pseudolysogénie est une forme de lysogénie particulière, caractérisés par la présence du génome du phage dans l'hôte sous forme de plasmide. Une hypothèse est qu'il s'agirait d'un état de non-choix entre cycle lytique et cycle lysogénique peu observé car trop instable. La copie du plasmide et sa transmission à la descendance de l'hôte dépendent du phage étudié : dans certains cas elle est incertaine, et cet état ne permet pas au phage d'assurer la production de virions (Weinbauer, 2004) ; en contrepartie il a été observé que la protection de l'hôte par immunité conférée est partielle (Jones et al., 1962). Dans d'autres, comme le phage P1 (Sternberg and Hoess, 1983), des mécanismes moléculaires particuliers permettent au plasmide d'être transmis en permanence à la descendance de l'hôte : les couples de protéines toxines-antitoxines sont un exemple parmi d'autres (Jensen and Gerdes, 1995). La pseudolysogénie a été observée avec des propriétés très diverses, laissant supposer qu'il s'agit en fait de plusieurs états différents regroupés sous le même nom.

Infection chronique L'infection chronique commence comme la phase lytique, mais au lieu de conduire à la lyse de la cellule hôte, les virions sont constamment exportés vers le milieu extérieur et continuent à être produits dans l'hôte. L'hôte est ici utilisé sans être détruit. Cet état n'a jusqu'à présent été que peu observé chez les procaryotes. Il est possible que l'équilibre nécessaire au maintien d'une telle situation ne puisse se maintenir suffisamment longtemps pour que des observations répétées aient lieu.

1.2.4 Évolution

Il est connu que les génomes de phages sont soumis à une sélection très forte et donc à une évolution très rapide. En effet, les mécanismes de défense bactériens existent, par exemple l'évolution rapide des récepteurs à la surface des bactéries, qui peut permettre aux bactéries d'échapper à certains phages. Ceux-ci doivent donc évoluer au moins aussi vite pour ne pas être privés d'hôte : on parle de Reine Rouge, en référence à "De l'autre côté du miroir", où Alice et la reine rouge doivent courir toutes les deux pour rester côte à côte sans bouger (van Valen, 1973). Cette pression de sélection, combinée à la grande population des phages et à leurs nombreuses opportunités de recombinaison, est un des facteurs qui font évoluer les phages extrêmement vite. Leur fort taux de reproduction démultiplie également les possibilités évolutives explorées. Finalement, le fait d'avoir plusieurs hôtes potentiels, et donc d'avoir leur matériel génétique traité par différentes molécules (par exemple, plusieurs polymérases différentes), réduit l'adaptation du phage à un système de réplication particulier, et augmente les probabilités d'erreurs, donc le taux de mutation. Le groupement de ces facteurs fait des phages des fusées évolutives, gagnant en permancence de petits avantages qui leur permettent de surpasser leurs concurrents.

L'évolution des phages a dans les dernières années été secouée par de nombreuses découvertes, qui ont relancé les hypothèses sur l'origine des virus, et par là même sur l'origine de la vie. Une des découvertes les plus importantes fut celle des "morons", séquences d'ADN présentes sur certains phages et pas sur leurs homologues proches, et qui a permis de confirmer l'hypothèse de mosaïcisme (Hendrix, 2002, 2003; Hendrix et al., 2000; Juhala et al., 2000). C'est une extension de l'hypothèse d'évolution modulaire, qui avait été développée dans les années 70. L'évolution modulaire était basée sur l'observation que les gènes de phage remplissent une gamme de fonction très restreinte, 11 chez les phages lambdoïdes, et supposait que l'évolution de certains phages pouvait avoir lieu par recombinaison de modules entiers entre séquences de phages de la même famille. Dans ce cas les recombinaisons se faisaient entre séquences homologues, situées entre les gènes. L'idée était qu'un phage pouvait être construit à partir de n'importe quel ensemble de modules, tant que toutes les fonctions étaient représentées. De nombreuses observations confirmant cette théorie ont eu lieu, voir Hendrix et al. (2000) pour une revue.

Le mosaïcisme pousse ce raisonnement plus loin, et suppose qu'en plus des recombinaisons homologues, des recombinaisons non homologues doivent arriver fréquemment dans les phages. Ces recombinaisons peuvent entraîner des disruptions de séquences, et donc la mort du phage, mais certaines d'entre elles peuvent conduire à un brassage génétique favorable pour le phage, voire à l'acquisition de nouvelles fonctions. Ce brassage aurait pour conséquence la disparité des séquences observées aujourd'hui au niveau des génomes de phages. Sachant qu'il a été estimé, de plus, que ces phages réalisent 10^{25} infections par seconde à l'échelle mondiale (Pedulla et al., 2003), et que, à l'intérieur de chaque hôte, il existe une micro-population de phages durant chaque événement d'infection, les opportunités de transfert horizontal et de recombinaison sont donc très grandes pour les phages, ce qui expliquerait le brassage génétique de leurs génomes.

Il est assez naturel, au vu du niveau d'évolution des séquences phagiques, de les imaginer comme des innovateurs génétiques. Les rares séquences non fonctionnelles chez les phages étant soumises au même niveau de recombinaison et de mutation, voire plus, que les séquences fonctionnelles, l'hypothèse a été émise que ces séquences pouvaient être le berceau de nouvelles séquences génétiques (Hendrix, 2002). Dans une version un peu

différente, il a été suggéré que les séquences non fonctionnelles issues d'anciens prophages dégénérés pouvaient également jouer ce rôle, donnant encore aux phages la paternité de nouveaux gènes, même chez les procaryotes (Daubin and Ochman, 2004).

Le renouveau de l'étude de l'évolution des phages, guidé par le séquençage de nouveaux organismes, a donné lieu à de nombreuses théories expliquant l'origine des phages. Nous passons brièvement en revue les plus récentes :

- Forterre (2006) a développé un modèle dans lequel trois cellules sont les trois ancêtres communs des domaines du vivant. Ces cellules, originaires d'un monde à ARN, verraient leur matériel génétique devenir de l'ADN grâce à des phages, qui auraient changé de matériel génétique pour mieux résister à la chaleur et avoir des génomes plus stables.
- Hendrix et al. (2000), avec une version modifiée d'une hypothèse classique d'échappement – dans lesquelles les virus naissent à partir de portions de cellules autosuffisantes –, ont proposé que les génomes de phages aient pu se construire entièrement par accrétion de "morons", de façon combinée à la création d'une protéine autoassemblante de la capside par les cellules environnantes. Dans ce modèle les phages, avant d'être des parasites, sont des vecteurs de gènes pour les autres cellules.
- Koonin et al. (2006) ont proposé un modèle dans lequel les phages ont une origine pré-cellulaire, en tant que séquences d'ARN sans membrane en permanente interaction. Cette hypothèse expliquerait la présence de gènes essentiels retrouvés dans presque tous les génomes viraux, tout en gardant une origine non monophylétique, ce qui serait contredit par le niveau de brassage génétique des phages.
- Claverie (2006) a offert un modèle de réduction génomique, disant que les phages sont des parasites intracellulaires qui ont subi une brutale réduction génétique. Ces arguments sont basés sur la similarité entre le métabolisme des parasites intracellulaires et celui des virus durant la phase d'infection, et la découverte récente de génomes viraux très grands, comprenant de nombreuses protéines du système de traduction (Raoult et al., 2004). Mais d'autres travaux ont montré que le contenu génétique de ces virus pouvait provenir d'acquisitions massives et non pas d'une réduction, laissant la question en suspens (voir Filee et al. (2007) ou mon travail sur l'acquisition d'ARN de transfert par des phages, page 151).

Chapitre 2

Le système de traduction bactérien

2.1 Expression génique

Cette thèse est focalisée sur l'étude du système de traduction chez les bactéries. Cependant, il est nécessaire de comprendre dans quel cadre les processus de traduction s'insèrent pour bien appréhender leur rôle, et les contraintes auxquels ils sont soumis. En effet, si ces processus font partie des mécanismes de la synthèse protéique, ils ne sont pas seuls en jeu. Avant d'étudier en détails la traduction, nous allons donc la replacer dans le contexte plus général du fonctionnement cellulaire.

Toute cellule vivante, comme on l'a vu à la section précédente dans le cas des bactéries, contient à la fois de l'ADN, de l'ARN et des protéines. L'information génétique est au départ contenue dans l'ADN. Cette information est stockée à plusieurs niveaux : dans les séquences codantes pour des protéines ou des ARN fonctionnels ; dans la façon précise dont les gènes sont codés ; dans la structure de l'ADN et ses repliements, qui permettent l'accès ou non à certaines séquences par des protéines ; et finalement dans les modifications qui peuvent affecter ses bases, la méthylation par exemple. Lors de l'étape de transcription, un des deux brins de l'ADN est transcrit en ARN. Cette étape se fait sans perte d'information, puisque la séquence de l'ARN est le complémentaire exact de la séquence d'ADN utilisée au départ, au remplacement des thymines par des uraciles près. Ensuite, lors de la traduction, la séquence d'ARN va être traduite en une séquence d'acides aminés, un polypeptide. Lorsque le polypeptide sera replié, et aura éventuellement subi des modifications post-traductionnelles, il sera fonctionnel et on parlera de protéine mature. Ce changement de séquence, des nucléotides aux acides aminés, occasionne une perte d'information, au sens où la donnée de la séquence d'acides aminés du polypeptide synthétisé ne permet pas de retrouver intégralement la séquence d'ADN utilisée au départ. Ce point sera étudié plus en détails au chapitre 3. Cet enchaînement représente de façon générale l'expression génique, ou comment le contenu de l'ADN peut être exprimé de façon visible par des protéines. Il faut noter que toutes les séquences d'ARN ne sont pas traduites, certaines sont directement fonctionnelles : il s'agit par exemple des ARN de transfert, des ARN ribosomaux et, en général, des ARN non-codants.

Cette représentation unidirectionnelle des choses a beaucoup évolué depuis sa conception. En effet, on s'est rendu compte qu'en sus du flot unidirectionnel décrit ci dessus, de très nombreux processus viennent se greffer et réguler la transcription et la traduction. Les plus connues sont les interactions véhiculées par les facteurs de transcription. Ces protéines interagissent physiquement avec l'ADN, et permettent, de réprimer ou d'en-

clencher la transcription d'un gène, ou d'un opéron. Leur rôle est essentiel, et permet de comprendre, *in vivo*, comment l'expression d'un gène peut directement avoir des effets positifs ou négatifs sur l'expression d'un autre gène. Leur découverte, et plus particulièrement la découverte de leur très grand nombre, a donné naissance à un pan de la biologie des systèmes, dont l'objet est l'étude des réseaux transcriptionnels formés par les ensembles de gènes en interaction.

De plus, les dernières années ont donné lieu à d'autres découvertes qui modifient notre compréhension du transfert d'information biologique. En effet, l'existence des facteurs de transcription a permis de montrer que les protéines pouvaient avoir un effet sur l'expression génique. Mais les ARN restaient dans ce cadre une étape uniquement transitoire dans l'expression génique. De nouvelles classes d'ARN, de type non codants, ont été découverts ces dernières années (voir par exemple le numéro spécial de Science (n° 309) ou encore Eddy (1999)). Comme leur nom l'indique, ces ARN ne sont pas les intermédiaires entre un gène et une protéine : ils ne sont pas traduits, et ne codent pour aucune protéine. Leurs propriétés sont uniquement régulatrices : ces ARN, par exemple les microARN ou les ARN interférants, ont des rôles multiples, et peuvent s'apparier avec d'autres molécules d'ARN, codantes, activant leur traduction ou leur dégradation, ou encore les empêchant d'être traduites normalement. Ces ARN peuvent donc réprimer l'expression de certains gènes, sans que ce mécanisme n'implique de protéine. L'avantage de ce type d'action est l'économie de la traduction d'une protéine, opération, qui, on va le voir, est très coûteuse énergétiquement. Le dogme central n'est donc, à l'heure actuelle, qu'une approximation très simplifiée de la réalité cellulaire.

2.2 Les molécules impliquées dans la traduction

La traduction est le processus permettant, à partir d'un ARN messager, d'obtenir une protéine. Une grande partie des fonctions cellulaires est remplie par ces protéines : la structure de la cellule est partiellement constituée de protéines, elles servent également dans tous les processus de reconnaissance de l'environnement et de signalisation, et dans les processus de gestion de l'énergie. L'ADN polymérase, qui réplique l'ADN et est à l'origine de toute forme de reproduction, est une protéine. Seul le processus de traduction lui-même emploie en grande quantité des ARN fonctionnels. En effet son origine est extrêmement ancienne, et on la suppose antérieure au dernier ancêtre commun, LUCA (Last Universal Common Ancestor). Les protéines impliquées dans la traduction sont également très vieilles, comme le montrent des développements récents de la recherche de l'ensemble de gènes minimal nécessaire à la survie d'un organisme. Ils ont montré que, parmi les 60 gènes essentiels de ce génome minimal, presque tous codent pour des protéines impliquées dans la traduction (Koonin, 2003), et sont homologues entre tous les êtres vivants. La traduction est donc un processus général et essentiel à la vie, et énergétiquement très coûteux (environ 80% des dépenses énergétiques d'une cellule sont dues à la synthèse de nouvelles protéines). Nous allons présenter le système de traduction chez les bactéries en deux parties : tout d'abord un inventaire des différentes molécules impliquées dans le processus, puis une description de leur rôle dynamique et de leurs interactions.

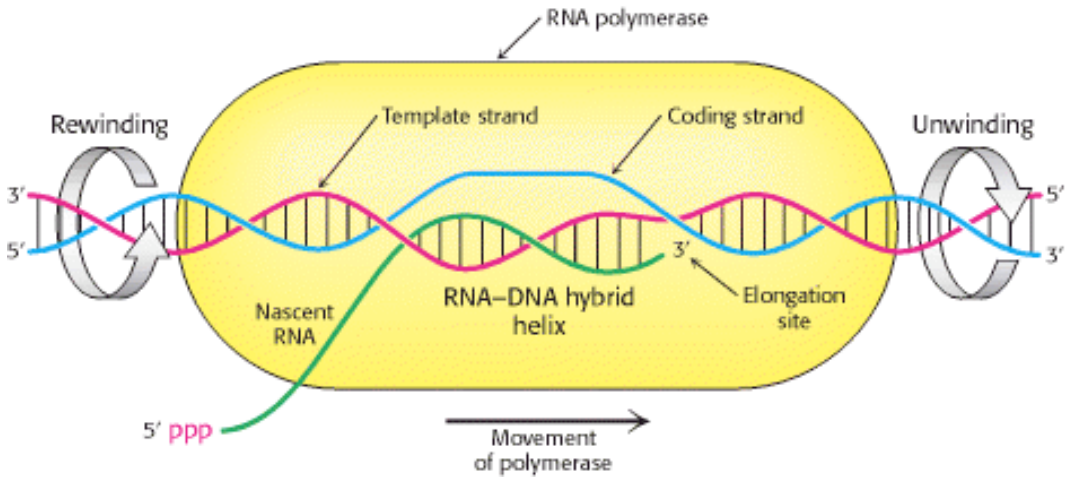


FIG. 2.1 – Bulle de transcription chez les procaryotes. On voit que la polymérase avance de 3' vers 5' sur le brin complémentaire de celui contenant la séquence codante, ce qui permet à l'ARNm d'être une copie conforme de celle-ci.

2.2.1 ARN messager et transcription

L'ARN messager (ARNm) est une molécule d'acides ribonucléiques monocaténaire, c'est à dire formée d'un seul brin. Il sert à transporter de façon intermédiaire l'information entre l'ADN[1] et les protéines. Il est formé de 4 bases différentes, trois similaires à celles trouvées sur l'ADN (adénine A, cytosine C, guanine G) et une différente : l'uracile U, qui remplace la thymine T.

L'ARNm est produit à partir de l'ADN double brin lors de l'étape de transcription. Cette phase commence par la reconnaissance de séquences promotrices, situées en amont du gène transcrit, par une ARN polymérase. Éventuellement grâce à d'autres protéines (les facteurs de transcription), la polymérase va se fixer un peu en amont de la séquence codante du gène, au niveau d'une séquence promotrice conservée. Là, son activité d'hélicase va lui permettre de désapparier les deux brins d'ADN l'un de l'autre localement. La polymérase va se fixer sur le brin complémentaire au brin codant, et avancer dans la direction 3' vers 5'. Ce mouvement entraîne une lecture de la séquence dite codante dans le sens 5' vers 3' (voir Fig. 2.1). A chaque nouvelle base parcourue, l'ARN polymérase va ajouter à l'ARNm qu'elle construit l'acide ribonucléique complémentaire de celui qu'elle lit, synthétisant ainsi l'ARNm dans le sens 5' vers 3'. Les règles de complémentarité sont les mêmes que pour l'ADN, avec U remplaçant T et se liant à A. Ainsi, la polymérase créant une chaîne complémentaire au brin complémentaire de la séquence transcrite, l'ARNm a finalement la même séquence que le brin transcrit d'ADN, à l'exception que les thymines ont été remplacées par des uraciles. La transcription se termine quand la polymérase atteint un signal terminateur. Il est constitué d'une structure tige-boucle dans la structure secondaire de l'ARNm synthétisé, souvent riche en GC, suivie d'une queue poly-U. La structure tige-boucle a pour effet de ralentir l'ARN polymérase, tandis que la queue

[1]Pourquoi est ce l'ADN qui contient l'information génétique et pas l'ARN ? Diverses hypothèses ont été avancées, la plus consensuelle étant que l'ADN est thermodynamiquement plus stable. Mais l'hypothèse selon laquelle, à l'origine de la vie, l'ARN aurait contenu l'information génétique, est très soutenue à l'heure actuelle. C'est l'hypothèse du "monde ARN".

poly-U affaiblit les interactions entre la polymérase, l'ADN et l'ARN. On connaît deux types de terminateurs : ceux qui sont *rho*-dépendant et ceux qui sont *rho*-indépendant. Dans le premier cas il y a en plus présence d'une portion de séquence riche en C et pauvre en G dans l'ARN synthétisé avant le terminateur. La protéine *rho* va s'apparier à cette séquence, et remonter le long de l'ARN dans le sens 5' vers 3' jusqu'à rentrer en contact avec la polymérase qui est bloquée au site terminateur, et briser le complexe polymérase-ARN-ADN, provoquant la fin de la transcription. Dans le second, la présence de la protéine *rho* n'est pas nécessaire, et la dissociation de la polymérase et de l'ADN a lieu naturellement (Carafa et al., 1990).

L'ARNm est une copie non seulement du gène transcrit, mais aussi d'une partie de la région en amont et en aval de la séquence codante à proprement parler. Cela permet à l'ARNm de contenir une portion de séquences régulatrices, qui vont être utilisées par exemple lors de l'amorçage de la traduction. De plus, l'ARNm peut contenir une ou plusieurs séquences codantes à la suite, et on le qualifiera respectivement de monocistronique ou de polycistronique. Dans le second cas, il s'agit d'un ARNm qui a été produit par une polymérase traduisant un opéron complet, soit un enchaînement de plusieurs gènes consécutifs très proches sur l'ADN, sous le contrôle du même promoteur. Cette configuration, on l'a vu, est très courante chez les bactéries. Elle permet à l'ARN polymérase de transcrire plusieurs gènes à la suite les uns des autres sans se détacher du chromosome.

L'ARNm est une molécule qui, au contraire des protéines, est très instable. On pense qu'il peut commencer à se dégrader au bout d'une à deux minutes environ, soit moins de temps qu'il n'en faut à l'ARN polymérase pour transcrire complètement un long gène ou un opéron (la vitesse de transcription chez *E. coli* est de 40 nucléotides par seconde, soit 20 à 30 secondes pour transcrire un gène de longueur moyenne). La dégradation commence par le coté 5' de l'ARNm, soit la partie qui a été créée le plus tôt, et implique plusieurs enzymes, des ribonucléases. Grâce à l'absence de noyau autour de l'ADN, et donc au libre accès à l'ARNm par les protéines cytoplasmiques, on observe un couplage entre transcription et traduction chez les procaryotes : la traduction de l'ARNm commence avant même qu'il ne soit complètement synthétisé, ou que la polymérase se soit détachée de l'ADN. La quantité d'ARNm dans une cellule est très variable, en fonction du taux de croissance de la cellule et de l'activité transcriptionnelle, mais toujours faible, dû à sa rapide dégradation. En général, un gène est transcrit un grand nombre de fois au cours d'un cycle cellulaire, ce qui permet d'expliquer le taux d'erreur de 10^{-4} de l'ARN polymérase : une erreur au niveau nucléotidique en moyenne, sur un gène de 1 kb, pourrait être grave si une seule copie de l'ARNm était synthétisée. Mais le nombre de copies d'ARNm est grand, diminuant l'importance de chacun. De plus, à cause des propriétés du code génétique, l'erreur a de très fortes chances de simplement changer un codon en un de ses synonymes, ou de très peu modifier la séquence protéique. Ce taux d'erreur est à comparer à celui de 10^{-9} atteint par la réplication de l'ADN, grâce à l'activité de relecture de l'ADN polymérase, qui vérifie le nucléotide ajouté à la séquence d'ADN : un tel taux est nécessaire au maintien de l'information génétique, car dans le cas de la réplication chaque copie de l'ADN doit contenir toute l'information, et chaque mutation est potentiellement délétère.

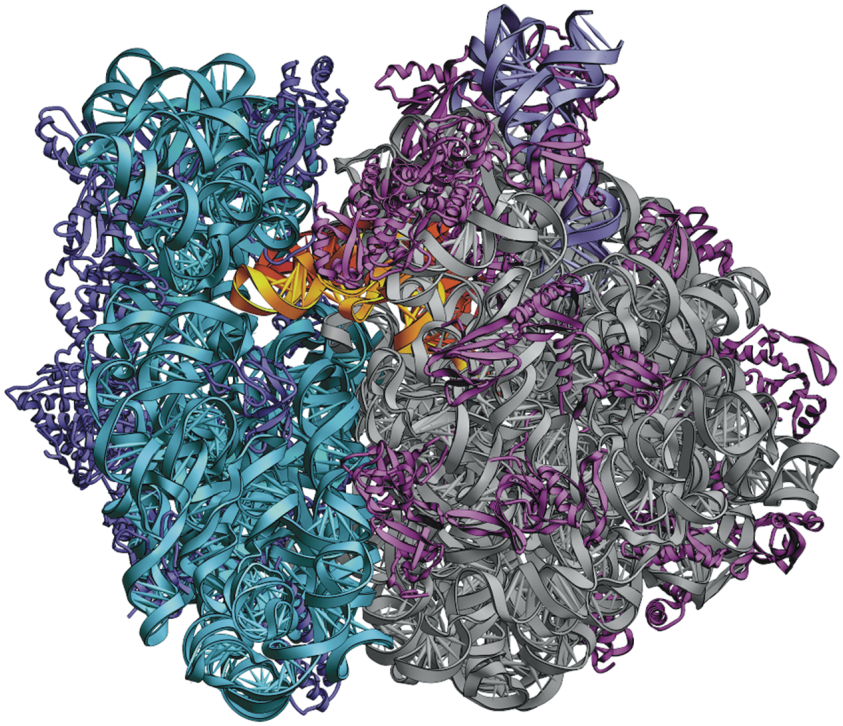


FIG. 2.2 – Structure des deux sous-unités ribosomales reliées. À gauche, la sous-unité 30S, et à droite, la sous-unité 50S. Au centre en jaune, un ARNt imbriqué dans la structure.

2.2.2 Ribosome

Le ribosome est le cœur de la traduction. C'est une ribonucléoprotéine, ensemble de molécules complexes formée de deux sous-unités. Chacune est composée d'environ deux-tiers d'ARN pour un tiers de protéines, en masse. Ceci en fait une grosse molécule, de 200 Å de diamètre environ. À eux seuls, les 20 000 ribosomes présents chez *E. coli* pèsent le quart du poids total de la cellule. Le ribosome bactérien est appelé ribosome 70S. Cette dénomination lui vient de son coefficient de sédimentation, première mesure avec laquelle il a été identifié. Les deux sous-unités sont de tailles inégales : la grande sous-unité est appelée 50S, et la petite sous-unité, 30S (Fig. 2.2). La structure exacte du ribosome est connue depuis 1968, et le ribosome est l'une des premières molécules complexes dont on ait la structure complète. La grande sous-unité ribosomale se compose de deux molécules d'ARN, les ARN 23S et 5S, et de 34 protéines, dont 32 différentes. La petite est formée d'un seul ARN, dit 16S, et de 21 protéines différentes. Les trois ARN proviennent du même transcrit, qui est clivé et modifié post-transcriptionnellement. Une remarquable propriété est que ces ARN sont truffés de courtes séquences complémentaires, et donc sont tous les trois partiellement repliés de façon extrêmement précise et conservée (Fig. 2.3). Cet agencement permet à certaines séquences, que l'on suppose fonctionnelles, d'être non appariées et accessibles depuis l'extérieur du ribosome. En particulier, trois sites fonctionnels sont présents sur le ribosome, appelés sites A, P et E. Ces sites sont des emplacements à l'intérieur du ribosome, dans lesquels les ARN de transfert vont passer successivement au cours de la traduction. Les ARNt peuvent transiter d'un site à l'autre grâce à des changements de conformation du ribosome et à des réactions biochimiques

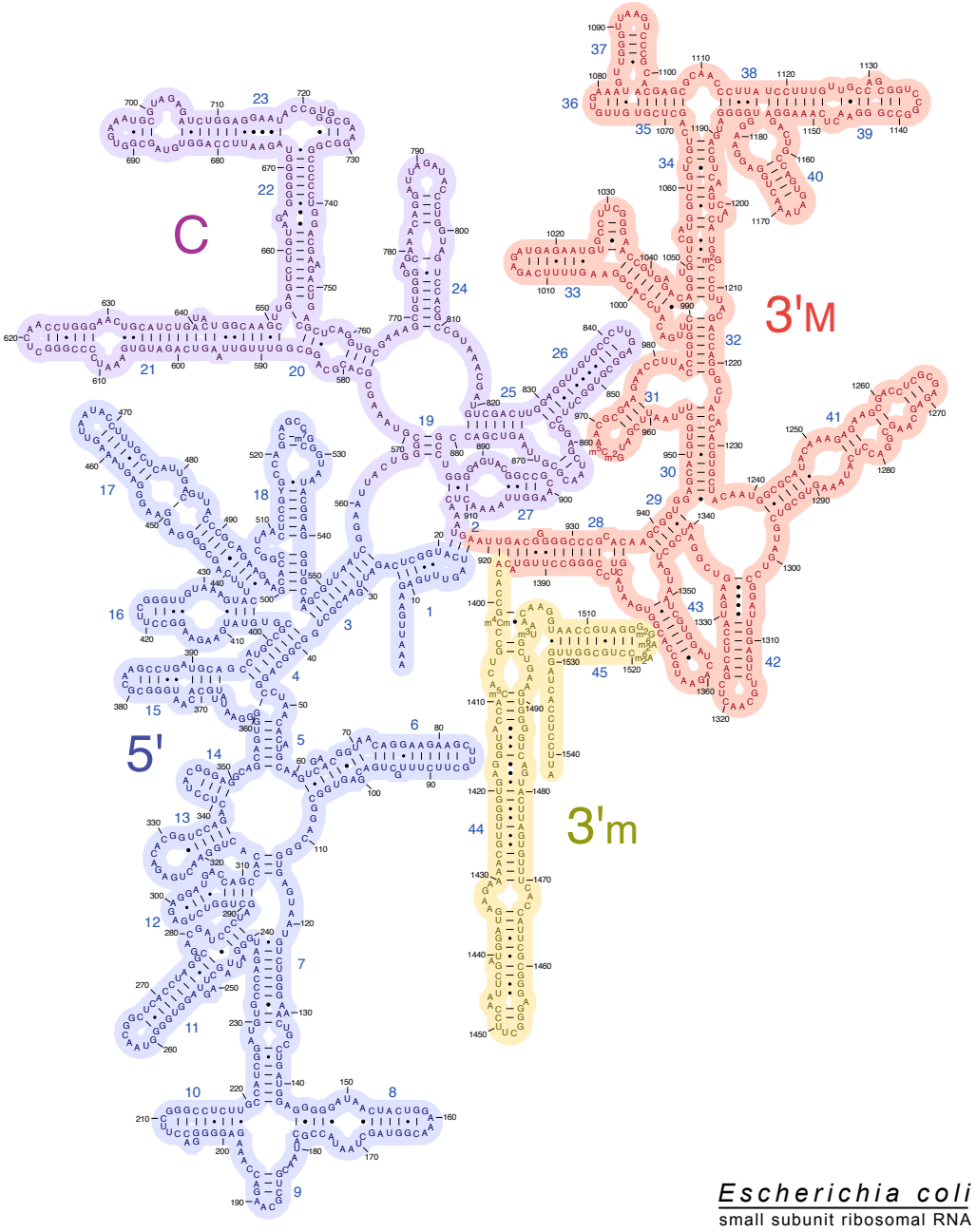


FIG. 2.3 – Schéma d'un ARN ribosomal 16S d'*E. coli*. On remarque la complexité des séquences et le grand nombre de bases appariées.

très fines, dont certaines sont provoquées par des molécules extérieures, comme EF-G ou LepA (Qin et al., 2006). Finalement, le centre actif du ribosome a une activité catalytique peptidyl-transférase, permettant de créer une liaison peptidique entre deux acides aminés.

Les gènes codant pour des protéines et ARN ribosomaux sont très souvent regroupés en opérons, dits opérons ribosomaux (Guy and Roten, 2004). Ces gènes essentiels sont souvent proches de l'origine de réplication chez la bactérie, en une ou plusieurs copies. L'intérêt d'un tel emplacement réside dans le fait que cette partie du chromosome est copiée en premier lors de la réplication, démultipliant le nombre de copies effectives que l'organisme possède de ces gènes et influant sur le dosage génique. Cet effet peut être très important dans le cas d'une croissance sur milieu riche : dans cette situation on estime que 3 fourches de réplication se suivent sur le chromosome d'*E. coli*, augmentant d'autant le nombre de copies des gènes localisés proches de l'origine de réplication. Les protéines et les ARN ribosomaux sont très conservés au cours de l'évolution, et on les retrouve chez toutes les espèces. Ceci est particulièrement vrai pour les gènes codant pour les ARN 16S, à cause des contraintes structurelles que doivent respecter ces ARN. C'est d'ailleurs sur la base de séquences d'ARN 16S qu'ont été effectuées les premières analyses phylogénétiques à grande échelle. Ces analyses peuvent en effet s'étendre aux eucaryotes, qui possèdent un ARN 18S homologue à l'ARN 16S.

2.2.3 ARN de transfert

Les ARN de transfert (ARNt) sont des acides ribonucléiques longs de 73 à 93 bases. Une de leurs particularités est que, en plus des ribonucléotides classiques (A, U, C et G), leur forme finale contient de très nombreuses bases modifiées. Les modifications peuvent être assez simples, comme une méthylation ou une diméthylation (de telles modifications sont aussi observées sur d'autres ARN), mais peuvent aller jusqu'a l'emploi de bases complètement différentes de celles trouvées classiquement dans les ARN. Un remplacement des plus courants est celui de l'adénine A par l'inosine I. En tout, plus de 60 types de modifications différentes sont recensées. Ces modifications, aussi bien les changements légers sur une base que son remplacement, sont le résultat de l'action d'une gamme d'enzymes très spécifiques. En effet, contrairement à l'ARNm, les ARNt sont modifiés de façon intense après la transcription de la séquence d'ARN primaire. En plus des modifications au niveau des bases, l'ARN produit lors de la transcription est clivé à la fois en 5' et en 3', puis lié à une séquence CCA en 3'. Finalement, à cause de sa séquence particulière, contenant de courtes parties complémentaires, l'ARNt se replie en forme dite de "feuille de trèfle", composée d'une tige et de trois boucles (Fig. 2.4 et 2.5). Alors seulement l'ARNt est fonctionnel et stable.

Dans sa forme repliée, qui est celle observée *in vivo*, la moitié environ des bases sont appariées. De plus, chez les Archaea, des bases conservées supplémentaires ont été identifiées (Mallick et al., 2005). Les bases non appariées se trouvent majoritairement dans les trois boucles, et beaucoup d'entre elles sont un rôle fonctionnel au cours des interactions que l'ARNt a avec les ribosomes et les aa-ARNt synthétases. Les boucles D et TψC s'apparient essentiellement avec les aa-ARNt synthétases, et leur composition détaillée en bases permet une sélection sur les interactions possibles (Söll and RajBhandary, 1995). La troisième boucle, dite boucle de l'anticodon, contient trois bases encadrées d'une uracile et d'une purine. Ces trois bases, numérotées 34, 35 et 36 d'après leur position dans la séquence de l'ARNt, sont appelées anticodon. Elles sont la signature de l'ARNt, car

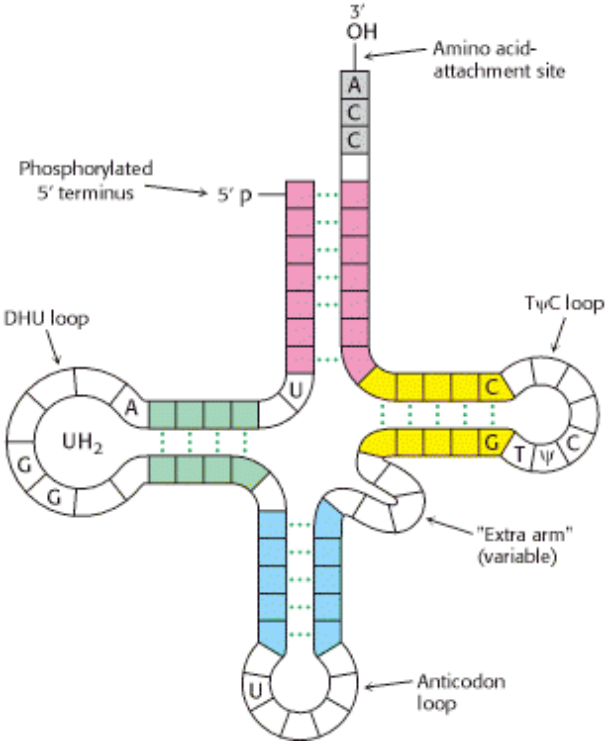


FIG. 2.4 – Schéma d'un ARN de transfert replié. On voit la structure en feuille de trèfle, ainsi que les bases conservées qui son indiquées dans la séquence.

elles représentent le ou les codons que va reconnaître l'ARNt lors de son interaction avec l'ARNm durant la traduction. La tige, quant à elle, est appelée bras accepteur. C'est à la séquence CCA qui y est liée que va s'apparier un acide aminé, formant une molécule appelée aminoacyl-ARNt, ou aa-ARNt. Cette réaction, à la base de l'existence du code génétique, est catalysée par une aa-ARNt-synthétase.

Les séquences d'ARNt, parfois notées ADNt, sont souvent regroupées sur les chromosomes bactériens. Chez les bactéries, le nombre de séquences présentes sur un génome est très variable, avec un minimum de 30 chez *Ureaplasma urealyticum*, ce qui est très proche du minimum théorique nécessaire pour reconnaître les 61 codons, et un nombre maximal supérieur à 100 (par exemple 107 ARNt détectés chez *Bacillus cereus ATCC 14579* (Reis et al., 2004)). En plus de cette variété, la redondance des gènes d'ARNt est très variable, avec certaines espèces ne possédant que des ADNt en unique exemplaire, et d'autre ayant de multiples copies de chaque ADNt. Ces ADNt sont transcrits de nombreuses fois, puisque on trouve de 100 à environ 5000 copies de chaque ARNt dans une cellule. Le nombre précis d'ARNt présent dans la cellule dépend de l'anticodon considéré, de l'espèce, des conditions environnementales et du taux de croissance de la cellule. Au total, on estime que le nombre total d'ARNt présents chez *E. coli* est compris entre 50000 et 70000, en phase de croissance sur milieu riche (Dong et al., 1996). Les variations sur la concentration d'ARNt dans la cellule jouent un grand rôle dans la régulation de la traduction, comme on le verra au chapitre suivant.

Les séquences des ADNt sont très variables, même si la structure des ARNt est conservée grâce à une sélection purificatrice intense. Une hypothèse expliquant la di-

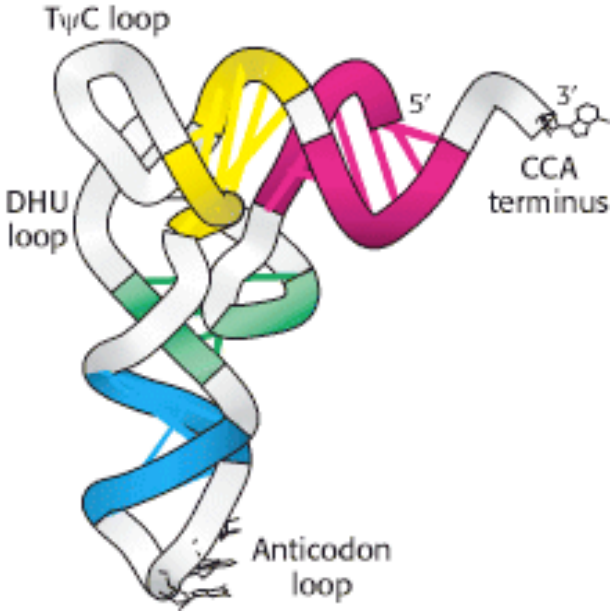


FIG. 2.5 – Structure d'un ARNt. Cette forme en L est très importante, car utilisée par d'autres molécules lors du processus de traduction pour prendre la place de l'ARNt au niveau du site A du ribosome.

versité des séquences d'ADNt est que chaque séquence doit être définie de façon à ce que l'ARNt ne puisse interagir qu'avec certaines aa-ARNt synthétases, mais les règles précises régissant ces interactions restent à définir, et dépendent de beaucoup de paramètres. D'un point de vue évolutif, ces contraintes sur la séquence et la structure rendent l'étude des ARNt très intéressante. Entre espèces proches, les ARNt sont très conservés, avec des homologies de séquences pouvant atteindre 98% (Withers et al., 2006). Mais à plus grande échelle, la diversité des séquences observées pour des ARNt ayant le même anticodon laisse à penser qu'ils n'ont pas d'ancêtre commun. Ceci a conduit à des modèles d'évolution par changement de l'anticodon, qui expliqueraient la brisure de la phylogénie à grande échelle entre ces ARNt (Saks et al., 1998). D'autres modèles, comme l'évolution d'ARNt par ligation de deux séquences courtes d'ARN (Nagaswamy and Fox, 2003), ont également été proposés dans cette optique. Mais l'étude de l'évolution des ARNt doit dans tous les cas être faite en parallèle avec celle de l'évolution des aa-ARNt synthétases et des ribosomes, leur rôle étant trop intriqué dans le système de traduction pour supposer une évolution indépendante (Agris et al., 2007). Ceci, bien sûr, permet de démultiplier les modèles possibles : voir par exemple (Taylor, 2006) ou (Ambrogelly et al., 2007) pour des hypothèses récentes, ainsi que la section sur l'évolution du code génétique page 67.

2.2.4 Aminoacyl-ARNt synthétase

Les aminacyl-ARNt synthétases (ou aa-ARNt synthétases, ou encore synthétases) sont des protéines possédant trois sites fonctionnels : un site de reconnaissance de l'anticodon, un site catalytique et un site de relecture ("edition site" en anglais). Leur rôle est de catalyser la réaction *acide aminé + ARNt* $\rightarrow$ *aminoacyl-ARNt*, et de vérifier que l'acide aminé chargé sur un ARNt est bien celui correspondant à l'anticodon de l'ARNt. Tout d'abord

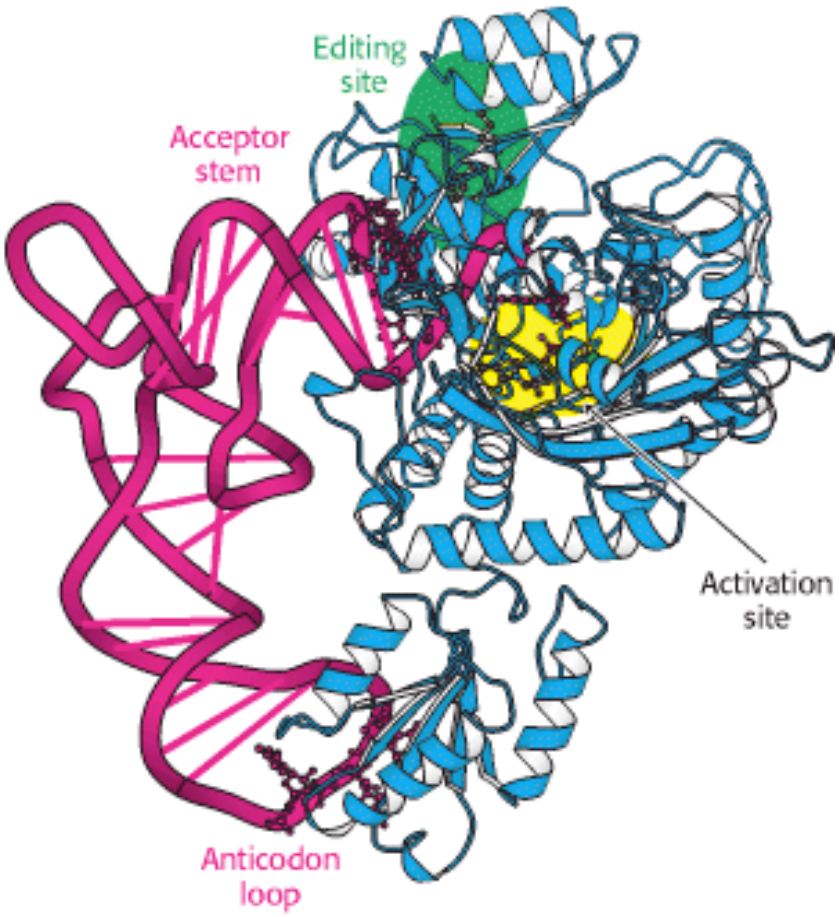


FIG. 2.6 – Synthétase appariée à un ARNt. À gauche on reconnaît la structure de l'ARNt vue précédemment. On voit le site de reconnaissance de l'anticodon à une extrémité de la synthétase, ainsi que le site catalytique et le site de relecture à l'autre.

l'acide aminé en question est reconnu par l'aa-ARNt synthétase, puis adénylé, et reste lié à elle (Fig. 2.6). Ensuite, un ARNt va se lier à la synthétase. Le site de reconnaissance de l'anticodon va s'apparier avec l'anticodon porté par l'ARNt, et, si l'appariement se fait de manière complémentaire ou presque[1], cela va déclencher un changement de conformation au niveau du site catalytique permettant le chargement de l'acide aminé par l'ARNt. Ceci permet au final de passer de trois espèces chimiques dissociées (acide aminé, ARNt, et aa-ARNt synthétase) à une aa-ARNt synthétase et un aminacyl-ARNt. À cause de contraintes stériques et réactionnelles, chaque synthétase ne peut se lier qu'avec un acide aminé particulier. Il y a donc 20 synthétases différentes dans une cellule, une correspondant à chaque acide aminé. Ensuite, la synthétase de chaque acide aminé ne peut réagir qu'avec certains ARNt (voir par exemple Rogers and Soll (1988)), particulièrement en fonction de l'anticodon qu'ils portent, et chaque ARNt n'est reconnu que par une synthétase, à quelques très rares exceptions près (Hohn et al., 2006). Cette reconnaissance spécifique à la fois d'un acide aminé et de quelques codons définit le code génétique (voir chapitre 3).

Le rôle du site de relecture est de vérifier qu'un ARNt est chargé par le bon acide

[1]Certaines synthétases peuvent reconnaître plusieurs ARNt différents, qui malgré leurs anticodons différents sont chargés par le même acide aminé.

aminé. La synthétase peut en effet se lier à un aminoacyl-ARNt formé par l'ARNt qu'elle reconnaît, grâce à la fois à l'anticodon de l'ARNt et à certaines bases sur ses boucles D et TψC. Une fois ce lien effectué, l'acide aminé chargé par l'ARNt va se trouver en face du site de relecture, dans lequel il va pouvoir s'insérer s'il correspond à celui que la synthétase en question peut charger. Dans ce cas l'aminoacyl-ARNt est relargué intact. Si l'acide aminé ne peut pas s'insérer dans le site de relecture, une erreur de chargement est détectée : l'ARNt testé n'est pas chargé avec un acide aminé correspondant à son anticodon. Dans ce cas l'aa-ARNt synthétase va hydrolyser le lien entre l'acide aminé et l'ARNt. Ce mécanisme permet de faire diminuer le taux d'erreur de chargement au niveau de l'ARNt à moins de 10^{-4}. Ce mécanisme de relecture est essentiel, et son dysfonctionnement peut avoir des conséquences graves au niveau de l'organisme (Bacher and Schimmel, 2007).

On peut classer les synthétases en deux grandes classes d'enzymes relativement similaires du point de vue de la séquence et de la structure, chacune comprenant 10 des 20 aa-ARNt synthétases. Les deux classes sont structuralement différentes, ne s'associent pas au même côté de l'ARNt et n'hydrolysent pas l'ATP avec le même mécanisme. Les aa-ARNt synthétases de classe II sont dimériques ou tétramériques, tandis que les enzymes de classe I sont le plus souvent monomériques. De plus, il a été observé que les acides aminés correspondants aux deux classes de synthétases (à savoir Ala, Asn, Asp, Gly, His, Lys, Phe, Pro, Ser, Thr pour la classe I et Arg, Cys, Gln, Glu, Ile, Leu, Met, Trp, Tyr et Val pour la classe II) ne sont pas répartis aléatoirement : les acides aminés ayant des propriétés plutôt similaires tendent à se trouver dans la même classe (Cavalcanti et al., 2004). Ceci laisse à penser que les deux classes ont peut être évolué à partir de deux molécules originelles, en même temps que les acides aminés insérés dans le code génétique et les ARNt correspondants. Finalement, malgré ce classement en deux catégories, il faut noter que les mécanismes de reconnaissance de l'anticodon et de l'acide aminé varient beaucoup dans chaque classe, et que des codons très similaires ne sont pas forcément reconnus par les mêmes interactions, ni par des synthétases appartenant à la même classe. Par exemple les codons GAC, codant pour Asp, et GAG, codant pour Glu, ne sont pas reconnus par des synthétases appartenant à la même classe.

Les gènes codant pour les synthétases sont essentiels (voir par exemple Rocha and Danchin (2003b)). Ils sont soumis à plusieurs processus de régulation très intéressants, parmi lesquels celui dit d'atténuation de la transcription (Gollnick and Babitzke, 2002). Le même processus a lieu lors de la régulation des opérons des voies biosynthétiques des acides aminés. L'idée est que, si la cellule est en manque de l'acide aminé qui doit être chargé sur la synthétase, la surexpression des gènes codant pour les synthétases en question permet d'augmenter la concentration de synthétases totale dans la cellule, et donc la concentration de synthétases chargées. Indirectement la concentration d'ARNt chargés est augmentée, à cause du déséquilibre engendré entre les réservoirs de synthétases chargées ou non. Ce processus permet donc d'éviter qu'une carence d'un acide aminé ne bloque la traduction, en autorisant la cellule à consommer au maximum les réserves d'acide aminé qui lui restent. Une régulation semblable agit sur les voies de biosynthèse de l'acide aminé en question, déclenchant sa synthèse à partir d'autres molécules. Dans ce cas le processus est relativement simple, et est résumé sur la Fig. 2.7.

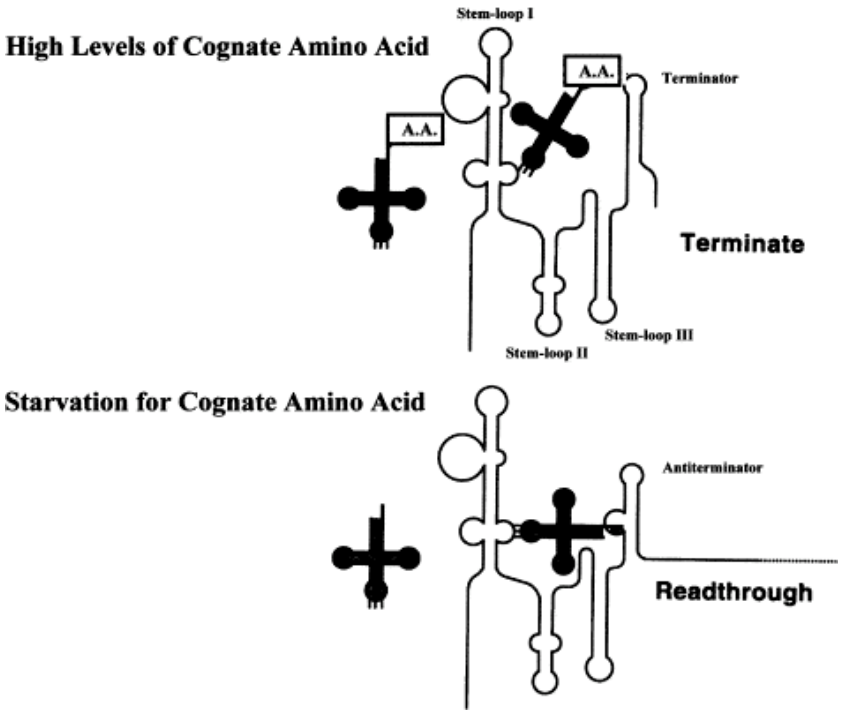


FIG. 2.7 – Atténuation de la transcription utilisant un mécanisme contrôlé par un ARNt. Voir le texte pour plus de détails.

Quand l'acide aminé est présent en grande quantité, ainsi que la synthétase, la majorité des ARNt pouvant le charger sont sous forme aminacyl-ARNt. Dans ce cas (haut de la figure), l'aa-ARNt ne peut pas interagir avec la boucle de l'ADN formant la séquence terminatrice de régulation, et la transcription ne peut pas dépasser cette boucle : le gène de l'aa-ARNt synthétase, situé en aval, n'est pas transcrit. Au contraire, quand l'acide aminé ou la synthétase sont présents en faible quantité (bas de la figure), les ARNt correspondants ne sont pas chargés, et l'ARNt peut interagir avec la boucle d'ADN, formant une séquence antiterminatrice, et autorisant la polymérase à atteindre le gène de la synthétase et à le transcrire. Ce mécanisme de régulation, entre autres, permet d'ajuster la concentration de synthétases dans la cellule en fonction des besoins exacts, et de la disponibilité en acide aminés. Les estimations sur le nombre de synthétases chez *E. coli* varient en fonction de l'acide aminé considéré et des conditions environnementales, mais sont de l'ordre de 500 par synthétase (Neidhardt et al., 1977). Cette estimation doit cependant être manipulée avec précautions, car elle précède la découverte de plusieurs des mécanismes de régulation des synthétases (Söll and RajBhandary, 1995).

2.2.5 Les acides aminés

Les acide aminés sont les éléments formateurs des protéines. Chimiquement, ce sont des chaînes carbonées caractérisés par la présence d'un groupe acide carboxylique (-COOH) et d'une amine (-NH_2), rattachés au carbone terminal, dit carbone α. Les acides aminés présents dans les protéines ne sont en fait qu'une sous classe très restreinte de cet ensemble, partageant tous la même chiralité. De plus, les protéines observées *in vivo* sont très souvent formées du même ensemble de 20 acides aminés. Récemment, il a été découvert

chez certaines bactéries des acides aminés supplémentaires, comme la sélénocystéine et la pyrrolysine, et la possibilité d'en découvrir d'autres est étudiée (Ambrogelly et al., 2007; Lobanov et al., 2006; Masashi et al., 2007). Des acides aminés non présents *in vivo* ont également été insérés dans certaines protéines par modification génétique, tout en gardant la fonctionnalité protéique (Link et al., 2003). Ceci permet d'affirmer que le choix des 20 acides aminés présents dans les protéines n'est pas nécessairement fonctionnel, et de nombreuses théories évolutives sur l'origine et l'emploi de ces acides aminés ont été proposées. Notamment, l'absence dans les protéines du vivant d'acides aminés existants en tant qu'étapes intermédiaires dans certaines voies métaboliques, comme l'ornithine ou la norleucine, peut s'expliquer par le fait que ces acides aminés se cycliseraient une fois chargés sur l'ARNt, et ne pourraient être incorporés dans les protínes. Cependant, cette explication n'est pas suffisante pour expliquer l'absence dans les protéines de certains acides aminés, comme le 2-aminobutyrate. On pourra consulter par exemple Danchin (1990, 2007), Doring et al. (2001) ou encore Sekowska (1999) et leurs références pour une discussion plus approfondie à ce sujet.

Les acides aminés ont des propriétés chimiques différentes, dépendant de leur composition et de leur structure. Ces propriétés sont celles qui, *in fine*, donneront aux protéines leurs fonctions et leur structure. Nous allons ici passer brièvement en revue les différentes propriétés des 20 acides aminés classiques (Fig. 2.8). Les acides aminés les plus simples sont l'alanine et la glycine, qui ne portent respectivement qu'un groupe méthyle ou un hydrogène sur le carbone α. Ces deux acides aminés auraient pu être synthétisés dans des conditions prébiotiques (Miller and Urey, 1959), voire être les blocs de base des premières protéines (Trifonov, 2004). Ces deux acides aminés, très simples, sont massivement employés dans les protéomes bactériens ; par exemple chez *E. coli* l'alanine représente 9.4% des acides aminés employés, et la glycine 7.3%.

En progressant sur l'échelle de la complexité, les acides aminés suivants ont un radical formé uniquement d'une chaîne carbonée. Il s'agit de la leucine, l'isoleucine et la valine, que leur chaîne carbonée rend très hydrophobes. De plus grandes chaînes carbonées conduisent à des circularisations, avec la chaîne carbonée relié à l'atome d'azote dans le cas de la proline, ou la présence de structures aromatiques dans le cas de la phénylalanine, la tyrosine et le tryptophane. La phénylalanine, comme les précédents acides aminés, est très hydrophobe, tandis que les autres sont plus hydrophiles, à cause des groupes réactifs portés sur leurs anneaux aromatiques : un groupe hydroxyle dans le cas de la tyrosine, et un groupe indole dans le cas du tryptophane. À cause, respectivement, de leur structure en anneau, et de leurs longues chaînes carbonées, la phénylalanine, la tyrosine, le tryptophane ainsi que l'isoleucine et la valine sont moins présents à l'intérieur des hélices α, des structures hélicoïdales classiques dans les protéines. À la place, ces acides aminés ont une forte propension à former des structures planaires dans les protéines, nommées feuillets β, dans lesquelles on trouve également beaucoup de proline.

Les autres acides aminés sont caractérisés par un niveau de réactivité beaucoup plus élevé. On trouve tout d'abord la méthionine et la cystéine, caractérisés par la présence d'un atome de soufre remplaçant un carbone de la chaîne, et également très hydrophobes. Les atomes de soufre de deux cystéines peuvent se coupler pour former un pont disulfure, extrêmement stable, dans une protéine. Ensuite viennent la thréonine et la sérine, avec un groupe hydroxyle très réactif, et qui à l'instar de la tyrosine et du tryptophane sont plus hydrophiles à cause de ce groupe réactif. D'autres acides aminés sont très hydrophiles, à cause d'une amide portée en bout de chaîne. L'amide est chargée positivement à pH

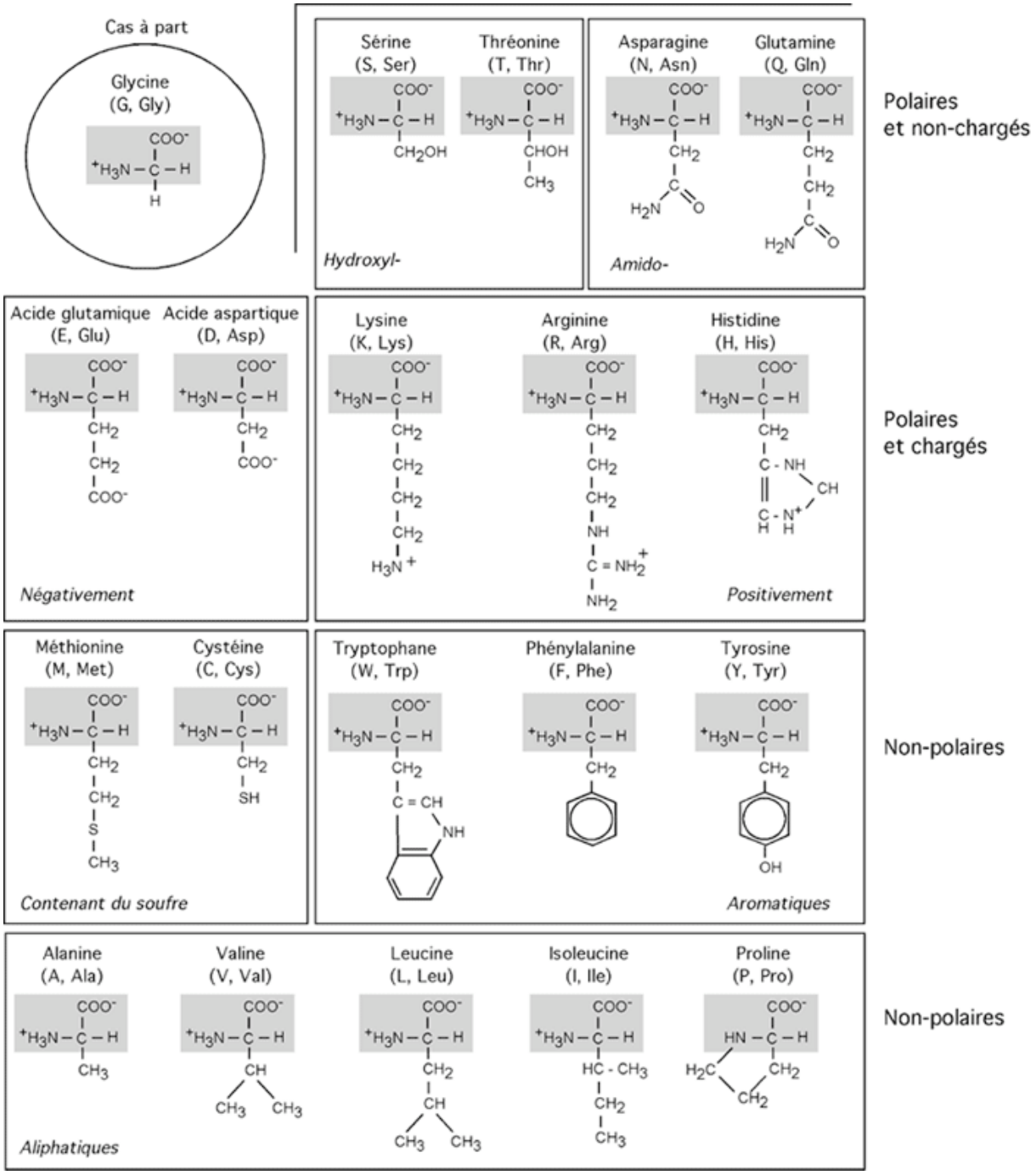


FIG. 2.8 – Tableau récapitulatif des propriétés des 20 acides aminés trouvés dans les organismes vivants.

neutre, expliquant l'hydrophilie de ces acides aminés, à savoir l'arginine, l'histidine et la lysine. Les quatre derniers acides aminés se répartissent en deux paires d'acides aminés semblables, dans lesquelles le premier porte un groupe acide carboxylique en bout de chaîne carbonée, tandis que le deuxième le remplace par un groupe carboxamide. Ces deux paires sont (acide aspartique/asparagine) et (acide glutamique/glutamine). Les deux acides sont très réactifs en solution.

Les acides aminés ont donc des propriétés réactionnelles très différentes, en fonction de leur structure. Cette structure implique également un coût : les acides aminés les plus simples, comme l'alanine et la glycine, nécessitent moins de ressources pour être synthétisés – ils ont un coût métabolique plus faible (Akashi and Gojobori, 2002; Baudouin-Cornu et al., 2001) – et donc peuvent être employés plus souvent. Ceci est un des nombreux paramètres à prendre en compte lors de l'étude de la composition protéique d'un organisme (Pascal et al., 2005, 2006) et de son évolution (Rocha, 2006), avec par exemple son adaptation à son environnement (Tekaia et al., 2002).

Les 20 acides aminés peuvent être synthétisés par *E. coli* et de nombreux autres microorganismes, mais ceci n'est pas le cas chez les eucaryotes supérieurs : par exemple, pour l'homme, 9 acides aminés sont essentiels et ne peuvent pas être synthétisés à partir d'autres molécules. La biosynthèse des acides aminés est régulée, comme celle des des synthétases, par une atténuation transcriptionnelle (Gollnick and Babitzke, 2002). Celle-ci n'implique pas les ARNt comme éléments de régulation, mais les ribosomes. Le principe est le même : les gènes codant pour des enzymes de la voie de biosynthèse d'un acide aminé sont regroupés en un opéron, précédé par une structure complexe comprenant un terminateur de transcription. La transcription est démarrée un peu avant ce site, créant un début d'ARNm que les ribosomes vont commencer à traduire. Cette séquence d'ARNm contient plusieurs codons correspondant à l'acide aminé régulé. En cas d'excès de l'acide aminé en question, les ribosomes peuvent traduire sans difficultés ces séquences et vont aller percuter l'ARN polymérase bloquée sur le terminateur : le processus s'arrête sans que les gènes de la voie de biosynthèse ne soient transcrits. Mais en cas de déficit de l'acide aminé en question, les ribosomes sont bloqués au début de l'ARNm, laissant assez de temps pour que se forme une structure antiterminatrice et que la polymérase puisse transcrire les gènes de la voie de biosynthèse.

2.2.6 ARN transfert-message

Les ARN transfert message (ARNtm) sont des acides ribonucléiques partiellement non codants, d'une longueur comprise entre 300 et 400 bases (Fig. 2.9). Leur structure, complexe, peut être divisée en deux grandes sous-parties (Haebel et al., 2004; Zwieb et al., 1999) :

- Une sous-partie ayant toutes les caractéristiques fonctionnelles du bras accepteur des ARNt, qui peut être chargée d'une molécule d'alanine comme un ARNt^{Ala} par une Ala-ARNt synthétase.
- Une sous-partie codante, similaire à un ARNm. La séquence codante peut différer selon les espèces, et n'a été vérifiée expérimentalement que chez *E. coli*, où elle code pour la chaîne d'acides aminés AANDENYALAA.

La présence des ces deux sous-parties différentes a valu son nom à l'ARNtm, identifié pour la première fois en 1978 (Lee et al., 1978) et appelé tout d'abord ARN SSrA ou ARN 10S. Une seule séquence d'ARNtm a été identifiée dans chaque espèce bactérienne,

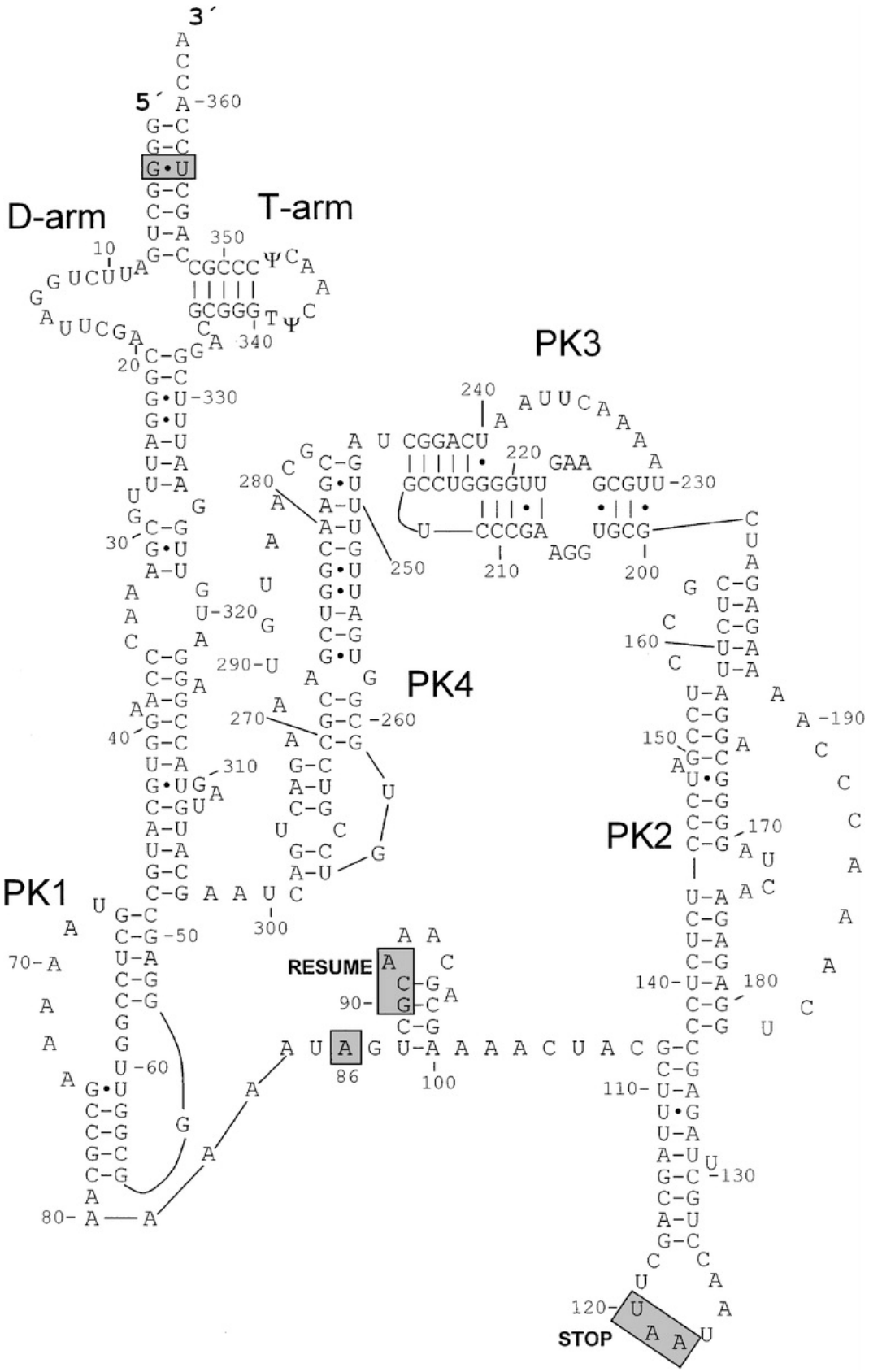


FIG. 2.9 – Séquence et structure secondaire d'un ARNtm. On distingue en haut à gauche la partie similaire à un ARNt avec deux boucles et un bras accepteur, et en bas à droite la partie codante, dont le début et la fin sont encadrés.

mais ces séquences, quoique non essentielles, se retrouvent dans toutes les bactéries, à de très rares exceptions près (par exemple *Rickettsia prowazekii*). Au contraire, on n'a pas identifié d'ARNtm chez les Archaea. Malgré la présence d'une seule séquence par génome, le nombre d'ARNtm présent dans une cellule est très grand, de l'ordre de 13000 (Altuvia et al., 1997).

Le rôle des ARNtm est la libération des ribosomes bloqués durant la traduction (Withey and Friedman, 2003). Ce rôle est intimement relié à leur structure et à leur capacité à être reconnus successivement comme ARNt puis comme ARNm par le ribosome. Leur fonctionnement est détaillé dans la section "Terminaison" de la section suivante, page 59.

2.2.7 Autres molécules impliquées

De nombreuses autres protéines sont impliquées dans le processus de traduction. Leur importance étant moindre pour notre étude, nous les mentionnons rapidement ici, pour mémoire :

- Les facteurs de d'amorçage, au nombre de 3 connus chez les bactéries, nommés IF-1, IF-2 et IF-3. Ils sont utilisés durant l'amorçage de la traduction, c'est à dire la phase de recrutement du ribosome par l'ARNm, et servent à stabiliser les interactions ribosome-ARNm.
- EF-Tu est un facteur qui se lie à l'aminoacyl-ARNt et le stabilise, avant qu'il ne soit délivré au site A du ribosome. Il y en a environ 70 000 dans une cellule, soit autant que tous les ARNt réunis.
- EF-G est un facteur d'allongement utilisé par le ribosome durant la translocation. Il y en a environ 20 000 par cellule, soit le nombre de ribosomes.
- Les facteurs de terminaison RF1, RF2 et RF3. Ce sont eux qui reconnaissent les codon stop et permettent au ribosome de se dissocier de l'ARNm quand la lecture de celui ci est terminée. Ils sont peu nombreux, environ 600 de chaque classe par cellule.
- Le facteur de recyclage des ribosmomes, qui permet la libération des sous-unités ribosomales de l'ARNm à la fin de la traduction.
- Les protéines chaperons comme GroEL ou le "trigger factor". Ces protéines servent, après la phase de synthèse peptidique, à faciliter le repliement du polypeptide en protéine fonctionnelle.
- Le GTP, carburant de la traduction. La traduction d'un ARN en peptide a un coût énergétique non négligeable : au total 4 molécules de GTP sont hydrolisées pour chaque acide aminé ajouté au peptide, soit 2 pour charger l'aminoacyl-ARNt, 1 pour délivrer l'aminoacyl-ARNt au ribosome, et 1 durant la phase de translocation. De plus, les phases d'amorçage et de terminaison de la traduction nécessitent elles aussi l'hydrolyse de molécules de GTP.

2.3 Fonctionnement dynamique du système

La traduction est l'étape durant laquelle un ARNm est lu par le ribosome, ce qui permet l'assemblage des acides aminés en une protéine. Elle est nommée ainsi car, durant cette étape, il n'y a pas simplement copie d'une séquence nucléique vers son complémentaire, mais traduction depuis un alphabet composé de triplets de 4 nucléotides différents dans un

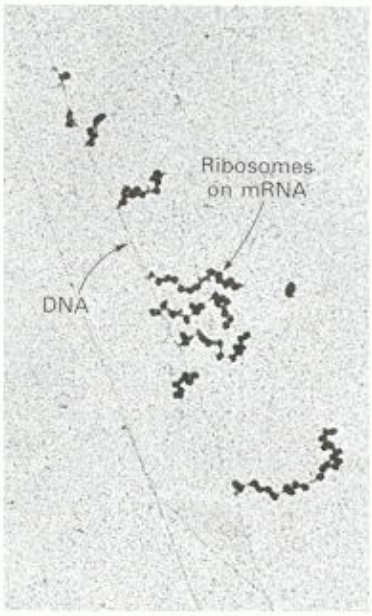


FIG. 2.10 – Polysome sur un ARN messager. On voit que l'ARN messager est encore en train d'être synthétisé au niveau de l'ADN alors que les ribosomes ont déjà commencé la traduction.

autre alphabet, dont les éléments de base sont les 20 acides aminés. C'est le code génétique, et les molécules qui l'incarnent dans la cellule, ARNt et synthétases, qui définissent le lien entre les séquences de triplets – les codons – et les acides aminés.

Presque tout le processus de traduction se passe au niveau du ribosome. Toutes les autres molécules viennent interagir de différentes façons avec le ribosome durant cette étape. Nous allons décrire le processus pour un seul ribosome, mais il est connu que, chez les bactéries, plusieurs ribosomes traduisent séquentiellement le même ARNm, les uns à la suite des autres (Fig. 2.10). De plus, chez les procaryotes, l'ARNm commence à être traduit avant même d'être complètement transcrit : pendant que l'ARN polymérase continue à synthétiser l'ARNm en allongeant son extrémité 3', le ribosome avance également sur l'ARNm dans le même sens, après s'être apparié à lui au niveau de l'extrémité 5'. C'est le couplage entre transcription et traduction (Lewis et al., 2000; Mascarenhas et al., 2001), qui est spécifique aux procaryotes, et dont on reparlera au chapitre 6 dans le cadre de l'analyse des corrélations entre gènes proches.

La traduction de l'ARNm, chez les procaryotes comme chez les eucaryotes, se décompose en 3 phases successives :

- L'amorçage est l'étape durant laquelle le ribosome et l'ARNm s'apparient, et la fabrication du peptide est débutée par la mise en place du premier acide aminé, une méthionine.
- L'allongement est une phase itérative, qui va se répéter autant de fois qu'il y a d'acides aminés à accrocher au peptide, et durant laquelle un nouvel acide aminé est lié au polypeptide naissant.
- La terminaison est la phase de relarguage du peptide finalement créé, et de dissociation du ribosome et de l'ARNm.

Nous allons détailler ces trois phases, du point de vue des procaryotes (Fig. 2.11). Chez les

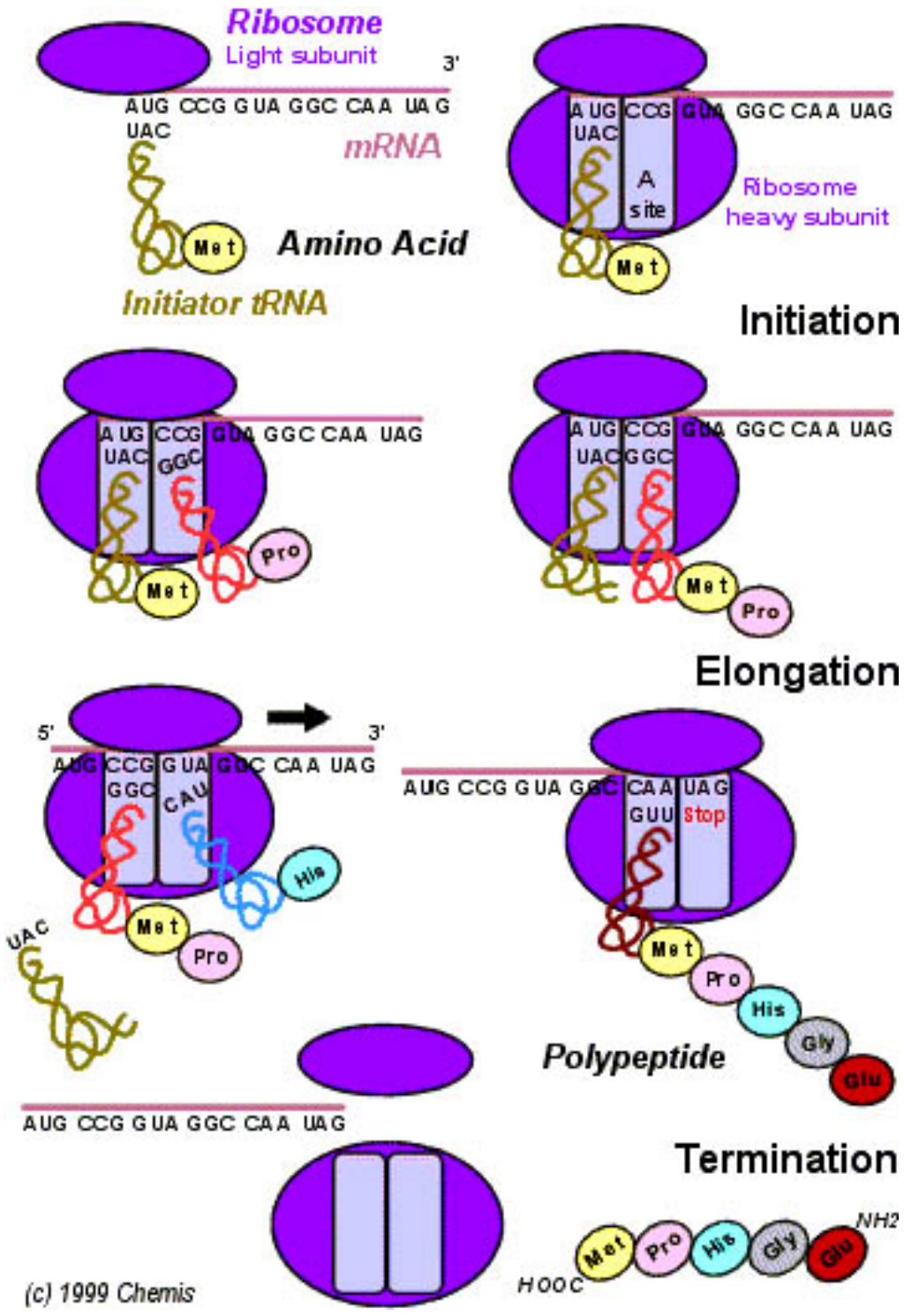


FIG. 2.11 – Schéma du fonctionnement général de la traduction chez les procaryotes. Les deux premières images représentent la phase d'amorçage, avec l'arrivée de l'ARNt amorceur et l'accrochage d'une méthionine formylée en tête du peptide. Les trois suivantes imagent la première phase d'allongement. La terminaison a lieu quand un codon stop est en face du site A.

FIG. 2.12 – Exemples de séquences de Shine-Dalgarno dans différents gènes et organismes. Les bases où s'apparie l'ARN 16S et l'ARNt amorceur sont surlignées.

eucaryotes, de très nombreux co-facteurs supplémentaires sont mis en jeu, et les processus, notamment l'amorçage[1], diffèrent quelque peu. Pour une excellente revue des processus de traduction chez les organismes eucaryotes, se référer à (Kapp and Lorsch, 2004).

2.3.1 Amorçage

L'état des lieux avant le début de la traduction est le suivant : les deux sous-unités ribosomales sont disjointes, et peuvent être considérées comme libres dans le cytoplasme. L'ARNm est encore relié à l'ADN, en train d'être synthétisé, ou bien déjà relargué par l'ARN polymérase ; dans tous les cas son extrémité 5' est libre. Les ARNt présents dans la cellule sont sous forme aminoacyl-ARNt, chargés par l'acide aminé correspondant à leur anticodon par les synthétases. Les autres sous-facteurs diffusent librement dans le cytoplasme.

Comme dit précédemment, la séquence de l'ARNm commence en amont du codon marquant le début de la séquence codante proprement dite. Ce codon est presque toujours le codon AUG, parfois GUG ou CUG (Giglione et al., 2004), selon les espèces. Cette diversité de codon d'amorçage, à elle seule, suggère que la reconnaissance du début d'une séquence ne peut pas dépendre que de ce codon. Si c'était le cas la traduction pourrait être commencée au milieu d'une séquence codante, sur la base de la seule présence de ce codon, voire se produire sur des ARN non codants, ce qui entraînerait la fabrication de résidus protéiques potentiellement toxiques pour la cellule. L'amorçage de la traduction doit donc être contrôlé de manière stricte, et être représenté par une séquence plus complexe sur l'ARNm. Cette séquence reconnue par l'ARN 16S est peu variable pour tous les ARNm

[1]Le terme amorçage sera employé dans cette thèse pour d'écrire la première phase de la traduction, bien que le terme "initiation" puisse être trouvé dans beaucoup d'ouvrages. Ce dernier est un anglicisme flagrant, qui tend malheureusement à être de plus en plus employé à l'heure actuelle. De même, "élongation" sera remplacé ici par "allongement", qui décrit beaucoup mieux le processus auquel sont soumis les polypeptides en formation. Ce vocabulaire a déjà été utilisé dans certains ouvrages français ou dans des traduction très soignées d'ouvrages anglais, comme par exemple Cooper (1999), dont on pourra se servir comme d'une référence sur la question.

d'une espèce, car elle doit être reconnue par une séquence précise sur l'ARN 16S, qui est spécifique. Elle est appelée séquence de Shine-Dalgarno et est située une dizaine de nucléotides en amont du codon AUG sur l'ARNm (Fig. 2.12).

La phase d'amorçage va permettre aux deux sous-unités du ribosome et à l'ARNm de s'assembler en une seule structure. Cette construction se fait de telle sorte que le ribosome soit bien placé au début de la séquence codante de l'ARNm (Yusupova et al., 2006). Tout d'abord, l'unité 30S du ribosome est recrutée par l'ARNm grâce aux facteurs de d'amorçage IF-1 et IF-3. Son placement exact sur l'ARNm est un élément clé de l'amorçage de la traduction chez les bactéries. En effet, la sous-unité 30S du ribosome est placée sur l'ARNm de telle sorte que l'ARN 16S qu'elle contient soit partiellement complémentaire à la séquence de Shine-Dalgarno. Les interactions entre ribosome et séquence de Shine-Dalgarno sont nécessaires à la stabilisation du complexe en formation, assurant une grande précision positionnelle lors de l'amorçage de la traduction.

Une fois la sous-unité 30S mise en place, il reste à ajouter au complexe la sous-unité 50S. Mais si celle-ci arrive trop tôt, à savoir avant qu'un aa-ARNt amorceur ne soit relié à l'ARNm, le ribosome formé ne sera pas apte à poursuivre la traduction. Un des rôles des facteurs d'amorçage est d'empêcher stériquement la liaison précoce des deux sous-unités. Avant l'arrivée de la sous-unité 50S, le complexe formé par le facteur d'amorçage IF-2, l'aminoacyl-ARNt amorceur formyl-méthionine-ARNt$_f$ chargé d'une méthionine formylée, et une molécule de GTP, vont se fixer sur l'unité 30S et l'ARNm. Grâce au placement de l'unité 30S relativement à l'ARNm, l'anticodon de l'ARNt$_f$ amorceur peut s'apparier de façon complémentaire au codon AUG (ou au codon d'amorçage de manière plus générale) de l'ARNm. L'ARNt$_f$ amorceur est un ARNt spécial, qui ne peut être chargé que par une méthionine formylée, et qui ne sert que durant cette phase précise de la traduction. Il est très conservé, même chez les eucaryotes (Marck and Grosjean, 2002).

L'achèvement de la phase d'amorçage a lieu lors de l'arrivée de la sous-unité 50S au niveau du complexe déjà existant. Ceci provoque l'hydrolyse de la molécule de GTP en GDP, et l'énergie libérée est utilisée pour libérer les facteurs d'amorçage et permettre à la sous-unité 50S de se lier à la sous-unité 30S, formant une structure qui enferme l'ARNm et l'ARNt$_f$ amorceur. Dans cette conformation, l'ARNt$_f$ amorceur occupe le site P du ribosome, tandis que les sites A et E sont vides.

2.3.2 Allongement

La phase d'allongement va permettre de faire croître la séquence peptidique commencée par la méthionine formylée. Elle se décompose en plusieurs étapes, et se déroule de manière itérative. À chaque itération, un acide aminé est rajouté à la chaîne peptidique.

La conformation du complexe ribosomal permet aux deux sites P et A d'être localisés en face de deux triplets nucléotidiques successifs sur l'ARNm. Le site P est occupé après l'amorçage, ainsi que pendant tout le processus, comme on va le voir. Le site A, lui, sert de point d'entrée aux ARNt dans le complexe ribosomal. L'allongement commence par le recrutement au site A d'un aminoacyl-ARNt à l'anticodon complémentaire du codon situé en face du site A. Des liaisons hydrogène sont alors mises en place entre le codon et l'anticodon, ainsi qu'entre le ribosome et l'aa-ARNt stabilisant la liaison codon-anticodon. Si un aa-ARNt ne portant pas le bon anticodon essaie d'entrer au site A, les liaisons hydrogène ne peuvent pas se former entre codon et anticodon, et l'aa-ARNt est rejeté par le ribosome. La fréquence d'erreur est très faible, de l'ordre de 10^{-4}. L'hypothèse a été

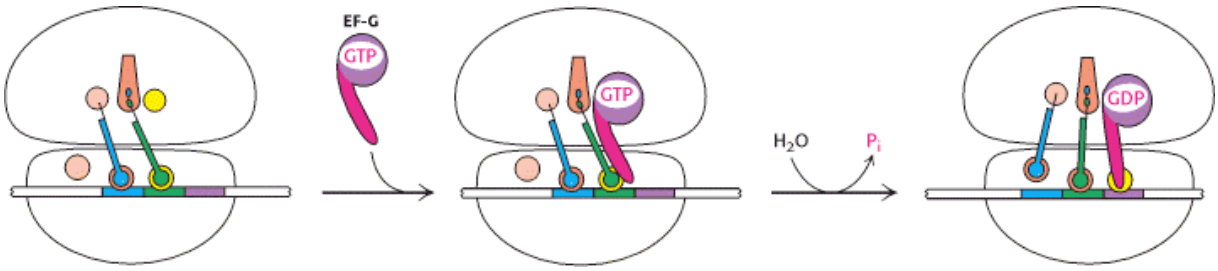


FIG. 2.13 – Schéma simplifié de l'étape de translocation. À gauche, la situation avant l'arrivée du complexe EF-G–GTP, avec un ARNt dans le site A et un dans le site P. Au milieu, EF-G prend la place de l'ARNt au site A grâce à sa structure qui imite celle d'un ARNt. À droite, l'hydrolyse du GTP permet la translocation proprement dite, avec déplacement des ARNt du site P au site E et du site A au site P.

émise que les rejets d'aa-ARNt non correspondants à l'anticodon pouvaient être le facteur dominant du temps nécessaire à la traduction de chaque codon, mais des expériences ont montré que ce n'était pas le cas (Bilgin et al., 1988). L'aa-ARNt recruté au site A arrive lié au facteur d'allongement EF-Tu et à une molécule de GTP. La liaison entre aminoacyl-ARNt, ARNm et ribosome est permise grâce à l'hydrolyse de la molécule de GTP en GDP, et par la présence de EF-Tu, sans laquelle la liaison n'est pas formée. Ce coût énergétique à la formation des liaisons entre codon et anticodon, et la nécessité de la présence de EF-Tu, sont deux facteurs qui permettent d'améliorer la précision atteinte au niveau de la reconnaissance entre codon et acide aminé.

Une fois l'aminoacyl-ARNt en place, il y a formation d'une liaison covalente entre le site amine de l'acide aminé porté par l'ARNt au site A, et l'extrémité carboxyle du peptide en formation, qui était attachée à l'ARNt présent au site P. Cette réaction est énergétiquement favorisée par le complexe ribosomal, et ne requiert pas d'énergie. En pratique, on peut la concevoir comme le transfert du peptide en formation, de l'ARNt présent au site P sur l'aa-ARNt nouvellement arrivé au site A. Ici on voit l'importance que la méthionine placée en tête dans le peptide soit formylée à sa position amine : si ce n'était pas le cas, une réaction parasite pourrait avoir lieu, l'amine de la méthionine attaquant le site carboxyle de l'aminaocyl-ARNt du site A. Le résultat des deux réactions combinées serait un produit peptidique cyclique et dissocié des deux ARNt présents dans le ribosome. Ceci préviendrait toute synthèse protéique ultérieure et serait délétère. La formylation de la méthionine en tête empêche donc cette réaction de se produire.

La phase d'allongement se termine par la translocation des deux ARNt. L'ARNt chargé de la séquence peptidique passe du site A vers le site P, repoussant l'ARNt présent au site P vers le site E. Elle s'achève par le relarguage de l'ARNt au site E dans le milieu cytoplasmique. La translocation de l'ARNt du site A vers le site P se fait sans que la liaison codon-anticodon soit brisée. Si on se place dans le référentiel du ribosome, ceci a pour effet de faire avancer l'ARNm en même temps que l'ARNt, et donc de décaler d'un codon le triplet en face du site A. Ce mécanisme permet de faire progresser la synthèse protéique, de telle sorte que chaque codon n'est traduit qu'une seule fois dans le ribosome, et que le même mécanisme peut s'appliquer à tous les codons de l'ARNm : en effet, après la translocation, la situation est la même qu'avant la phase d'allongement, avec un ARNt relié à la chaîne peptidique dans le site P, et le site A vide. La translocation

requiert le facteur EF-G et l'hydrolyse d'une molécule de GTP supplémentaire. D'un point de vue structurel, EF-G est très similaire au complexe ternaire EF-Tu-GTP-ARNt. On suppose qu'il agit en se plaçant dans le site A du ribosome, comme un ARNt, et qu'ensuite l'hydrolyse du GTP permet la translocation (voir Fig. 2.13). La similarité structurelle de EF-G et EF-Tu permet de supposer qu'ils sont associés aux mêmes sites sur le ribosome, et donc qu'ils sont mutuellement exclusifs : tout d'abord EF-Tu se lie au ribosome durant l'arrivée du nouvel aa-ARNt pour favoriser la liaison codon-anticodon, puis EF-G prend sa place durant la translocation.

2.3.3 Terminaison

La fin d'une séquence codante est marquée sur l'ARNm par des codons particuliers. Ces codons sont dits codons stop, et sont UAA, UGA ou UGG chez la très grande majorité des bactéries. Lorsqu'un tel codon est placé en face du site A, il ne peut pas s'apparier avec un ARNt possédant un anticodon complémentaire, car ceux-ci n'existent pas. À la place, une autre molécule, ressemblant structuralement à un ARNt, va venir se placer au niveau du site A. Ces protéines, nommées facteur de relarguage, portent là où devrait se trouver un acide aminé sur un ARNt, une molécule d'eau. Quand cette protéine se place au niveau du site A, la molécule d'eau qu'elle porte se trouve assez proche de la liaison entre le peptide et l'ARNt au site P pour l'hydrolyser, relâchant ainsi le peptide dans le cytoplasme. Ce fonctionnement, extrêmement simple, n'est possible que parce que le ribosome empêche en temps normal toute molécule d'eau d'accéder à la fragile liaison entre l'ARNt et le peptide, en l'isolant complètement du milieu extérieur. Après le relarguage du polypeptide, les sous-unités ribosomales 30S et 50S vont se dissocier à l'aide d'une molécule nommée "ribosome recycling factor", du facteur IF-3 qui va stabiliser la sous-unité 30S dissociée du complexe, et de l'hydrolyse d'un GTP porté par le facteur d'allongement EF-G. Alors seulement la traduction à proprement parler du polypeptide est achevée.

Déblocage des ribosomes par les ARNtm Il peut arriver qu'un ribosome soit très longtemps bloqué lors de la synthèse protéique. Cela peut faire suite à un problème au niveau de l'ARNm, comme l'absence de codon stop ou une succession de codons rares[1](Roche and Sauer, 1999), ou à une carence en acides aminés ou en ARNt dans l'organisme, ou finalement à l'action d'un antibiotique comme la tétracycline ou le chloramphénicol. Dans cette situation, trois problèmes principaux se posent à la cellule :

- Le ribosome bloqué, ainsi que tous les autres engagés dans la traduction en amont du site de blocage, ne peuvent plus être utilisés.
- L'ARNm bloqué également ne sera jamais traduit.
- Le produit peptidique de la traduction partielle de l'ARNm peut être toxique pour la cellule, s'il est relâché dans le milieu cytoplasmique.

L'ARNtm a pour fonction de répondre à certains de ces problèmes. Il n'agit sur un complexe ribosomal que si celui-ci est bloqué anormalement. Les mécanismes permettant de reconnaître un ribosome dans une telle situation sont encore inconnus, mais on sait qu'ils impliquent une protéine secondaire, SmpB (Roche and Sauer, 1999). Le déblocage se

[1]Un codon rare est un codon employé à une faible fréquence, et qui ne peut souvent être reconnu que par des ARNt peu concentrés dans la cellule. Les problèmes posés par les codons rares sont abordés en détail dans la section 3.5

déroule au départ comme si un ARNt classique état employé. L'ARNtm est recruté par le ribosome avec l'aide de deux protéines SmpB (Hallier et al., 2006). Sa partie structuralement semblable à un ARNt va occuper le site A du ribosome, comme un aa-ARNt recruté normalement. Ensuite, l'alanine transportée par l'ARNtm va être reliée à la chaîne peptidique. Jusque ici tout s'est passé comme lors d'une étape d'allongement classique. Mais la translocation est différente : pendant que l'ARNtm est déplacé dans le site P du ribosome, la partie codante de l'ARNtm va se placer en face du site A, délogeant l'ARNm qui était traduit. Puis la traduction de la partie messager de l'ARNtm va avoir lieu normalement, si elle n'est pas bloquée par une cause extérieure (c'est par exemple le rôle de certains antibiotiques). Ainsi, le peptide dont la synthèse avait été bloquée, n'est pas achevé, et se voit attaché une queue peptidique, dont la séquence est codée par la partie codante de l'ARNtm. Cette partie codante se terminant par un codon stop, la synthèse protéique va s'achever pour ce ribosome, qui va se détacher de l'ARNm. Les ribosomes bloqués en amont vont donc pouvoir reprendre la traduction, permettant, si le problème était local, de donner naissance à la protéine codée par l'ARNm orignal. Par la suite, les enzymes de dégradation présentes dans la cellule vont reconnaître spécifiquement la séquence de la queue du peptide relâché après le blocage et vont le dégrader, éliminant le risque de toxicité pour la cellule représenté par le polypeptide non achevé. Ce mécanisme permet, au final, de débloquer les ribosomes présents sur un ARNm, et autorise la traduction de cet ARNm, si le blocage était dû à des causes naturelles, par exemple une carence en acides aminés, ou une translocation erronée.

Une propriété importante de l'ARNtm dans notre étude est le fait qu'il n'est en général pas en compétition avec les autres ARNt lors de phases de traduction normale(Moore and Sauer, 2005). En effet, la protéine SmpB doit être recrutée par le ribosome pour que l'ARNtm soit à son tour recruté, et le processus de déblocage enclenché. Aucune expérience à ce jour n'a permis d'identifier comment SmpB était recrutée par le ribosome bloqué, mais il a été observé que cela n'arrive pas durant les temps d'attente normaux du ribosome où le site A est vide. Au vu du grand nombre d'ARNtm (13000) dans la cellule (Altuvia et al., 1997) comparé au faible nombre (Dong et al., 1996) des ARNt les plus rares (~ 500), cela évite qu'une compétition cinétique empêche presque toute traduction de codon traduit par un ARNt peu concentré.

2.3.4 Modifications post-traductionnelles

Une fois relâché dans le cytoplasme, le polypeptide n'est pas encore qualifié de protéine. Il est souvent nécessaire qu'il subisse des modifications, dites post-traductionnelles, avant d'être pleinement fonctionnel. Une des premières modifications faites sur le polypeptide, pendant qu'il est encore en formation au niveau du ribosome, est le clivage de la méthionine formylée en tête. En effet, une fois que la longueur du peptide dépasse 10 ou 15 acides aminés, le risque de circularisation du peptide est exclus, car structuralement la méthionine ne peut plus accéder à l'intérieur du site A du ribosome. Cette méthionine est donc clivée par une méthionine aminopeptidase dans de nombreux cas (Giglione et al., 2004), en particulier en fonction de l'acide aminé suivant sur le gène. En effet il a été montré que l'acide aminé en N-terminal jouait un rôle crucial sur la durée de demi-vie d'une protéine (Shrader et al., 1993), ce qui implique un contrôle strict sur la détermination des protéines cibles du clivage.

Le deuxième changement qui intervient sur presque tous les polypeptides est le replie-

ment. Pour être fonctionnelle, une protéine doit présenter certains sites à certains emplacements, et au contraire ne pas exposer ses régions hydrophobes pour éviter la formation d'agrégats. Ceci implique qu'elle doit se replier de façon particulière, afin d'obtenir une forme fonctionnelle. Les différents repliements possibles pour une protéine dénaturée ont été largement étudiés, et on suppose qu'ils sont dirigés par une baisse d'énergie libre du polypeptide (hypothèse d'Anfinson). Une classe de protéines nommées chaperons facilite le repliement des polypeptides. Elles isolent le polypeptide dénaturé du milieu extérieur, et facilitent son repliement en offrant un substrat sur lequel le polypeptide dénaturé va fixer ses séquences hydrophobes exposées. Un très bel exemple est la protéine chaperon HsP60, qui forme une véritable boîte avec un couvercle (Ellis, 2006; Tang et al., 2006), dans laquelle le polypeptide dénaturé entre et la protéine repliée ressort. Ces chaperons peuvent même agir pendant que le polypeptide est construit, directement à la sortie du complexe ribosomal, comme dans le cas du "trigger factor" chez *E. coli* (Kaiser et al., 2006), ou plus tard, après relarguage du polypeptide. De plus, ils peuvent également jouer leur rôle lors de situations de stress cellulaire, qui peuvent conduire à la dénaturation de protéines fonctionnelles. Dans ce cas elles sont utilisées comme un système de réparation des protéines.

Finalement, de nombreuses protéines doivent être localisées spécifiquement dans la cellule. Par exemple des protéines doivent se trouver dans la membrane interne ou externe, ou dans l'espace intermembranaire. Ces protéines doivent être conduites à leur emplacement définitif pour pouvoir remplir leur rôle. Des processus plus rapides que la simple diffusion jusqu'à la cible existent chez les bactéries, comme des mécanismes de transport adaptés, par exemple le translocon, dont une partie de la structure a été résolue récemment (Mitra et al., 2005), qui amène directement les protéines du ribosome au-delà de la membrane interne.

Chapitre 3

Code génétique et usage de codons

3.1 Le code génétique

On a vu au chapitre précédent que les synthétases chargeaient spécifiquement certains ARNt avec un acide aminé particulier, en fonction de l'anticodon porté par ces ARNt. On a également vu que les ARNt servaient d'intermédiaires entre l'ARNm et le peptide : ils permettent l'insertion de l'acide aminé qu'ils portent dans la séquence peptidique quand le complexe ribosmomal lit sur l'ARNm un codon complémentaire à leur anticodon. Cette combinaison de la spécificité des synthétases pour les anticodons des ARNt, et de ces derniers pour le codon complémentaire sur l'ARNm, crée une correspondance entre les codons présents sur l'ARNm – et donc sur l'ADN duquel il a été transcrit – et les acides aminés formant les protéines. Cette correspondance, véhiculée par la machinerie cellulaire de la traduction, est connue sous le nom de code génétique (Fig. 3.1).

Le code génétique a été identifié par une collaboration internationale, chapeautée par F. Crick, au cours des annés 60 (Crick, 1966). Sa propriété la plus importante a été rapidement mise à jour : non seulement le code est le même pour tous les gènes d'un organisme – car le système traductionnel est le même –, mais il est universel, c'est à dire qu'il est le même pour tous les organismes vivants, à de très rares exceptions près (Fox, 1987). Les différences en question concernent en particulier l'attribution des codons terminateurs, qui peuvent parfois coder pour un acide aminé, ou de certains codons particuliers qui ne sont pas utilisés dans quelques génomes. Le changement de sens d'un codon est, quant à lui, très rare. L'exemple le plus commun des codes alternatifs est fourni par les génomes mitochondriaux[1] (Sengupta et al., 2007). D'autres codes sont également obtenus par inclusion chez certains organismes de la sélénocystérine et de la pyrrolysine à la place d'un terminateur ou d'un autre acide aminé. Mais même ceux-ci ne diffèrent que relativement peu du code génétique standard, auquel on se référera par la suite.

Le code génétique fait correspondre 64 codons, correspondant aux 64 possibilités d'écriture d'un triplet dans un alphabet à 4 bases, à 20 acides aminés en général. Si on enlève les 3 codons terminateurs (TGA, TAG et TAA), auxquels aucun acide aminé ne correspond, il reste 61 codons pour 20 acides aminés : le code génétique est dégénéré, et plusieurs codons, dits synonymes, codent pour le même acide aminé. Comme on le voit sur la figure 3.1, les codons synonymes sont toujours très peu différents les uns des autres, avec rarement plus qu'un simple changement de la troisième base distinguant deux d'entre eux.

[1] Les mitochondries sont des organelles eucaryotes, qui ont la propriété de posséder leur propre génome.

	T	C	A	G
T	TTT Phe (F) TTC " TTA Leu (L) TTG "	TCT Ser (S) TCC " TCA " TCG "	TAT Tyr (Y) TAC TAA **Ter** TAG **Ter**	TGT Cys (C) TGC TGA **Ter** TGG Trp (W)
C	CTT Leu (L) CTC " CTA " CTG "	CCT Pro (P) CCC " CCA " CCG "	CAT His (H) CAC " CAA Gln (Q) CAG "	CGT Arg (R) CGC " CGA " CGG "
A	ATT Ile (I) ATC " ATA " **ATG** Met (M)	ACT Thr (T) ACC " ACA " ACG "	AAT Asn (N) AAC " AAA Lys (K) AAG "	AGT Ser (S) AGC " AGA Arg (R) AGG "
G	GTT Val (V) GTC " GTA " GTG "	GCT Ala (A) GCC " GCA " GCG "	GAT Asp (D) GAC " GAA Glu (E) GAG "	GGT Gly (G) GGC " GGA " GGG "

FIG. 3.1 – Le code génétique standard.

Ces propriétés structurelles du code seront étudiées plus en détails dans quelques lignes.

3.1.1 Les règles de reconnaissance floue entre ARNt et ARNm

Malgré la dégénérescence du code génétique, on pourrait supposer que le système de traduction a évolué de façon à associer à chaque codon un ARNt ayant l'anticodon correspondant, tout comme chaque acide aminé est directement chargé sur ses ARNt par une synthétase bien précise[1]. Quelques valeurs numériques permettent rapidement de voir que cette méthode de reconnaissance extensive n'est pas employée par les organismes bactériens : en effet de nombreux procaryotes contiennent au total moins de 61 ARNt, par exemple *Haemophilus influenzae* ou *M. leprae*, ce qui rend impossible toute reconnaissance fidèle et extensive, de cette façon, des codons portés par l'ARNm. De plus, parmi les organismes possédant plus de 61 ARNt, certains anticodons ne sont pas représentés parmi les ARNt. Par exemple *E. coli* contient 86 ARNt, mais au total seulement 39 anticodons sont utilisés. De même pour *Vibrio fischeri* ES 114, qui malgré ses 119 ARNt n'utilise que 41 anticodons différents.

Les ARNt présents dans l'organisme ont donc une mission difficile : ils doivent être capables, en plus du codon complémentaire à leur anticodon, de reconnaître d'autres codons, afin que ceux ci puissent être traduits. Il s'agit de l'hypothèse de reconnaissance floue, émise par F. Crick également (Crick, 1965). De plus, ils ne doivent pas s'apparier

[1]Dans certains cas, comme l'asparagine, c'est un acide aminé proche qui est chargé sur l'ARNt, puis modifié (Di Giulio, 2005a).

avec les codons non synonymes. Les critères de reconnaissance doivent donc être très précis. Nous allons les résumer ici (voir Table 3.a).

La reconnaissance entre codon et anticodon est un processus complexe, déterminée non seulement par l'appariement de bases complémentaires, mais aussi par l'hydrolyse d'une molécule de GTP, et la stabilisation des interactions entre ARN ribosomal, ARNt et ARNm (Agris, 2004). La multiplicité des interactions et la spécificité de la structure du ribosome et des ARNt permettent une reconnaissance précise de certains codons par des anticodons non complémentaires, notamment au niveau de la première base de l'anticodon, tout en assurant un taux d'erreur à l'insertion très faible, de l'ordre de 10^{-4}.

Tout d'abord, il faut savoir que les deux premières bases du codon, correspondant aux bases 35 et 36 sur l'ARNt[1], doivent être parfaitement complémentaires aux deux premières bases de l'anticodon pour que la reconnaissance ait lieu. L'intérêt d'un tel mécanisme peut être compris aisément en examinant la figure 3.1 : à l'exception des acides aminés dégénérés 6 fois, chaque acide aminé est caractérisé par les deux premières bases de ses codons. Le problème de la reconnaissance floue peut donc se poser ainsi : comment un anticodon peut-il s'apparier avec plusieurs codons ayant des troisièmes bases différentes, sans forcément reconnaître toutes les troisièmes bases possibles, ce qui poserait un problème au niveau des codons dégénérés 2 ou 3 fois ? Deux autres bases sur l'ARNt ont un rôle essentiel dans ce mécanisme : il s'agit de la base 34, celle reconnaissant la troisième base du codon, et de la base 37, qui peut stériquement influencer le lien entre la base voisine 36 et la première base de l'anticodon. L'importance de cette base a donné lieu à l'appellation d'anticodon étendu pour les bases 34 à 37 de l'ARNt (Yarus, 1982). En ce qui concerne la base 37, on résumera simplement en disant qu'il s'agit en règle générale d'une purine (A, G ou un de leurs dérivés) et qu'elle est très souvent soumise à des modifications sur l'ARNt qui augmentent la spécificité de ce dernier pour certains codons.

Anticodon 34	Base reconnue
A	Non utilisé
I	U,C,A
C	G
G	C,U
U	A,G
U_{modif}	G ou A,G ou A,G,U,C

TAB. 3.a – Règles de reconnaissance floue entre l'anticodon 34 de l'ARNt et la troisième base du codon. U_{modif} représente les trois possibilités de mutations observées sur U, avec les trois reconnaissances associées.*Inspirée de Agris et al. (2007)*

Au niveau de la base 34, des règles claires ont été observées. La base C_{34} – qui dénote une base C en position 34 sur l'ARNt – ne s'apparie qu'avec son complémentaire G, et G_{34} s'apparie avec les pyrimidines C et U. L'adénine A_{34} n'est quasiment jamais observée dans l'ARNt ; même au niveau de la séquence génétique, seulement 12% des gènes codant pour un ARNt reconnaissant 4 codons synonymes ont une adénine en position 34 (Sprinzl

[1]Les bases de l'anticodon sont numérotées d'après leur place dans la séquence de l'ARNt. Elles portent les numéros 34, 35 et 36, et s'apparient respectivement à la dernière base du codon sur l'ARNm, celle du milieu et la première.

and Vassilenko, 2003). La reconnaissance des codons finissant par T s'effectue à la place grâce à une inosine I_{34}, car l'inosine permet au complexe ribosomal d'être plus stable une fois l'ARNt déplacé au site P (Agris, 2004). Au niveau des ARNt reconnaissant les codons 4 fois dégénérés, c'est U_{34} qui est observé le plus souvent, suivi par G_{34} (Rocha, 2004). U peut normalement s'apparier avec les purines A et G, mais des modifications chimiques de U_{34} permettent à certains ARNt de reconnaître tout ou partie des 4 bases A, C, T et G, résolvant le problème des codons 4 fois dégénérés. Pour les codons deux fois dégénérés, la possibilité pour G_{34} de s'apparier avec les deux pyrimidines C et T, et pour U_{34} de reconnaître les purines, permet également à un seul ARNt de servir pour deux codons différents. Il est connu que pour la phénylalanine, la tyrosine et l'asparagine, trois acides aminés codés par seulement deux codons (respectivement TTT/TTC, TAT/TAC et AAT/AAC), un seul ARNt dont l'anticodon contient toujours G_{34} reconnaît les deux codons. Ceci n'est possible que grâce à la structure particulière des codons deux fois dégénérés, qui se terminent toujours soit par une purine, soit par une pyrimidine.

Les règles de reconnaissance floue sont très légèrement différentes dans les génomes mitochondriaux, et ceux de certaines bactéries, comme *Mycoplasma capricolum*. Ceci autorise ces génomes, très réduits, à minimiser le nombre d'ARNt nécessaires à la traduction des 61 codons sens : *M. capricolum* n'utilise au total que 29 ARNt différents, sur un minimum de 26 théoriquement possible pour ne pas commettre d'erreur de décodage (Marck and Grosjean, 2002).

3.1.2 La structure du code génétique est-elle optimale ?

Le code génétique n'est pas organisé de manière aléatoire. On a déjà remarqué précédemment que les règles de reconnaissance floue ne fonctionnaient que grâce à la structure particulière du code, les codons synonymes étant "proches" dans le code. Un code génétique aléatoire, complètement brassé, n'aurait pas cette propriété, et nécessiterait un appareillage moléculaire beaucoup plus important, en termes d'ARNt du moins, pour que la traduction puisse avoir lieu. On peut donc considérer le code génétique comme optimisé de ce point de vue, à savoir de façon à minimiser le nombre d'ARNt différents nécessaires à la traduction.

Mais le code génétique est structuré également à d'autres niveaux. De très nombreux travaux sur ce thème ont été poursuivis. Nous n'en citerons que quelques-uns. Une des thématiques qui a été la plus explorée est celle visant à montrer que le code génétique est construit de manière à être robuste face à des mutations ponctuelles (Freeland and Hurst, 1998; Goodarzi et al., 2007, 2005; Haig and Hurst, 1991), de manière à ce que :

- Une mutation simple conduise le plus souvent possible sur un codon synonyme. Ce serait une autre raison pour laquelle les codons synonymes sont associés en blocs, et pour laquelle les acides aminés dégénérés 6 fois ont des codons situés dans la même colonne[1]. Ceci implique qu'une mutation ponctuelle sur certains codons de l'arginine (CGA et CGG) ou de la leucine (CTA et CTG) a 5 chances sur 9 de toujours coder pour le même acide aminé. On peut estimer l'importance de cette structure en sachant que, en général, l'usage des acides aminés est proportionnel au nombre de codons qui leurs correspondent, et donc que ces acides aminés sont très employés en général.

[1] À l'exception de la sérine, le seul acide aminé dont tous les synonymes ne soient pas adjacents à une mutation près.

– Si une mutation simple provoque un changement d'acide aminé, le nouvel acide aminé doit avoir des propriétés physico-chimiques semblables à celles de celui duquel il dérive. Ceci signifie que des codons proches correspondent à des acides aminés similaires. Un exemple est l'ensemble des codons de la leucine, l'isoleucine et la valine, trois acides aminés très semblables, qui sont tous dans la même colonne du code.

D'autres particularités du code ont également été observés. Tout d'abord, une possibilité est que la redondance du code permette un certain choix au niveau des bases employées dans les séquences codantes (Shabalina et al., 2006), et que ces bases puissent servir dans la structure secondaire de certains ARNm. En effet le repliement partiel des ARNm peut être employé pour réguler la traduction (Kozak, 2005). Ensuite, il a également été observé que le nombre d'acides aminés présent dans le code pouvait, d'une certains façon, présenter le meilleur rapport entre maximisation de la diversité des codons et minimisation des chances de mutation non synonyme (Gultepe and Kurnaz, 2005). Finalement, le cadre de lecture d'une séquence – qui représente le choix de la position de départ, et donc le découpage en triplets de la séquence nucléotidique – peut être modifié par erreur au cours de la traduction. Des travaux récents sur la robustesse du code génétique face à des décalages du cadre de lecture du ribosome ont tiré des conclusions intéressantes mais contradictoires. Dans ce cas, rare dans la nature (avec une fréquence de 5.10^{-5}), certaines analyses ont montré que le code génétique continuerait cependant à synthétiser une protéine aux propriétés physico-chimiques globalement semblables à celle qui aurait dû naître (Chechetkin, 2006). Ensuite, d'autres résultats ont au contraire conclu que ce type d'erreur conduirait la traduction à s'arrêter plus vite qu'avec d'autres codes génétique, en montrant que la probabilité qu'un codon stop apparaisse dans un cadre de lecture décalé, à l'intérieur de séquences codantes, était beaucoup plus grande pour le code génétique réel que pour un code aléatoire (Itzkovitz and Alon, 2007; Seligmann and Pollock, 2004).

3.1.3 Origine et évolution du code

La question de l'origine, et de l'éventuelle évolution, du code génétique, a intéressé les chercheurs depuis longtemps. Une hypothèse est celle d'une version accidentelle du code, qui serait restée "gelée" après sa création, pour maintenir la fonction des protéines des organismes de l'époque. En effet, tout changement brutal dans le code après l'émergence de la vie pourrait avoir des conséquences dramatiques sur les protéines synthétisées, et il a donc longtemps été supposé que le code n'avait pu évoluer qu'aux origines de la vie, et peu ou plus après. La question de l'évolution du code s'est donc confondue avec celle de son origine. Et même sous cette dénomination unificatrice, les problèmes étudiés sont multiples : ils recouvrent l'évolution des molécules qui représentent le code, à savoir les ARNt et les synthétases, la question des conditions dans lesquelles le code a émergé, et le problème du lien entre la structure du code génétique et les acides aminés qui le composent.

Trois théories principales sur la façon dont le code s'est construit essayent d'expliquer sa structure (Di Giulio, 2005a) :

– La théorie physico-chimique se base directement sur la robustesse du code génétique, et fait l'hypothèse que c'est la sélection naturelle qui a agi sur les codes possibles pour ne sélectionner que celui qui minimisait les erreurs dues à des mutations ponctuelles. Un tel code aurait pu évoluer en minimisant les conséquences des inévitables

d'erreurs traductionnelles dans les organismes l'utilisant, leur donnant un avantage sélectif. Si elle a l'avantage d'expliquer directement l'optimisation structurelle étudiée au paragraphe précédent, elle ne parvient cependant pas à expliquer les bases des autres théories.

– La théorie stéréochimique est basée sur l'idée que les codons et les acides aminés pour lesquelles ils codent doivent d'une certains façon pouvoir interagir, où avoir des propriétés similaires. Cette théorie a été formulée avant même que le code génétique ne soit connu (Gamow, 1954). Plus récemment, des expériences en sa faveur ont montré que certains sites d'attachement des protéines sur des séquences d'ARN incluaient, de façon très significative, les triplets codant pour les acides aminés par lesquels a lieu l'interaction entre ARN et protéine : par exemple, le codon UAA correspondant à l'isoleucine peut justement servir de point d'attache sur la séquence d'ARN à une molécule d'isoleucine (Yarus, 2002).
– La théorie de coévolution se base sur l'étude des voies métaboliques de la biosynthèse des acides aminés. Il a été montré (Wong, 1975, 2005) que les voies de biosynthèse des acides aminés différents, dont les codons sont proches, se recouvrent partiellement. On peut supposer qu'au début de la vie, il n'existait qu'un nombre restreint d'acides aminés, et tous les codons servaient pour eux. Petit à petit de nouveaux acides aminés ont été synthétisés, par évolution des voies biosynthétiques, et ils ont pu progressivement entrer en compétition avec les anciens acides aminés, jusqu'à ce qu'une séparation nette des codons représentant chacun ait lieu. Cette théorie présente l'avantage d'expliquer l'évolution du code génétique jusqu'à son état actuel sans partir de ce dernier, mais expliquer également la robustesse du code : si on suppose l'acquisition graduelle d'acides aminés par les organismes, il faut également faire l'hypothèse que les remplacements d'anciens acides aminés par des nouveaux n'ont pas donné pas lieu à des changements trop importants au niveau des fonctions protéiques. Ceci amène à la conclusion que les acides aminés codés par des codons proches doivent effectivement être similaires, puisqu'ils ont servis à une certaine époque pour la même fonction.

Sans adhérer particulièrement à l'une ou l'autre de ces théories, d'autres travaux ont permis de faire des hypothèses sur le milieu dans lequel le code génétique s'était formé, et l'ordre dans lequel les composants du système de traduction et les acides aminés ont été acquis. La question de l'environnement dans lequel le code s'est formé est liée à celle des origines de la vie, dans une conception de ces origines où les protéines n'apparaissent qu'une fois le code approximativement fixé. De très nombreuses études ont été menées, basées sur la comparaison des séquences génétiques d'organismes vivant dans des milieux différents, l'étude des substitutions d'acides aminés entre des molécules conservées de ces organismes (Di Giulio, 2007, 2000, 2005b) ou encore sur des approches ancrées dans la chimie (Granick, 1957; Wachtershauser, 1988; Wächtershäuser, 2007). Les résultats obtenus attendent encore une confirmation, et si certains tendraient à affirmer que le code génétique s'est développé dans un milieu chaud, acide et anaérobie, renforçant l'image d'une soupe primitive chaude à l'origine de la vie, d'autres feraient pencher la balance en faveur d'une origine de la vie plus minérale, partiellement basée sur la chimie de surface de certains minéraux (Danchin, 1990; Wachtershauser, 1988).

L'ordre d'acquisition des composants du système de traduction, et l'influence qu'ils ont pu avoir les uns sur les autres, ont également longuement été débattus. Il est en général supposé que les ARN ribosomaux sont les vestiges des premières molécules du

vivant apparues sur Terre, à cause de leur rôle catalytique dans le processus de traduction. Concernant les ARNt et les synthétases, de récentes analyses phylogénétiques font supposer que les ARNt ont peut-être précédé les synthétases, menaçant la théorie de coévolution du code et des acides aminés (Hohn et al., 2006). Finalement, l'étude de l'ordre d'intégration des acides aminés dans le code, elle, a fait l'objet de très nombreuses études physiques, chimiques et biologiques pendant 50 ans, dont la synthèse tendrait vers une théorie générale (Trifonov, 2004) : l'intégration des acides aminés y est expliquée en parallèle avec l'évolution du code génétique, en incluant les acides aminés de manière progressive dans les protéines au fur et à mesure que les codons sont employés dans les séquences du vivant. Cette théorie n'explique cependant pas le choix exact des acides aminés présents dans le code : en effet de très nombreux acides aminés ne sont utilisés par les organismes que pour la biosynthèse d'autres, et n'apparaissent pas dans les protéines. Même si certains choix peuvent s'expliquer, comme on l'a vu page 49, les raisons de l'existence d'un alphabet bien précis d'acides aminés restent encore à découvrir (Lu and Freeland, 2006).

Pour finir, nous dirons simplement que l'évolution du code pourrait bientôt devenir l'évolution des codes, car la vision de l'unicité du code a récemment été remise en cause (Vetsigian et al., 2006). Leurs simulations supposent l'existence d'un grand nombre de codes différents aux origines de la vie, et les mettent en interaction par le transfert horizontal de gènes entre organismes. Dans ce cadre, la sélection naturelle favoriserait les organismes capables d'utiliser au mieux les séquences qu'ils peuvent acquérir, et donc ceux dont le code est le plus proche, en terme de produit protéique, de ce qu'ils peuvent recevoir de leur environnement. Cela induirait une convergence des codes utilisés. Cette hypothèse, attrayante alors que le transfert horizontal est de plus en plus considéré comme un moteur évolutif, permettrait de plus d'expliquer pourquoi le code génétique est optimisé, et de s'affranchir de la notion d'"accident conservé".

3.2 Définition du biais d'usage de codons

Le biais d'usage de codons est défini comme l'emploi à des fréquences différentes de codons synonymes par un organisme. Par exemple, chez *B. subtilis*, les codons CCX, où X représente n'importe quel base, codent tous pour une proline. Mais leurs fréquences d'usage moyennes sur les séquences codantes sont différentes :

- CCT 0.29
- CCC 0.09
- CCA 0.19
- CCG 0.43

Un autre exemple, chez *E. coli* K12, est l'emploi des codons pour l'arginine : sur les 6 codons, 2 seulement (CGC et CGT) totalisent plus de 78% des codons utilisés, tandis que le rassemblement des 3 codons AGA, AGG et CGA forme moins de 12% des codons employés. De tels exemples peuvent être multipliés sans difficulté. Le même phénomène d'usage différencié des codons synonymes se retrouve dans pratiquement toutes les espèces bactériennes, pour beaucoup d'acides aminés. Il existe à la fois chez les eucaryotes et les procarytes, mais nous nous focaliseront uniquement sur le domaine bactérien, qui présente déjà une grande diversité. Pour une étude chez les eucaryotes, on pourra se référer à Akashi (2001) et aux références citées. Le biais d'usage de codons (que l'on appelera par

la suite biais de codons, ou biais d'usage du code), s'il est très répandu, n'est pas le même pour toutes les espèces. Par exemple, la lysine est codée par deux codons, AAA et AAG. Les fréquences d'emploi de ces deux codons dans différents organismes peuvent être très variables, comme présenté dans la table 3.b.

	Anaeromyxobacter dehalogenans	*Deinococcus radiodurans*	*Aquifex aeolicus*	*Escherichia coli* K12	*Buchnera aphidicola*
AAA	1%	29.3%	48.1%	76.6%	91%
AAG	99%	70.7%	51.9%	23.4%	9%

TAB. 3.b – Usage de codons pour la lysine.

La première hypothèse qui peut venir à l'esprit pour expliquer le biais de codons observé chez les organismes bactériens est également la plus simple, à savoir supposer que le phénomène est aléatoire et dû à la dérive génétique. Si l'on suppose les mutations entre codons synonymes neutres au niveau phénotypique, c'est à dire qu'on suppose que l'adaptation d'un organisme à son milieu ne dépend que de son contenu protéique, les fréquences relatives des codons synonymes dans la population peuvent varier sans contraintes, et ne sont soumis à aucune forme de sélection. C'est la théorie de l'évolution neutraliste (Kimura, 1968; King and Jukes, 1969). Dans ce cas, que l'on peut modéliser comme un processus de Markov, il est possible d'observer de très grands écarts entre les fréquences de deux codons synonymes, malgré l'absence de sélection. Ce phénomène est simplement dû à l'aléatoire inhérent du système, et au rééchantillonage des codons à chaque génération à partir des génomes parentaux.

Cependant une observation vint contredire cette vision des choses au début des années 80. En effet, il fut remarqué que tous les gènes – connus à l'époque – d'un même organisme présentaient tous le même usage biaisé de codons (Grantham et al., 1980). Cette observation est contradictoire avec la théorie neutraliste, car pour que tous les gènes aient le même biais, il faut supposer qu'ils sont tous soumis à une pression de sélection commune. Ce travail a mené à la formulation de "l'hypothèse du génome", qui dit que l'unité évolutive, au niveau des pressions d'appliquant sur le biais de codons, est le génome et non le gène. Par la suite, d'autres travaux ont montré que cet usage biaisé de codons était particulièrement important dans les gènes fortement exprimés (Gouy and Gautier, 1982; Grantham et al., 1981), impliquant un lien entre usage de codons et expression génique, et des analyses à plus grande échelle gène par gène ont été menées pour comprendre comment le biais d'usage de codons se répartissait entre les séquences d'un génome (Gautier et al., 1985; Gouy et al., 1985).

3.3 Mesures du biais de codons

Différentes manières de mesurer le biais d'usage de codons ont été développées depuis 1981, de plus en plus sophistiquées. L'idée générale est de pouvoir mesurer le biais de codons, à l'échelle du gène ou du génome, par un simple indicateur chiffré. Les mesures du biais de codons de gènes individuels permettent de classer les différents gènes d'un même génome. Une fois le score du gène obtenu, on peut le comparer à celui des séquences des

gènes fortement exprimés connus de son génome, et en inférer son niveau d'expression (Karlin and Mrazek, 2000) ou son origine, car un gène ayant un biais de codon très différent du reste de son génome a probablement été acquis récemment (Koonin et al., 2001; Médigue et al., 1991). Il a également été proposé d'employer le biais de codons comme un moyen de détecter de nouveaux gènes et d'aider à l'annotation (Karlin, 2001; States and Gish, 1994). Au niveau des génomes, les mesures du biais peuvent permettre d'estimer les facteurs de sélection agissant sur l'organisme dans son ensemble, et ainsi de comparer les organismes entre eux.

Nous allons maintenant passer en revue les indicateurs du biais de codons. Ils sont trop nombreux pour qu'on les cite tous, on ne décrira donc que ceux qui présentent un intérêt historique ou technique. Les premiers ont tous pour caractéristique de nécessiter de connaître *a priori* les codons préférés, ou majeurs, de l'organisme, c'est-à-dire les codons qu'il utilise plus fréquemment dans ses gènes fortement exprimés. Parmi ceux-ci on trouve le "codon biais index", qui calcule l'excès d'usage des codons préférés de l'organisme dans chaque gène (Bennetzen and Hall, 1982) :

$$CBI = \frac{N_{maj} - \langle N_{maj} \rangle}{N_{tot} - \langle N_{maj} \rangle}, \tag{3.1}$$

où les moyennes $\langle N_{maj} \rangle$ correspondent aux valeurs attendues pour l'usage des codons majeurs si l'usage du code était aléatoire, N_{maj} est le vrai nombre de résidus codés par des codons préférés, et N_{tot} est la longueur de la protéine. On a $\langle N_{maj} \rangle = \sum_{i=1}^{17} r_i n_i$, la somme sur tous les acides aminés i de la fraction de codons préférés r_i pour cet acide aminé multipliée par le nombre de résidus n_i de cet acide aminé. La somme se fait sur 17 acides aminés, car la méthionine, le tryptophane, et l'acide aspartique[1] sont exclus.

Un autre indicateur de cette lignée a été développé la même année (Ikemura, 1981b). La mesure en question est la fraction des codons optimaux F_{op}, et se calcule simplement comme le rapport du nombre de codons préférés employé dans le gène divisé par le nombre total de codons :

$$F_{op} = \frac{N_{maj}}{N_{tot}}. \tag{3.2}$$

Cette mesure est très fortement corrélée au CBI. Le principal défaut de ces indicateurs est qu'il nécessite de savoir quels sont les codons majeurs avant l'analyse, et qu'en fonction des choix faits pour déterminer ceux-ci les résultats ne seront pas les mêmes. De plus, certains acides aminés ne peuvent pas être tenus en compte, car il est impossible de déterminer le codon majeur pour eux. Finalement, le nombre des codons majeurs n'étant pas le même pour chaque acide aminé dans chaque espèce, les comparaisons inter-espèces faites avec ces indicateurs sont à prendre avec précaution.

Ces difficultés techniques ont conduit à aborder le problème d'une autre façon, statistiquement plus fiable et plus facile à généraliser, consistant à comparer l'usage de codons dans un génome à un usage aléatoire, pour détecter la présence de biais. Les premiers indicateurs de cette gamme sont le "codon preference biais" CPB (McLachlan et al., 1984), et le "codon preference statistic" CPS (Gribskov et al., 1984), qui est le rapport de deux vraisemblances : celle de trouver un codon particulier dans un gène fortement exprimé et celle de le trouver dans une séquence aléatoire de même composition protéique. En notant

[1] L'exclusion de l'acide aspartique est due à son absence de biais d'usage de codons mesurée dans le papier original de Bennetzen et al.

f_i^j la fréquence relative d'usage du codon i pour coder l'acide aminé j, et r_i^j la fréquence relative d'usage du codon i pour un acide aminé j avec un usage du code non biaisé (par exemple, $r_{TTT}^{Phe} = 0.5$ car seulement 2 codons codent pour la phénylalanine), pour un gène de longueur L, on a :

$$CPS = \left(\prod_{k=1}^{L} \frac{f_k^j}{r_k^j} \right)^{1/L}, \tag{3.3}$$

où k représente successivement chaque codon dans le gène. Le calcul du CPB est un peu plus complexe, et on le définit comme le z-score de la probabilité d'observer le gène avec une distribution multinomiale des codons utilisant les fréquences d'emploi d'un génome non biaisé r_i^j.

Le principal inconvénient de ces deux indicateurs est de ne pas tenir en compte la composition de base du génome, et de donner de très grands scores pour des génomes à la composition biaisée en acides aminés ou en nucléotides. De plus, ils ne sont pas normalisés, et très sensibles à la longueur des gènes : deux gènes employant uniquement des codons majeurs n'ont pas forcément la même valeur de CPS ou de CPB selon leur longueur et leur génome d'appartenance.

Une normalisation inspirée du CPS a été proposée par la suite (Sharp and Li, 1987) : il s'agit du "Codon Adaptation Index", ou CAI. Utilisant un ensemble de gènes fortement exprimés prédéfini – ce qui évite d'avoir à désigner les codons majeurs tout en les incluant comme référence –, cette mesure est restée très utilisée, et est encore aujourd'hui une référence en matière de mesure d'usage de codons. Le CAI se calcule en deux étapes ; il faut tout d'abord calculer, *sur les gènes fortement exprimés*, l'adaptativité w_{ij} de chaque codon. On a, avec les notations précédentes :

$$w_{ij} = \frac{f_i^j}{\max_i(f_i^j)}. \tag{3.4}$$

L'adaptativité du codon le plus employé pour chaque acide aminé est donc 1. Ensuite, pour n'importe quel gène, on peut calculer le CAI comme suit :

$$CAI = \left(\prod_{k=1}^{L} w_k \right)^{1/L}. \tag{3.5}$$

L'avantage du CAI est qu'il garde l'idée du CPS, à savoir mesurer le biais directement en comparant les fréquences d'usage de codons, et pas à partir d'un sous-ensemble prédéterminé de codons majeurs. De plus, cet indicateur permet de tenir compte des différences d'usage entre les codons majeurs, tous considérés équivalents auparavant. Finalement, il est normalisé, ce qui permet – prudemment – de faire des comparaisons inter-gènes et inter-espèces.

Le seul défaut du CAI est de nécessiter la comparaison a un set de gènes prédéfini. Historiquement, les gènes fortement exprimés étaient utilisés, car le biais que l'on voulait détecter était un biais lié à l'expression génique. Mais l'emploi d'autres ensembles de gènes permet de détecter des biais liés à d'autres contraintes. Ceci a permis, par des comparaisons systématiques de groupes de gènes du même génome entre eux, de détecter le biais dominant du génome, qu'il soit lié à l'expression génique, à sa composition, à la position des gènes sur les deux brins d'ADN (Carbone et al., 2003) ou encore au style de vie de l'organisme (Willenbrock et al., 2006).

Récemment, le CAI a vu le développement d'analogues qui tiennent explicitement en compte le contenu en ARNt dans la cellule. En effet, l'hypothèse selon laquelle le biais d'usage de codons est corrélé au contenu cellulaire en ARNt a reçu beaucoup de soutien[1]. Des indicateurs ont donc été développés, qui mesurent l'emploi de codons non plus par rapport à leur emploi dans les gènes fortement exprimés, mais par rapport au contenu en ARNt les reconnaissant dans la cellule. Il s'agit du tAI (Reis et al., 2004) et de la mesure S pour "codon biais Strength" (Sharp et al., 2005). L'étendue de leurs valeurs sur l'ensemble des génomes bactériens a permis de faire des hypothèses générales sur l'efficacité du biais d'usage de codons chez les bactéries, mais de nombreuses différences d'interprétation subsistent.

Pour finir, deux autres indicateurs sont fréquemment utilisés. Basés sur l'idée de comparer l'usage des codons à un usage non biaisé, comme le CPB et le CPS, il s'agit du "Nombre effectif de codons" $\hat{N}_c$ (Wright, 1990) et de sa version normalisée par rapport au contenu en GC, $\hat{N}_c'$ (Novembre, 2002). L'idée est de calculer, sur 61 codons possibles, combien sont "réellement" utilisés par un gène, en donnant un poids à chaque codon en fonction de sa fréquence d'emploi relativement à ses synonymes. La méthode de calcul est inspirée de la génétique des populations, et on a :

$$\hat{N}_c = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6}, \tag{3.6}$$

où F_n est l'hétérozygotie moyenne attendue sur les codons dégénérés n fois, que l'on calcule ainsi :

$$F_n = \frac{1}{Q_n} \sum_{a=1}^{Q_n} \frac{\left(n_a \sum_{j=1}^{n} p_j^2 - 1\right)}{n_a - 1}. \tag{3.7}$$

Dans cette équation Q_n est le nombre d'acides aminés dégénérés n fois, n_a le nombre total de codons employés pour cet acide aminé, et p_j la fréquence d'emploi du codon j codant pour a dans le gène (et non pas relativement à ses synonymes). La première version de cet indicateur avait le défaut de varier avec le contenu en G+C du gène dans lequel on le mesurait, et bien que cette variation soit quantifiée, elle empêchait l'usage de statistiques bien définies dans les analyses. Une version normalisée a été développée, qui annule ce biais. Un des intérêts de $\hat{N}_c'$ sur le CAI est le fait qu'il est encore moins sensible aux faibles longueurs de gènes et aux effets d'échantillonnage. Un autre est qu'il autorise une mesure, incluant les biais compositionnels du génome, du niveau auquel un gène ou un génome est biaisé en terme d'usage de codons. Ceci permet de comparer des génomes complets pour chercher une sélection potentielle, ce qui est soumis à caution en employant le CAI.

D'autres méthodes ont été employées pour étudier le biais de codons dans un organisme, qui ne résument pas le biais de codons en une seule valeur, gardant ainsi plus d'information. Il s'agit majoritairement de méthodes d'analyse des correspondances : un historique de ce qui a été fait est présenté au chapitre 6.

On peut voir que le calcul d'indices pour mesurer de façon effective le biais de codons dans un gène (ou un génome), a suscité de nombreuses publications, et s'est affiné avec le temps. Le problème reste que chaque auteur publiant son propre indice, avec ses biais, avantages mais aussi inconvénients, il est difficile de comparer les résultats des analyses

[1] voir la section suivante, "Causes du biais de codons".

faites avec ces différents indices. On pourra se référer pour plus d'exemples d'indicateurs à Gladitz et al. (2005), Zhao et al. (2003), Karlin et al. (1998) ou Merkl (2003) et leurs références.

3.4 Causes du biais de codons

Les mesures du biais de codons sont construites pour détecter des gènes fortement exprimés, en comparant les gènes en question au reste du génome, et pour quantifier le biais d'usage de codons dans un organisme, en comparant son biais à celui d'autres génomes. Une question fondamentale qui se pose, depuis la découverte du biais d'usage de codons, est la nature des processus qui le causent, à savoir :

- Quels sont les processus qui agissent sur le contenu en codons d'un génome ?
- Pourquoi ces processus n'ont-ils pas la même influence sur tous les gènes d'un organisme ?

Nous allons exposer les principales hypothèses qui ont tenté de répondre à ces questions, ainsi que les preuves apportées par l'étude des génomes. Ces hypothèses ont été en concurrence pendant près de 20 ans, ce qui a été résumé dans un article récent par la phrase "the study of codon usage is one of the most controversial areas of molecular evolution" (Reis et al., 2004). Le plus probable est que tous les processus décrits ici ont une influence sur le biais de codons de tous organismes, et que c'est leur importance relative au sein de chacun qui définit exactement l'usage du code à l'intérieur du génome.

3.4.1 Les biais de composition

Une façon d'expliquer le biais de codons présent dans les génomes bactériens est de supposer qu'il est dirigé par la composition globale du génome. Ceci présente l'avantage, contrairement à la théorie neutraliste, de donner une raison pour laquelle chaque génome a un biais de codons qui lui est propre, et pour laquelle les variations inter-génomiques peuvent être très grandes : si les organismes ne sont pas soumis aux mêmes pressions sélectives, leurs compositions génomiques peuvent varier, créant ainsi les disparités observées entre leurs biais d'usage de codons.

a) Le biais de GC

La première hypothèse qui peut être faite, la plus simple, est de tenter d'expliquer le biais d'usage de codons par le contenu en GC. En effet, on a vu que le pourcentage en GC pouvait énormément varier d'un organisme à l'autre. Un exemple simple est d'imaginer un organisme très biaisé, avec un fort pourcentage en GC : dans ce cas, pour coder une histidine, à laquelle correspondent les codons CAT et CAC, l'organisme emploiera très souvent le codon CAC, simplement parce que, à grande échelle, son génome contient plus de C que de T. Ce type de raisonnement, basé sur les corrélations entre la troisième base des codons et le contenu général du génome, est à la base de toute une série de travaux (Sueoka, 1962, 1988, 1992), qui ont permis de montrer que :

i) il existe effectivement un biais de composition en GC dans les génomes, qui peut être expliqué par des mutations préférentielles de AT $\rightarrow$ GC propres à chaque organisme.

ii) il est possible d'estimer le niveau de sélection auquel une séquence est soumise en comparant son contenu en GC à celui d'une séquence neutre du même génome, par exemple une séquence intergénique.

Par la suite, plusieurs travaux viendront appuyer cette hypothèse, selon laquelle le contenu en GC est la principale cause de variation de biais d'usage de codons entre les génomes : l'analyse de certains génomes montrera des cas très clairs où le biais de codons est dirigé par le faible contenu en GC, à l'échelle du génome (Charles et al., 2006), et des analyses comparatives de nombreux génomes montreront qu'il est possible de prédire globalement l'usage de codons dans un génome à partir de son contenu en GC (Chen et al., 2004; Knight et al., 2001). Les raisons du biais de composition en GC, quant à elles, sont moins claires ; cependant plusieurs hypothèses ont été formulées pour expliquer le contenu en GC dans différentes classes d'organismes. Il a été observé que le contenu en GC des bactéries est plus faible que celui des eucaryotes, et une hypothèse explicative est que les liaisons A-T étant plus fragiles, elles permettraient un appariement moins stable mais plus rapide au niveau de la liaison codon-anticodon de l'ARNt lors de la traduction, résultant en une accélération de la traduction chez les bactéries, au détriment de la précision privilégiée par les eucaryotes, et matérialisée chez eux par des liaisons G-C (Pluhar, 2006). Une autre observation est que les organismes parasites ou viraux ont un faible contenu en GC relativement à leur hôte, ce qui pourrait s'expliquer par la plus grande disponibilité des bases A et T, ou encore par leur moindre coût de synthèse, dans les organismes hôtes (Rocha and Danchin, 2002). Le contenu en GC est également lié au métabolisme de l'organisme (Naya et al., 2002).

b) Le biais de composition lié au brin.

Un autre biais compositionnel est celui qui différencie la composition des deux brins d'ADN. En l'absence de biais mutationnel particulier, on peut en effet attendre un équilibre de composition des bases complémentaires à l'intérieur de chaque brin : la fréquence d'emploi de la base A, notée [A], doit être égale à celle de la base T, et de même on doit avoir [C]=[G], avec les mêmes notations (Sueoka, 1995). Cette règle, nommée "Parity Rule 2", ou PR2[1], est à bien différencier de la règle de Chargaff, qui dit que A est complémentaire à T et C à G : dans le cas de la PR2, on parle des fréquences sur chaque brin, et pas sur l'ensemble de l'ADN.

Il a été très tôt mis à jour que la PR2 est l'exception plutôt que la règle, aussi bien dans les génomes bactériens que ceux de phages (Kano-Sueoka et al., 1999; McLean et al., 1998; Rocha et al., 1999; Sueoka, 1995). Un enrichissement en bases G et T sur le brin précoce – le brin orienté de 5' vers 3' dans la direction de la réplication – a été observé. Cet enrichissement est suffisamment fort et cohérent, tout le long du chromosome, pour permettre l'identification de l'origine de réplication par le simple calcul du nombre cumulé de bases G en excès par rapport au bases C sur un brin (Lobry, 1996a,b). En effet la valeur d'un indicateur comme $\frac{[G]-[C]}{[G]+[C]}$ subit un changement de signe au niveau de l'origine de réplication, dû au fait qu'en traversant l'origine de réplication et en continuant à lire le chromosome dans le même sens, on passe du brin précoce au brin tardif (ou inversement), et donc que l'excès de GT d'un côté devient un déficit de l'autre. L'hypothèse a également été émise que les organismes qui n'affichent pas un tel biais, notamment les Archaea et

[1]La "Parity Rule 1" dit que les biais mutationnels sont les mêmes à l'intérieur de chaque brin, et est une condition nécessaire à la PR2.

certains eucaryotes, ont plusieurs origines de réplication (Nikolaou and Almirantis, 2005; Olsen and Woese, 1997).

Quelles sont les raisons d'un tel biais de composition entre les deux brins ? Leur principale distinction étant liée à la réplication, on peut supposer que le mécanisme qui induit une asymétrie dans leur contenu en G, et à un moindre niveau en T, est lié à la réplication. De très nombreuses causes ont été invoquées (pour une revue récente on pourra se référer à Rocha et al. (2006)), mais celle qui permet d'expliquer la plus grande partie des observations de la manière la plus simple est la désamination de la cytosine sur le brin précoce (Francino and Ochman, 1997; Frank and Lobry, 1999; Lobry, 1996b). L'idée est basée sur l'observation que la réaction de désamination de la cytosine, qui transforme C en T, est 140 fois plus rapide sur un simple brin d'ADN que sur un double brin (Frederico et al., 1990). Or, lors de la réplication, le brin complémentaire du brin tardif en formation est exposé sous forme simple brin beaucoup plus longtemps que le brin complémentaire au brin précoce. Ceci est dû à la différence de synthèse des deux brins, le brin précoce étant synthétisé de manière continue, et donc s'appariant directement avec son complémentaire, tandis que les fragments d'Okazaki du brin tardif vont mettre plus de temps à s'apparier avec leur complémentaire. Celui-ci est donc sujet à des désaminations en excès, qui vont, après correction par le système de réparation de l'ADN, induire des mutations C $\rightarrow$ T. Ces mutations, en faisant baisser [C] et en augmentant [T] dans le complémentaire du brin tardif – donc le brin précoce de la génération suivante –, l'enrichissent effectivement en G relativement à C et en T relativement à A.

Ces recherches ont également mis en avant un autre biais auquel seraient potentiellement soumis les génomes, en montrant que la désamination des cytosines pouvait également avoir lieu lors de la transcription, à cause de l'exposition prolongée durant cette opération du brin codant sous une forme simple brin. Ce biais peut *a priori* être isolé du précédent, car il affecte de la même façon les séquences sur le brin précoce et sur le brin tardif, et par contre n'affecte que les séquences codantes. Son importance relativement à l'usage de codons est donc théoriquement séparable de celle du biais lié à la réplication. Cependant, la propension, pour beaucoup d'organismes, à avoir une majorité de leurs gènes sur le brin précoce (et donc, à l'équivalence pour eux du brin précoce et du brin sens), combinée à la faible quantité de séquences intergéniques, a fait supposer que le contenu enrichi en GT du brin précoce pouvait être dû à des désaminations de cytosines ayant lieu à la fois durant la réplication et durant la transcription. L'influence de la transcription sur le biais d'usage de codon d'un génome a d'ailleurs été mise en évidence dans le cas du phage T4 (Kano-Sueoka et al., 1999).

Les multiples biais mutationnels auxquels sont soumis les génomes, dont nous n'avons présenté ici que les principaux, peuvent donc avoir une forte influence sur l'usage de codons des bactéries. C'est par exemple le cas de *B. burgdorferi*, dont l'usage du code est complètement dominé par le biais causé par la réplication, avec deux gammes de fréquences d'usage de codons bien différenciées sur les deux brins (McInerney, 1998). On peut donc, dans certains cas, uniquement à l'aide de mécanismes mutationnels et des biais compositionnels qu'ils entraînent, comprendre l'usage du code de certaines bactéries. Ce n'est cependant pas le cas pour toutes, et les causes sélectives du biais d'usage de codons sont présentées dans la suite.

3.4.2 Robustesse et évolutivité

Une théorie sur la nature de la sélection agissant sur le biais de codons, qui s'est développée ces dernières années, dit que l'usage du code pourrait être sélectionné pour des questions de robustesse. Cette théorie est le parallèle de celles développées sur le code génétique et sa robustesse face aux mutations : tout comme la structure du code semble optimisée pour que des mutations ponctuelles aient peu d'effet sur un organisme, un usage prioritaire des codons les plus stables, avec les mêmes critères, pourrait renforcer la robustesse de organismes. En pratique, cela nécessite l'estimation d'une mesure de robustesse pour chaque codon, mesure qui doit quantifier la gravité au niveau phénotypique d'une ou plusieurs mutations dans le codon en question. Une mutation synonyme, bien sûr, n'a aucune incidence, mais on peut imaginer que pour plusieurs codons synonymes, la même mutation ait des conséquences tout à fait différentes. Par exemple supposons que deux codons synonymes de la leucine, CTA et CTG, subissent une mutation à la première base C→A. Pour le premier codon, cette mutation a pour effet sur la protéine une mutation Leu → Ile, tandis que pour le second la leucine est remplacée par une méthionine. L'isoleucine et la méthionine sont deux résidus hydrophobes, mais la présence de soufre dans la méthionine la rend potentiellement beaucoup plus réactive (voir par exemple Levine et al. (1996)). Si une telle réactivité peut nuire à la stabilité de la protéine formée, on peut émettre l'hypothèse – sujette à caution – que l'emploi du codon CTA va être préféré à celui de CTG. On peut quantifier la différence induite par la mutation au niveau protéique en utilisant une distance entre les acides aminés (McLachlan, 1971), et en moyennant la distance moyenne des mutants au codon originel. Cependant, les résultats des analyses sont contradictoires (Archetti, 2004a,b; Marquez et al., 2005; Rocha, 2006). De plus, ces méthodes présentent de nombreux problèmes méthodologiques :

- Les méthodes basées sur une distance à une seule mutation ne peuvent s'appliquer que sur très peu de codons. En effet, de nombreux codons synonymes ont accès aux mêmes acides aminés après une mutation ponctuelle, à cause de la structure du code. Ceci implique qu'ils auront le même score dans n'importe lequel de ces modèles, et ne pourront avoir d'impact sur la robustesse de la protéine. Tous les codons dégénérés deux fois, à l'exception de la lysine (pour laquelle le codon AAG peut muter en méthionine, alors que le codon AAA ne peux pas), voient le même paysage mutationnel en terme d'acides aminés. Pour les acides aminés 4 fois dégénérés, les codons finissant par A et G, ou C et T, ne peuvent muter que vers deux codons synonymes, sauf dans le cas de la thréonine, avec le même mécanisme que pour la lysine. Ceci réduit drastiquement le pouvoir de discrimination entre codons de ces analyses.
- Un autre problème dans ces méthodes est de savoir quel score de similarité est donné aux mutations vers un codon stop. A priori toute mutation vers un stop doit être très délétère, mais ceci ne peux pas être mesuré par les différences entre propriétés d'acides aminés. Bien qu'il ait été montré (Archetti, 2004a) que l'importance du score donné aux codons stop puisse être négligeable, ceci reste néanmoins un problème théorique.
- Les scores obtenus par chaque codon dans ce type de mesure ne dépendent pas de l'organisme, uniquement du code génétique et des propriétés des acides aminés. Le biais d'usage du code étant différent selon les espèces, cette constatation pose le problème de l'importance réelle de la robustesse dans le biais de codons.

- Le coût énergétique, et la disponibilité des 20 acides aminés dans différents organismes peuvent moduler les taux de mutations réels entre acides aminés. Une mutation entre deux acides aminés très similaires peut être fortement contre-sélectionnée, car l'acide aminé cible est très rare pour l'organisme en question, ou nécessite des étapes de biosynthèse supplémentaires.
- La définition de la similarité entre acides aminés nécessite de nombreuses précautions. Premièrement, sachant que dans beaucoup de protéines, un changement local d'acide aminé n'altère pas la fonction (Beyer, 1997), on peut se demander si la similarité à l'échelle de l'acide aminé représente réellement une propension à garder la même fonction. Au contraire, sur certains sites, la similarité compte peu devant le fait de garder certains acides aminés précis, sous peine de perdre la fonction (par exemple, sur des sites de surface impliqués dans des interaction protéines-protéines, ou sur un site catalytique).
- Finalement un problème fondamental d'interprétation se pose : les codons pouvant muter vers des acides aminés très similaires sont-ils des codons transitoires, sélectionnés pour changer[1] ? Et à l'inverse, les codons pour lesquels toute mutation est très délétère sont-ils nécessairement soumis à une sélection purificatrice ? Ces hypothèses, développées récemment (Plotkin et al., 2006a,b, 2004), ont été soumises à de très fortes critiques (Dagan and Graur, 2005; Friedman and Hughes, 2005; Sharp, 2005; Stoletzki et al., 2005; Zhang, 2005).

3.4.3 Sélection des codons synonymes et traduction

Les autres hypothèses sélectives sur le biais d'usage de codons sont toutes basées sur l'influence du choix du codon sur le processus de traduction. L'acide aminé codé est le même, mais les détails du processus de son insertion dans le peptide ne sont pas forcément équivalents pour tous les codons. De ce point de vue, on peut considérer le codon comme transportant deux informations : l'acide aminé qui va être ajouté à la protéine naissante, et des détails sur la façon dont cet ajout doit avoir lieu. Ainsi, il n'y a perte d'aucune information, et le choix du codon devient un facteur important.

Une observation complémentaire faite à l'époque, sur les séquences disponibles, est que le contenu en ARNt dans la cellule est également biaisé, de telle sorte que les ARNt les plus courants sont ceux qui peuvent s'apparier avec les codons majeurs (Gouy and Gautier, 1982; Gouy and Grantham, 1980). Cette corrélation a été explicitement étudiée par la suite chez *E. coli*, mais également chez la levure *Saccharomyces cerevisiae* (Ikemura, 1981a,b, 1982). Ces travaux ont été confirmés par la suite grâce à des analyses à plus grande échelle sur les génomes procaryotes (Dong et al., 1996; Kanaya et al., 1999). Chez les organismes multicellulaires, par contre, les liens entre usage du code et concentration en ARNt sont moins clairs (Ikemura, 1985). Chez les phages, il a été observé que l'usage de codons correspondait au contenu cellulaire en ARNt de l'hôte, ce qui est cohérent puisque les phages, de manière générale, voient leurs gènes traduits par le système cellulaire de leur hôte (Sharp et al., 1985). Il a même été montré sur le phage T4 – qui possède 8 ARNt dans son génome – que les gènes précoces ont un usage du code équivalent à celui de leur hôte, alors que les gènes tardifs ont un usage légèrement biaisé de façon à mieux correspondre

[1] Une telle sélection pour des propriétés futures n'est pas envisageable, mais dans le cas d'une grande fréquence d'erreur lors de la traduction, une telle sélection effective pour des codons situés "loin" – en terme de nombre de mutations – de ceux codant pour des acides aminés différents, est possible.

au contenu en ARNt de T4. Ceci peut correspondre à un mécanisme adaptatif : au début de la phase d'infection, le contenu de la cellule en ARNt est celui de l'hôte, et les gènes précoces de T4 sont adaptés à cet environnement. À la fin de la phase d'infection, le contenu en ARNt a changé, puisque ceux codés par T4 ont été exprimés, et les gènes tardifs de T4 sont également adaptés de façon précise à son nouvel environnement (Cowe and Sharp, 1991).

Quel bénéfice y-a-t-il a avoir un usage de codons corrélé à son contenu en ARNt ? Il a été très tôt supposé que des différences d'abondance des ARNt pourraient être une contrainte sélective sur les codons synonymes. Plusieurs hypothèses principales sur la nature de cette sélection ont été émises.

a) Sélection sur la liaison codon-anticodon

La première hypothèse consiste à dire que les séquences codantes sont ajustées de manière à n'employer que certains codons, en fonction de leurs interactions avec l'anticodon de leur ARN de transfert (Grosjean and Fiers, 1982). L'idée est que les interactions codon-anticodon trop fortes et trop faibles sont contre-sélectionnées, respectivement parce que conduisant à un trop grand taux de rejet de l'ARNt par le ribosome (Andersson et al., 1986), ou amenant trop d'erreurs traductionnelles. Une expérience qui soutient cette hypothèse est la différence observée du temps nécessaire à la traduction de deux codons synonymes par le même ARNt (Sorensen and Pedersen, 1991; Thomas et al., 1988) : le codon GAG est traduit 3 fois plus lentement que son synonyme GAA, alors qu'ils sont reconnus par le même ARNt. Ceci implique une différence fondamentale entre les codons au niveau du complexe ribosomal et du lien avec l'anticodon. Les codons les plus fréquents, donc correspondants aux ARNt dont la concentration est la plus élevée, devraient respecter une homéostasie d'énergie de liaison avec leur anticodon (Bennetzen and Hall, 1982). En pratique, cela signifie que les codons utilisés dans les gènes fortement exprimés ne doivent en général pas utiliser de paires de G et C côte à côte, car les deux bases pourraient interagir pour renforcer les liaisons codon-anticodon durant la traduction. De plus, parmi ces codons, ceux terminant par C et T doivent être préférés aux autres pour les acides aminés 4 fois dégénérés, et les codons terminant par C doivent l'être pour les acides aminés deux fois dégénérés[1]. Si on le combine aux règles de reconnaissance floue, ce modèle implique des règles très strictes sur les codons employés. Son défaut est qu'il implique un usage de codons similaire pour tous les organismes, ce qui n'est pas observé ; cependant des analyses à grande échelle récentes ont montré que ce type de sélection correspondait bien à une grande partie des gènes fortement exprimés de nombreux génomes (Rocha, 2004).

b) Modélisation du temps d'attente des ribosomes durant la phase d'allongement

Un autre modèle, basé plus explicitement sur les abondances des ARNt dans la cellule, a été beaucoup utilisé. En modélisant les ARNt comme diffusant librement dans la cellule, on trouve que les codons correspondants aux ARNt les plus nombreux sont

[1] On peut mettre ceci en relation avec le fait que les seuls ARNt employés pour décoder la phénylalanine, la tyrosine et l'asparagine ont un anticodon commençant par G. Voir la section "Règles de reconnaissance floue" page 66.

en moyenne traduits plus vite (Bulmer, 1987b). Des observations du temps de traduction pour différents codons confirment aussi ce modèle, en montrant que la vitesse de traduction des codons usuels, pour lesquels beaucoup d'ARNt sont présents dans la cellule, est de 12 acides aminés par seconde, contre 2 acides aminés par seconde pour les codons rares (Sorensen et al., 1989). Dans ce cas l'accent est mis sur le fait que les concentrations des ARNt reconnaissant les deux types de codons sont différentes *in vivo*, et pas sur les propriétés intrinsèques de chaque codon. Ce modèle, qui est encore employé aujourd'hui (voir par exemple un article très récent, Man and Pilpel (2007)), dit que le temps nécessaire pour traduire chaque codon est inversement proportionnel à la concentration des ARNt capables de le lire. Cette hypothèse n'est justifiée que si le temps caractéristique d'attente de l'ARNt au niveau du ribosome est significativement plus long que le temps nécessaire pour créer le lien codon-anticodon et éventuellement rejeter un ARNt ne correspondant pas au codon en face du site A. Des expériences ont montré que c'est effectivement le cas, et que l'augmentation du nombre de "tests" avant la fixation au site A d'un ARNt complémentaire – réalisée par l'augmentation de la concentration des ARNt non complémentaires au codon devant être traduit – n'allonge pas la durée de la traduction (Bilgin et al., 1988).

Ce mécanisme de modulation du temps d'attente au niveau du ribosome pour chaque codon est utilisé dans les modèles de minimisation des risques de traduction erronée, et dans ceux de sélection pour la vitesse de traduction. Nous allons les présenter en détails ici, car ils constituent les deux hypothèses qui ont fait couler le plus d'encre sur la cause sélective du biais de codons.

c) Précision de la traduction

L'hypothèse selon laquelle certains codons sont préférentiellement utilisés par les organismes pour améliorer la précision de la traduction a été émise. Dans une formulation basée sur l'abondance relative des ARNt (Bulmer, 1991), l'usage des codons est ajusté aux concentrations des ARNt pour minimiser la probabilité qu'un codon soit reconnu par erreur par un ARNt non complémentaire et qu'un acide aminé inadéquat soit inséré dans la protéine. Ceci ne nécessite pas le choix de codons particuliers, seulement l'ajustement des abondances relatives des ARNt à celles des codons qu'ils reconnaissent – modulée par l'affinité entre les ARNt et les codons – qui minimise les temps d'attente pour un bon ARNt au niveau de chaque codon. En effet le coût énergétique dû à l'insertion d'un acide aminé inadéquat dans une protéine, ou pire à la terminaison précoce de sa synthèse, peut être significatif à l'échelle de la cellule. Cette façon de penser a de plus permis de comprendre une tendance observée chez les procaryotes, à savoir que le biais d'usage de codons est corrélé à la longueur des gènes. Ceci s'explique naturellement dans un cadre de sélection pour la précision : la perte d'énergie lors de la traduction avortée d'un ARN messager est proportionnelle à la longueur qui a déjà été traduite, et donc on s'attend à ce que le biais de codons soit beaucoup plus fort à la fin d'un gène qu'au début (Gilchrist and Wagner, 2006). Cette interprétation est intéressante, car elle est en opposition avec une vision basée sur la génétique des populations, qui dit que la longueur d'un gène est inversement reliée à la sélection sur chaque site (voir Akashi (2001) ou Kreitman and Comeron (1999) pour des revues sur le sujet). Mais la variation du biais de codon avec la longueur des gènes n'est pas la même pour tous les organismes, et chez *Drosophila melanogaster*, la mouche des fruits, il y a une corrélation négative entre biais de codons

et longueur des gènes (Moriyama and Powell, 1998).

Ces hypothèses sélectives du biais d'usage de codons basées sur la précision de la traduction semblent se vérifier sur certaines catégories de gènes et pour certains organismes, en particulier eucaryotes. Elles semblent particulièrement adaptées pour mesurer le biais de codons sur des sites fonctionnellement sélectionnés, pour lesquels l'acide aminé codé est essentiel à la fonction protéique (Akashi and Eyre-Walker, 1998; Stoletzki and Eyre-Walker, 2007).

d) Efficacité de traduction

Un autre cadre théorique a été proposé en parallèle des modèles de sélection pour la précision, celui d'une sélection pour l'efficacité ou pour la vitesse de traduction. Cette hypothèse de régulation de la vitesse de traduction par la concentration des ARNt est divisée en deux gammes de modèles.

Limitation du taux de traduction par le taux d'amorçage La première est celle des modèles se basant sur l'hypothèse que c'est l'amorçage de la traduction qui est le facteur limitant en temps, et que la durée de la traduction est majoritairement due au temps passé par l'ARNm à attendre qu'un nouveau ribosome ne vienne débuter la traduction sur son extrémité 5'. Dans ce cas, le taux de synthèse protéique dans l'ensemble de la cellule est affecté de plusieurs manières par le biais d'usage de codons. Un premier lien entre l'usage du code et la vitesse de lamorçage est lié à la transcription : si les bases employées pour fabriquer l'ARNm lors de la transcription sont peu courantes dans la cellule, on observera un faible taux de transcription, ce qui renforcera le problème du taux global de traduction, les ARNm étant moins nombreux. L'usage de codons peut, notamment au niveau de la troisième base, faire varier significativement la demande dans les différents nucléotides ; il devrait donc refléter le contenu cellulaire dans les différentes bases (Xia, 1996). Ensuite, pendant la traduction, l'emploi de codons traduits rapidement diminue le temps passé par les ribosomes sur l'ARNm, et augmente donc indirectement la concentration des ribosomes libres dans le cytoplasme. Ensuite, au niveau des premiers codons de l'ARNm, on observe l'effet contraire : des codons traduits lentement bloquent les ribosomes longtemps, et empêchent d'autres ribosomes de venir se fixer, augmentant également la concentration des ribosomes libres, en contrepartie d'une légère diminution de la vitesse de traduction du gène en question. On s'attend donc, si les ribosomes sont limités en nombre relativement aux ARNm, à ce qu'un biais d'usage de codons avec des codons rares en tête de gène et des codons rapides après facilite globalement la synthèse protéique à l'échelle de la cellule (Bulmer, 1991). Une telle partition du biais de codons a été observée (Liljenstrom and von Heijne, 1987). Par la suite, cependant, il a été remarqué que le biais d'usage de codons en tête des gènes était similaire quel que soit le niveau d'expression du gène (Bulmer, 1987a), prévenant toute forme de régulation dépendant spécifiquement du niveau d'expression. D'autres expériences soutiennent cependant l'hypothèse d'un taux de traduction limité par l'amorçage : par exemple le fait que les ribosomes présents sur un ARNm sont en moyenne écartés de près de 225 bases (Ingraham et al., 1983), alors qu'ils ne recouvrent que 30 bases de l'ARNm. Si le temps nécessaire à l'amorçage de la traduction était un facteur négligeable, on s'attendrait à avoir des ribosomes proches les uns des autres sur l'ARNm, ce qui n'est pas le cas.

Limitation du taux de traduction par le temps d'allongement La deuxième catégorie de modèles est celle qui suppose que la phase d'allongement est longue devant les autres étapes du processus de traduction, et que c'est le temps nécessaire à la traduction de chaque codon qui domine le temps total. Dans ce cas, l'explication du lien entre concentration en ARNt et usage de codons est très simple : le rapport entre la quantité d'ARNt dans la cellule et la quantité de codons que doivent traduire les ARNt est optimisé pour accélérer le processus de traduction en minimisant les temps d'attente en moyenne sur tous les codons. Plusieurs modèles ont calculé un tel rapport, et ont montré que le biais dans le contenu en ARNt doit être moins fort que le biais de codons (Bulmer, 1987b; Sharp et al., 1986). Précisément, en écrivant f_i^a la fréquence d'usage de codon i, et t_i^a le nombre d'ARNt reconnaissant i dans la cellule, pour deux codons synonymes i et k codant pour l'acide aminé a, la relation minimisant le temps total d'attente au niveau de l'ARNm pour ces deux codons est :

$$\sqrt{\frac{f_i^a}{f_k^a}} = \frac{t_i^a}{t_k^a}. \tag{3.8}$$

Ces modèles ont également montré comment un tel équilibre pouvait être obtenu, en supposant un usage de codon fixe et une population d'ARNt en évolution. Les modèles opposés, qui font évoluer le biais de codons à population d'ARNt fixée, trouvent un résultat différent : la traduction est plus rapide dans ce cas si, pour chaque acide aminé, seul le codon correspondant aux ARNt les plus représentés dans la cellule est employé (Xia, 1998).

Une remarque est à garder à l'esprit dans le cadre des modèles de sélection du biais de codons basée sur la vitesse : c'est le fait qu'un ARNm traduit vite ne sera pas nécessairement traduit plus souvent, et que c'est pourtant cette deuxième propriété qui est recherchée pour les gènes fortement exprimés. En effet, si beaucoup de gènes sont exprimés en parallèle (comme c'est le cas en milieu pauvre), le ribosome, après avoir synthétisé une protéine d'un l'ARNm particulier $\mathcal{A}$, va se dissocier et très probablement aller s'attacher sur un ARNm différent[1]. Au final, il n'y aura eu qu'une protéine correspondante à l'ARNm $\mathcal{A}$ de produite, ceci quelle que soit la vitesse de traduction de ce dernier : l'avantage gagné sur la concentration en ribosomes libres après la traduction est partagé entre tous les ARNm traduits au même moment (Andersson and Kurland, 1990; Kurland, 1991). Par contre, si la cellule est dans une phase de croissance en milieu riche, elle ne produit que très peu de protéines différentes. Dans cette situation, l'accélération de la traduction au niveau d'un ARNm $\mathcal{A}$ permet d'augmenter significativement le taux de production des protéines $\mathcal{A}$. Dans ce cas, le biais de codons permet une accélération significative de la synthèse protéique dans son ensemble, et ceci explique pourquoi ce biais est plus fort dans les gènes fortement exprimés : ce sont eux, et pratiquement eux seuls, qui s'expriment quand l'organisme est en phase exponentielle de croissance.

Pour terminer, nous mentionnerons simplement qu'il est difficile de différencier la sélection sur la précision de celle sur la vitesse, puisque toutes deux prédisent la relation

[1]Ce qui n'est pas le cas chez les eucaryotes, à cause du phénomène connu sous le nom de recyclage ribosomal : grâce à la structure circulaire des ARNm eucaryotes, les ribosomes, à la fin de la traduction, se retrouvent proches d'une position où ils peuvent commencer une nouvelle traduction sur le même ARNm. Voir Chou (2003) pour une belle modélisation de ce processus.

observée entre concentration cellulaire en ARNt et usage de codons. Il a été tenté de les réconcilier (Solomovici et al., 1997), ou d'expliquer le biais d'usage de codons par d'autres facteurs, comme des contraintes sur le repliement des ARNm (Jia and Li, 2005). Au vu des données génomiques actuelles, il est seulement possible de dire que le biais observé sur les génomes a une origine multifactorielle, une seule pression de sélection ne pouvant tout expliquer (Dethlefsen and Schmidt, 2005), et que chaque modèle étant corroboré par des observations sur au moins un organisme, il est difficile de généraliser les conclusions de chaque étude.

3.5 Le paradoxe des codons rares

Une question qui se pose naturellement, si on se base sur une hypothèse sélective pour expliquer la présence des codons majeurs, est de savoir pourquoi des codons rares sont néanmoins employés. Dans le cadre des modèles de sélection pour la précision, on peut se demander pourquoi certains gènes tolèrent mieux les erreurs que d'autres. Pour les modèles de sélection pour la vitesse, l'intérêt d'être traduit lentement n'est pas évident au premier abord. En plus des séquences de codons rares observées en tête de gène (Bulmer, 1987a; Liljenstrom and von Heijne, 1987), l'emploi de codons rares a lieu de façon homogène dans le génome. Sachant les risques encourus de ne pas parvenir à synthétiser la protéine codée si le ribosome reste bloqué trop longtemps sur l'un d'eux, à cause des ARNtm par exemple, ou encore d'une erreur d'insertion due à une surabondance d'ARNt non complémentaires (Bilgin et al., 1988), on peut se poser la question d'un éventuel rôle des codons rares dans le processus de traduction.

3.5.1 Le ralentissement de la traduction

En effet, il est aisé de passer de l'idée que les codons majeurs accélèrent la traduction à celle que les codons rares la ralentissent. Ici, il est nécessaire d'être très précautionneux dans le choix des termes. De la même façon que l'on a montré au paragraphe précédent que les codons majeurs n'accéléraient substantiellement la traduction que lors de la croissance sur milieu riche, on peut supposer que les codons rares ne la ralentissent pas énormément (Andersson and Kurland, 1990; Kurland, 1991; Sharp and Li, 1986). Le seul effet lors de la traduction d'un codon rare est la pause au niveau du ribosome, qui en moyenne doit attendre plus longtemps qu'un ARNt correspondant lui parvienne, en supposant que les ARNt diffusent librement dans la cellule. Au final, la protéine sera traduite quand même, et le seul cas où le taux de production protéique total serait affecté serait si l'ARNm correspondant à cette protéine devait être traduit de très nombreuses fois, donc dans le cas d'un gène fortement exprimé. Or la présence de codons rares dans des gènes fortement exprimés a très rarement été observée. L'hypothèse que des codons rares puissent être soumis à une sélection positive, afin de réduire le niveau d'expression protéique – par exemple dans des gènes régulateurs dont le nombre de protéines dans la cellule doit être très contrôlé – a été défendue durant les années 80. Mais l'augmentation de la taille des jeux de données a permis de montrer que le taux de substitutions silencieuses dans les codons rares était aussi élevé que dans les codons normaux, et beaucoup plus fort que dans les codons majeurs : il n'y a donc pas de sélection globale au niveau des codons rares. De plus, certaines observations (McNulty et al., 2003; Spanjaard and Duin, 1988)

sur les effets de codons rares successifs au niveau du ribosome, ont montré que la présence de codons rares avait plus souvent pour effet de provoquer des erreurs ou des décalages du cadre de lecture que de ralentir la traduction. Ceci implique que les codons rares ne peuvent pas être utilisés pour réguler la vitesse de la synthèse protéique. Par contre, le décalage du cadre de lecture peut être utile dans certains cas particuliers, permettant de synthétiser une protéine différente, et procurant un joli mécanisme de décodage alternatif de l'ARNm (Craigen and Caskey, 1986). L'augmentation du taux d'erreur traductionnelles lors de la synthèse d'une protéine codée par un ou plusieurs codons rares reste cependant problématique, mais est à mettre en perspective avec la faible fréquence des séquences de codons rares successifs dans les organismes.

3.5.2 Pause et repliement

Une autre hypothèse intéressante a été développée : celle que certains codons rares, placés à des endroits bien particuliers, pourraient avoir pour rôle de forcer une pause lors de la traduction. Pendant cette pause, le peptide déjà formé pourrait se replier dans une forme fonctionnelle, ce qui serait peut être plus difficile une fois toute la protéine traduite. Ceci permet de prédire que l'on devrait trouver des codons rares à des emplacements dans les séquences correspondants à des jonctions entre domaines fonctionnels protéiques, ce qui a été observé (Thanaraj and Argos, 1996a,b). Bien que l'analyse ait porté sur un jeu de données restreint, il a été montré que, chez *E. coli*, les emplacements des codons rares étaient fortement corrélés avec les positions des liens entres domaines protéiques, et même que ces codons rares étaient placés plutôt en 3' par rapport au domaine sur la séquence d'ARNm, permettant effectivement de supposer que le domaine se replie pendant que le ribosome est bloqué sur le codon rare. Cette observation est renforcée par le fait que la protéine chaperon ubiquitaire GroEL, qui aide au repliement protéique après la traduction, a été observée agissant directement sur le polypeptide en sortie du ribosome, et donc peut être avant la fin de la traduction.

De plus, dans un autre travail publié la même année par la même équipe, une analyse des liens entre usage de codons et structure protéique a révélé un autre résultat intéressant. Les zones traduites lentement (et donc celles qui contiennent le plus de codons rares) correspondent au niveau structurel à des feuillets β, et ceci même en éliminant les corrélations qui pourraient être dues à un usage d'acides aminés particulier. Or, lors du repliement protéique, la stabilisation des feuillets β a lieu après celle des hélices α : l'usage de codons rares permet donc peut-être que la traduction soit plus lente au niveau des séquences des feuillets β qu'au niveau de celles des hélices α, laissant le temps à ces dernières de se former avant de poursuivre la traduction de l'ARNm.

3.5.3 Usage dépendant du taux de croissance

Les modèles précédents nous donnent une relation entre la fréquence d'usage de chaque codon, et la concentration relative des ARNt correspondants. Or il est connu que la concentration en ARNt de chaque type dans la cellule est fortement dépendante du taux de croissance de la cellule, et des conditions extérieures dans lesquelles elle se trouve. Des études (Dittmar et al., 2005; Elf et al., 2003; Sorensen et al., 2005) ont montré que lors de situations de carence, certains ARNt voyaient leur concentration chuter, alors que d'autres ARNt isoaccepteurs gardaient un niveau d'expression constant. Ceci avait déjà

été observé dans les années 60 (Bock et al., 1966; Morris and DeMoss, 1965; Yegian and Stent, 1969) sans être expliqué. Ces travaux donnent un point de vue complètement neuf sur l'usage des codons, et modifient les notions de codons rares ou majeurs en codons sensitifs ou insensitifs aux carences.

On note respectivement t_i et f_i, le nombre d'ARNt reconnaissant le codon i, et la fréquence d'emploi du codon i relativement à ses synonymes, et α_i la fraction relative d'ARNt chargés dans la cellule. On suppose que, à l'équilibre, le taux de chargement des ARNt de chaque espèce est proportionnel au taux d'utilisation par les complexes ribosomaux de ces ARNt. Si on prend l'exemple simple de deux codons synonymes reconnus par deux ARNt différents, la condition d'équilibre précédente s'écrit comme l'égalité des rapport des concentrations des ARNt chargés et de leur taux d'utilisation :

$$\frac{(1-\alpha_1)t_1}{(1-\alpha_2)t_2} = \frac{f_1}{f_2}. \tag{3.9}$$

Cette équation représente simplement l'égalité des taux de deux processus chimiques, le chargement par les synthétases et le déchargement par le complexe ribosomal, pour deux ARNt chargés du même acide aminé. Elle permet de voir que, si l'on soumet la cellule à une carence dans l'acide aminé qui devrait être chargé sur ces ARNt, α_1 et α_2 ne tendent pas vers zéro à la même vitesse. L'un des deux ARNt va voir la concentration de ses éléments chargés chuter à zéro, tandis que pour l'autre elle restera non nulle. L'ARNt qui va voir sa concentration d'éléments chargés diminuer le plus est celui pour lequel le ratio t_i/f_i est le plus faible au départ. Si on se replace dans le contexte des modèles précédents, et que l'on suppose que la cellule est optimisée pour un usage de codons permettant de minimiser les temps d'attente au niveau du ribosome, alors on respecte la condition $t_1/t_2 = \sqrt{f_1/f_2}$. Dans ce cas, on voit que si la concentration en ARNt 1 est la plus importante (donc $t_1 > t_2$, et le codon 2 est le codon rare), on trouve $t_1/f_1 < t_2/f_2$. C'est donc l'ARNt reconnaissant le codon majoritaire, celui-la même qui est employé par les protéines essentielles, qui va voir sa concentration chuter à zéro en cas de carence, dans ce modèle. Et l'ARNt correspondant au codon rare qui va continuer à être exprimé, même si c'est à un taux plus faible. Donc une cellule optimisée pour accélérer sa traduction sur milieu riche l'est *naturellement* pour changer de codons majeurs en cas de carence. Ce changement peut avoir de nombreuses conséquences, dont la première est la diminution du taux de production des protéines ayant le biais de codons des gènes fortement exprimées, par exemple les protéines ribosomales. En effet, la cellule en situation de carence doit économiser ses acides aminés, et ne peut se permettre de faire augmenter son stock de ribosomes : elle doit avant cela synthétiser l'acide aminé manquant en quantités suffisantes pour rentabiliser l'investissement représenté par la fabrication de nouveaux ribosomes.

D'ailleurs, il a également été observé que les voies biosynthétiques des acides aminés, qui ne s'expriment qu'en situation de carence, ont leurs gènes codés majoritairement par des codons insensitifs, qui sont certains des codons "rares". Ces gènes sont donc traduits sans difficulté lors des situations de carence, au moins du point de vue des ARNt : leur usage de codons est optimisé de la même manière que celui des protéines fortement exprimées, mais de façon ajustée au contenu en ARNt réel que les ARNm voient au moment de leur expression, et pas celui qui est observé en phase exponentielle de croissance. De plus, cette hypothèse permet également d'expliquer le choix des codons employés dans les séquences promotrices des opérons des voies biosynthétiques : ce sont les codons les plus sensibles, donc ceux qui disparaissent le plus vite en cas de carence. Ce double usage

des codons permet aux opérons biosynthétiques d'être transcrits en cas de carence uniquement – grâce aux mécanismes d'atténuation transcriptionnelle décrits au chapitre 2 – tout en permettant que les gènes eux-mêmes soient codés avec des codons pour lesquels les ARNt chargés ne vont pas manquer.

Finalement, l'analyse de la séquence messager des ARNtm montre qu'elle est codée par des codons insensitifs. En cas de brusque carence dans le milieu, la traduction des molécules coûteuses va s'arrêter, nécessitant des ARNtm pour débloquer les ribosomes. Et l'usage de codons de la séquence messager des ARNtm est optimisé par rapport au contenu réel en ARNt dans lequel les ARNtm baignent lorsqu'ils sont utilisés. De ce point de vue, les ARNtm peuvent être considérés comme une réponse à un facteur de stress, la carence en un acide aminé.

Deuxième partie

Classification et théorie de l'information

Introduction

La classification s'intéresse au problème de la réduction des jeux de données. Le but de toute classification est de regrouper les données dans un ensemble de groupes de façon à ce que les données situées dans le même groupe soient les plus similaires possibles. Un exemple trivial est de partitionner la liste {rose, Marie, chien, chat, éléphant, Emmanuel, bleu, dromadaire, vert, Paul} en sous-listes. On obtient intuitivement :

- Marie, Paul et Emmanuel – les prénoms
- chien, chat, éléphant, dromadaire – les animaux
- vert, bleu, rose – les couleurs

Mais partitionner de façon automatique d'énormes jeux de données est un problème complexe. Rien que sur cet exemple simple, on peut identifier plusieurs des éléments nécessaires à toute classification. Il faut savoir, avant de classer, le nombre des groupes dans lesquels on veut séparer les données, et le critère de similarité que l'on veut employer. Dans notre exemple, tous deux ont été choisis par le lecteur *a posteriori*, après analyse des données. Le choix du critère de partition, même s'il paraît intuitif dans cet exemple, est une question complexe, car le critère de partition doit être objectif. On peut par exemple se poser la question, pourquoi "rose" est classé comme une couleur et pas comme un prénom ? L'absence d'une majuscule est dans ce cas un critère bien défini permettant de choisir. Mais comment classer "saumon" avec les critères précédents ? Est-ce un animal ou une couleur ?

Les usages de la classification sont multiples. En informatique, les techniques de classification servent à compresser les données, les images par exemple, comme on le verra au chapitre 5. Un autre usage, connu de tous, est l'"antispam", qui trie les courriers électroniques et infère de leur intérêt, les classant en deux catégories.

En statistiques, les données sont regroupées de façon à limiter le nombre de catégories afin de faciliter la compréhension et d'augmenter le pouvoir statistique à l'intérieur de chaque classe. La classification permet d'abréger la description nécessaire pour donner le sens des mesures statistiques faites sur, par exemple, 1000 personnes, sans avoir à décrire les particularités de chacun.

En biologie, et tout particulièrement en génomique, les techniques de classification ont de multiples usages, dûs à la grande taille des jeux de données utilisés. Les applications sont multiples, comme la classification de protéines par leur composition en acides aminés, ou la classification de données d'expression par leur similarité, dans le but d'identifier des gènes ayant les mêmes régulateurs. Une autre des applications les plus connues des techniques de classification en biologie est la réalisation de classifications phylogénétiques, retraçant l'évolution des organismes et permettant d'établir des liens de parenté entre espèces. C'est une forme de classification hiérarchique qui est employée dans ce cas. Une autre application est l'analyse de données d'expression. En effet, l'étude de données d'expression de gènes

conduit naturellement à vouloir identifier des groupes de gènes qui s'expriment ensemble, et donc à classer les données en groupes. De plus, dans ce cas, la méthode de classification employée doit être robuste et tenir compte du bruit qui est inévitablement présent dans les données.

Chapitre 4

Classification

4.1 Bases théoriques

Pour étudier la classification, nous allons tout d'abord exposer les bases théoriques qui la soutiennent, puis examiner un peu plus en détails les algorithmes les plus couramment utilisés. Prenons pour commencer un exemple formel. Supposons que l'on dispose d'un jeu de N objets, chaque objet ayant p caractéristiques. Ces caractéristiques peuvent être quantitatives (le nombre de segments sur une forme) ou qualitatives (la couleur de la forme). Pour simplifier le problème, nous allons considérer qu'on peut exprimer les variables qualitatives sous une forme numérique, en utilisant par exemple un code binaire de réponse à des questions qualitatives, comme "La forme est-elle bleue ?". Ceci nous permet de représenter chaque objet par un vecteur $\vec{x}_i$, $i = 1..N$ à p composantes dans $\mathbb{R}$, que l'on notera x_i^k, $k = 1..p$. Chaque objet peut être un point dans un espace p-dimensionnel, mais aussi n'importe quelle représentation à p dimensions de l'objet initial (image, expression d'un gène, texte...). On parlera un peu plus du problème de la représentation d'un objet à p dimensions par la suite, lors de l'étude de l'analyse des correspondances et de ses usages en classification.

Une classification des N objets est la partition des N vecteurs dans c classes. Cette partition peut être stricte – auquel cas on définit une fonction $f : \mathbb{R}^p \mapsto [1..c]$ qui à chaque vecteur associe la classe à laquelle il appartient – ou floue, c'est à dire que l'on évalue la probabilité pour chaque point d'appartenir à chacune des classes, $p(c|\vec{x}_i)$. Dans ce cas la normalisation $\sum_c p(c|\vec{x}_i) = 1$ doit s'appliquer pour chaque point.

Toute classification devrait tendre à regrouper entre elles les données les plus similaires, et à les isoler des autres. Mais la notion de similarité n'est pas une notion évidente, surtout en termes mathématiques. La définition d'un critère objectif de similarité a été l'objet de nombreuses publications, et nous allons présenter ici la discussion générale auxquelles elles ont donné lieu. Nous commencerons par présenter les problèmes associés à la définition de la similarité entre paires d'objets et du choix du nombre de classes. Par la suite, nous discuterons des mesures de dissimilarité entre groupes, et plus généralement de la caractérisation d'une classification, qui ne nécessite pas forcément l'emploi d'une mesure de similarité par paires, comme nous allons le voir.

4.1.1 Distances et similarité

Classer des objets en fonction de leur similarité implique d'avoir une mesure quantitative de cette similarité, ou inversement de leur dissimilarité d. C'est cette dernière qui est employée couramment, et dont nous allons détailler les propriétés. Nous allons également étudier les fonctions les plus courantes utilisées pour d. Sauf indication contraire, la discussion présentée ici est tirée de (Gordon, 1999).

Une mesure de dissimilarité est une fonction qui à deux objets $\vec{x}_i$ et $\vec{x}_j$, associe une valeur réelle, qui représente la différence des deux objets. Intuitivement, toute mesure de dissimilarité doit posséder certaines propriétés. Tout d'abord, la dissimilarité entre un objet et lui-même doit au moins être minimale :

$$d(\vec{x}_i, \vec{x}_i) \leq d(\vec{x}_i, \vec{x}_j) \, \forall i, j \in [1..N] \times [1..N], i \neq j. \tag{4.1}$$

Dans la pratique, on utilise des distances qui respectent la propriété suivante, plus contraignante :

$$d(\vec{x}_i, \vec{x}_i) = 0 \quad \forall i \in [1..N]. \tag{4.2}$$

Une autre propriété que doit respecter la dissimilarité découle des deux précédentes. La dissimilarité doit être définie positive :

$$d(\vec{x}_i, \vec{x}_j) \geq 0 \quad \forall i, j \in [1..N] \times [1..N]. \tag{4.3}$$

Finalement, comme une distance, la dissimilarité doit être symétrique :

$$d(\vec{x}_i, \vec{x}_j) = d(\vec{x}_j, \vec{x}_i) \quad \forall i, j \in [1..N] \times [1..N]. \tag{4.4}$$

Ces propriétés, intuitives, rendent la dissimilarité très proche d'une métrique. Mais la propriété d'inégalité triangulaire, si elle est souhaitable parfois, n'est pas nécessairement respectée. En fonction des données, il arrive parfois que cette inégalité représente mal les relations entre les objets. Nous la donnons néanmoins ici car elle est souvent employée :

$$d(\vec{x}_i, \vec{x}_j) + d(\vec{x}_j, \vec{x}_k) \geq d(\vec{x}_i, \vec{x}_k) \quad \forall i, j, k \in [1..N] \times [1..N] \times [1..N]. \tag{4.5}$$

L'égalité représentant "graphiquement" le cas où les trois points sont alignés.

Voici quelques exemples des dissimilarités les plus employées :

- Toutes les α-normes telles que $d(\vec{x}_i, \vec{x}_j) = \left(\sum_{k=1}^{p} w_k (x_i^k - x_j^k)^\alpha\right)^{1/\alpha}$, $\alpha \neq 0$, dont l'exemple le plus connu est la 2-norme, la norme euclidienne.
- Les mesures basées sur les corrélations entre éléments, par exemple $d(\vec{x}_i, \vec{x}_j) = \frac{1}{2}\left(1 - \frac{\sum_k (x_i^k - \bar{x}_i)(x_j^k - \bar{x}_j)}{\sqrt{(\sum_k (x_i^k - \bar{x}_i)^2)(\sum_k (x_j^k - \bar{x}_j)^2)}}\right)$, avec $\bar{x}_n = \frac{1}{p}\sum_k x_n^k$.
- La métrique de Canberra $d(\vec{x}_i, \vec{x}_j) = \frac{\|x_i - x_j\|}{\|x_i\| + \|x_j\|}$ où la norme est l'une des α-normes précédemment évoquées.

Les paramètres w_k dans la formule des α-normes sont les poids donnés à chaque dimension. En effet, on peut être amené à considérer que la dissimilarité entre deux objets dépend plus de leur différence sur quelques composantes précises que sur l'ensemble. Le choix des valeurs des $\{w_k\}$, $k = 1..p$ est un problème en soit.

Pour éviter d'avoir à choisir ces poids, une façon de normaliser les données, est de considérer des variables modifiées $\vec{y}_i$, $i = 1..N$, où la valeur de chaque composante k est

le z-score de la composante initiale x_i^k par rapport à la distribution de cette composante sur tous les objets. Cela donne, pour chaque composante y_i^k :

$$y_i^k = \frac{x_i^k - \frac{1}{N}\sum_{i=1}^{N} x_i^k}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i^k)^2 - (\frac{1}{N}\sum_{i=1}^{N} x_i^k)^2}}. \tag{4.6}$$

Ainsi, chaque composante y_i^k représente l'écart à la moyenne de la composante x_i^k correspondante, normalisé par l'écart-type de la distribution de cette composante. Cette méthode de normalisation ne peut s'appliquer que si les valeurs possibles de la variable k ont un spectre assez large. En particulier, pour une variable binaire, ou pour la représentation numérique d'une variable qualitative, il est préférable d'employer une approche basée sur des coefficients w_k biens choisis.

Hypothèse implicites et choix de la distance Un exemple simple permet de comprendre pourquoi le choix de la distance peut avoir des conséquences sur la classification. Pour regrouper des points dans un espace à deux dimensions, la norme euclidienne peut paraître tout à fait adaptée, mais des contre-exemples simples peuvent être trouvés. Supposons que l'on dispose de N points dans un plan, dont $\frac{N}{2}$ ont été générés le long de l'axe x, pris au hasard sur un intervalle $[0, a]$ et $\frac{N}{2}$ le long d'une autre horizontale, d'équation $y = e$, sur le même intervalle. Visuellement, on peut se retrouver dans trois situations différentes (voir Fig. 4.1.)

Dans les deux premiers cas, si $a \lesssim e$, un algorithme de classification utilisant la métrique euclidienne (la 2-norme), et cherchant 2 groupes, va correctement isoler les deux lignes, quoique déjà dans le deuxième cela dépende de la position exacte des points sur les horizontales. Mais si maintenant on se place dans la situation où $a \gg e$, alors le résultat va beaucoup dépendre de l'algorithme employé. En effet, le défaut de l'emploi de la 2-norme est l'hypothèse implicite que les données similaires sont regroupées dans des boules, et pas le long de lignes. Une classification qui peut être obtenue dans le cas $a \gg e$ est celle obtenue en bas de la figure 4.1. On voit que, dû au choix particulier de la distance, la classification obtenue n'est pas celle attendue. Ceci montre une notion très importante en classification : toute classification utilise implicitement un modèle sur la répartition des données. Ce modèle dépend à la fois de la dissimilarité d choisie, mais aussi du critère de partition (voir plus loin). Dans le cas des α-normes, le critère est que les groupes formés à la fin de la classification sont des boules. La seule *alpha*-norme qui pourrait être utilisée avec succès ici est la 1-norme telle que $d(\vec{x}_i, \vec{x}_j) = \|x_i^y - x_j^y\|$, où la composante y des $\vec{x}_i$ est la composante verticale. Mais l'emploi de cette distance est *a priori* contre-intuitif, et nécessite de connaître la classification attendue, ce qui réduit son intérêt. De plus, cet exemple permet de soulever un point important : celui de la subjectivité de la classification attendue. En effet, la figure en bas de 4.1 contient une information objective : celle que la séparation des points à gauche et à droite conduit à une minimisation de la distance de chaque point au centre de son groupe. Cette information doit être comparée soigneusement à la seule autre que nous possédons, à savoir la façon dont les groupes ont été générés, avant de pouvoir être rejetée comme ne correspondant pas à la structure des données. La classification donnée par un algorithme représente toujours un certain optimum, même si ce n'est pas celui attendu.

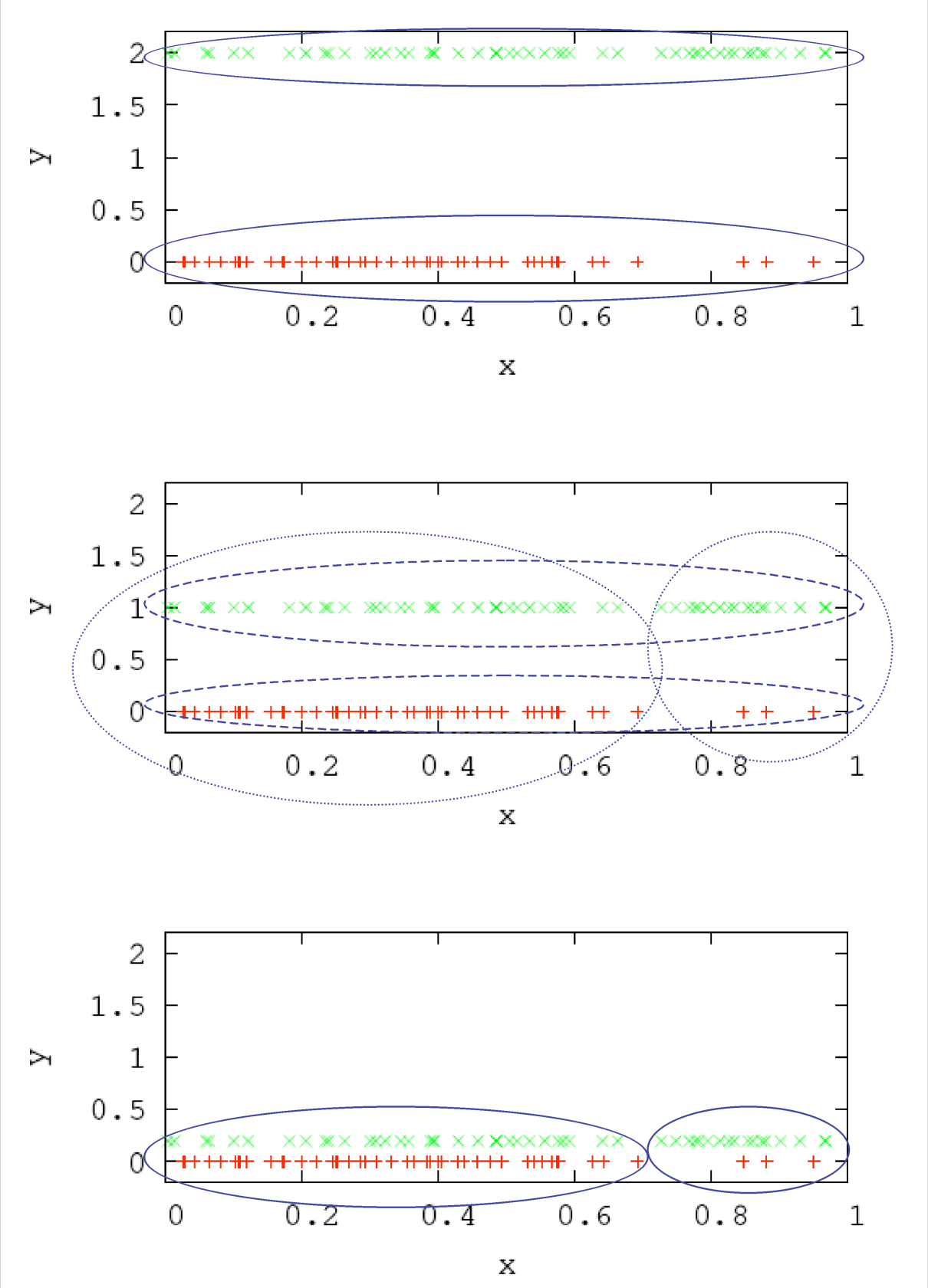


FIG. 4.1 – Limites de la classification euclidienne. On a choisi $a = 1$, et de haut en bas on a successivement $e = 2$, $e = 1$ et $e = 0.2$ (voir texte). En bleu, les classifications euclidiennes que l'on peut obtenir. Sur la figure du milieu, deux classifications donnant des scores proches sont désignés par les deux types de pointillés différents.

4.1.2 Choix du nombre de classes

Une fois la distance (ou dissimilarité) choisie, le deuxième paramètre important qui doit être défini avant la classification est le nombre de classes. En effet, il n'existe pas de critère simple pour définir le nombre de groupes distincts dans un jeu de données particulier. C'est un problème particulièrement difficile, et assez complexe à appréhender quand on voit la facilité avec laquelle l'esprit humain est capable de le résoudre : il ne faut pas plus de quelques secondes à quelqu'un pour déterminer le nombre de personnes présentes sur une photo, même si celle-si sont à moitié cachées. Or ce problème est numériquement très difficile à résoudre.

On verra dans la section suivante que les algorithmes de classification les plus utilisés se répartissent en deux classes : les algorithmes hiérarchiques, qui testent tous les nombres de groupes possibles, et les algorithmes de réallocation, qui donnent la configuration la plus exacte possible à un nombre de groupes donnés. Un des problèmes principaux vient du fait que les critères numériques simples, comme la somme des variances intra-classes, diminue avec le nombre de classes, et donc que ce critère ne permet de fixer un maximum ou un minimum global en fonction du nombre de classes employées. On retrouvera ce problème lors de la définition du critère de stabilité de notre méthode de classification, au chapitre 6.

De plus, le choix du nombre de groupes n'est pas soumis qu'à des contraintes objectives. On veut souvent garder le nombre de classes le plus bas possible, afin de pouvoir faire une synthèse succincte de chaque classe. Mais on veut également que les objets à l'intérieur de chaque classe soient les plus similaires possibles. Ceci montre déjà le compromis à faire entre la compression maximale, dont un exemple est la classification qui regroupe tous les points dans une seule classe, et donc qui minimise le nombre de groupes, et la précision maximale, la classification la plus détaillée possible, qui à chaque point associe son propre groupe. Ces deux extrêmes sans intérêt vont souvent optimiser les critères de partition les plus simples auxquels on pourrait penser... Une formalisation du problème de l'équilibre entre compression et précision, en suivant une approche à la fois de physique statistique et de théorie de l'information, est détaillée dans le chapitre suivant.

De nombreuses solutions ont été trouvées au problème du choix du nombre de groupes, par des approches statistiques classiques. Tout d'abord, certaines méthodes permettent de "deviner" de façon assez précise le nombre de classes. Ces méthodes font souvent appel à une visualisation des données, et à l'efficacité de l'observateur pour trouver le nombre de classes. Un exemple est l'analyse en composantes principales (voir Saporta (1990)), qui permet de visualiser en un faible nombre de dimensions un jeu de données à l'origine p-dimensionnel. Pour l'aide à la classification, l'application est de visualiser sur le plan en deux dimensions les données, et d'en inférer le nombre de classes. Cette méthode a été utilisée dans l'analyse des propriétés des acides aminés (voir Pascal et al. (2005) par exemple). L'idée est de projeter les données p-dimensionnelles sur le plan formé par les deux axes sur lesquelles la variance des données est la plus grande. Le calcul de ces deux axes revient en pratique au calcul des vecteurs propres ayant les plus grandes valeurs propres de la matrice de covariance des données. Les deux vecteurs trouvés définissent un plan, sur lequel on peut projeter les données. De plus, les valeurs propres de la matrice de covariance donnent le pourcentage de la variance totale des données expliqué par chaque axe. Il est fréquent qu'une grande partie de la variance soit expliquée par un ou deux axes. Ceci permet d'inférer avec une confiance relative le nombre de classes dans lesquelles

sont réparties les données. Néanmoins, cette méthode de réduction dimensionnelle, par principe, ne permet d'utiliser qu'une partie de l'information disponible dans le jeu de données pour le choix du nombre de groupes, et dans ce cas précis s'en remet à l'œil de l'observateur, forcément subjectif, pour définir le nombre de classes.

Une autre gamme de méthodes consiste à obtenir N classifications, de la plus triviale à la plus détaillée, et à appliquer un critère numérique pour définir le nombre de groupes optimal. Ce sujet a fait l'objet d'une nombreuse littérature, de laquelle je ne prendrai que quelques exemples pour illustrer la diversité des méthodes employées :

- Ma et al. (2006) classent des données d'expression de gènes en utilisant un algorithme bayesien, basé sur l'optimisation de la vraisemblance des données. Dans ce cas, il est possible de pénaliser la présence de groupes trop nombreux, et d'obtenir un critère qui maximise une fonction de la vraisemblance et de la pénalité, donnant un nombre de classes optimal. Ce critère, le "Bayesian Information Criterion" (BIC) (Schwarz, 1978), est un représentant d'une plus grande famille de critères bayesiens comprenant par exemple le MDL (Minimum Description Length). Ces critères apportent la notion conceptuellement importante d'équilibre entre précision et simplicité, qui sera détaillé dans le chapitre 5.
- Stone (1974) et Smyth (2000) calculent le nombre de groupes sur une partie seulement de leurs données, en utilisant un critère dépendant de la méthode d'analyse, et ensuite vérifient la cohérence de ce nombre de groupes en appliquant la même procédure sur l'autre partie des données. La similarité des paramètres trouvés dans les deux sous-groupes permet d'évaluer la justesse de la méthode.
- Tibshirani et al. (2001) développent une statistique, la "Gap statistic", qui permet de comparer toutes les classifications à k classes à la classification triviale à 1 classe. Un test statistique est ensuite développé, donnant une significativité à chaque classification relativement à son nombre de groupes et à l'étalement des données, et donc un indice de confiance pour le nombre de groupes.
- Roth et al. (2004) utilisent un critère de stabilité basé sur l'intuition qu'une bonne classification donne les mêmes résultats sur les vraies données et sur des données générées à partir de la même source, et comparent les résultats de la même méthode de classification sur ces deux jeux de données pour estimer le nombre de groupes optimal.
- Vogl et al. (2005) emploient une méthode totalement bayesienne, et considèrent le nombre de classes comme un paramètre dans son modèle bayesien pour obtenir sa distribution de probabilité.

On voit que ces critères sont très variés. Cependant, aucun d'entre eux ne s'est imposé comme meilleur que tous les autres, et on verra que des développements théoriques récents montrent que leur choix est un problème toujours d'actualité.

4.2 Les méthodes usuelles de classification

Nous allons présenter ici les critères et les algorithmes de classification les plus courants. Il faut tout d'abord remarquer que l'algorithme est indépendant à la fois de la distance choisie et du critère de partition à optimiser. Le choix des bonnes combinaisons algorithme-distance-critère est une question qui dépend très fortement du problème étudié et des contraintes qu'il pose. Une formalisation de cette question, ainsi que de nombreuses

références, peuvent être trouvées dans (Gordon, 1999), qui décrit quelles classifications pouvaient être effectuées en fonction du type de données pour ne pas induire de biais.

4.2.1 Critère de partition

La plupart des méthodes de classification commencent par l'établissement d'une matrice de dissimilarité, ou distance, entre les données. Le but final étant de regrouper des objets similaires, on a vu que le choix de la distance employée pour mesurer la différence entre deux objets est crucial. Ensuite, quels que soient les algorithmes employés, un critère de partition est optimisé. En général, ce critère est une mesure de l'exactitude de de la classification obtenue, et il est maximisé à chaque itération jusqu'à convergence.

Ce critère peut être de nature variable. On peut essentiellement comparer la dissimilarité au niveau des groupes, et maximiser l'isolation des groupes entre eux, ou maximiser l'homogénéité interne de chaque groupe, ou encore une combinaison des deux. Ces deux critères, quoique complémentaires, donnent lieu en pratique à des classifications différentes sur les mêmes jeux de données. À l'instar de la distance employée, le choix du critère de partition dépend des données étudiées, et une étude préliminaire de leur structure permet souvent d'inférer quel critère donnera les meilleurs résultats. Il faut remarquer ici que tous les critères de partition ne sont pas basés sur une matrice de dissimilarité et sur la minimisation ou la maximisation de distances. Des critères combinatoires, basés sur la répartition des éléments dans les groupes, peuvent être employés.

Les critères mesurant l'isolation d'un groupe relativement aux autres peuvent être directement basés sur la dissimilarité choisie, ce qui implique de généraliser les propriétés de la distance par paires à des groupes, ou basée sur un d'autres paramètres comme les probabilités d'appartenance à chaque groupe de chaque élément. En effet, si la dissimilarité d permet d'évaluer si deux objets doivent être regroupés ou non, elle ne permet pas de comparer directement si deux groupes d'objets dans une partition correspondent bien à deux entités distinctes. Cependant, une manière simple d'évaluer la dissimilarité inter-classes D est de définir un représentant $\vec{r}_i$, $i = 1..c$ pour chacune des c classes, et de définir la dissimilarité inter-classes à partir de celles des paires, comme :

$$D(\vec{x}_i, \vec{x}_j) = d(\vec{r}_i, \vec{r}_j), \tag{4.7}$$

ce qui reporte le problème à celui de la définition d'un représentant (voir par exemple (Mézard, 2007) pour une discussion récente). La notion de représentant est importante, puisqu'un représentant permet également de symboliser tous les éléments de sa classe, sans avoir à les décrire de façon exhaustive. Cet aspect sera abordé plus en détails par la suite. Une autre façon d'utiliser d pour calculer l'isolation d'un groupe α est de mesurer la distance minimale avec le membre le plus proche appartenant à une autre classe :

$$D = \min d(\vec{x}_i, \vec{x}_j),\ i \in \alpha, j \notin \alpha. \tag{4.8}$$

On peut ainsi créer de nombreux critères mesurant l'isolation d'un groupe au reste des données. Une dernière méthode est par exemple de mesurer la somme des distances d'un objet dans le groupe α à tous les objets extérieurs au groupe :

$$D = \sum_i \sum_j d(\vec{x}_i, \vec{x}_j),\ i \in \alpha, j \notin \alpha. \tag{4.9}$$

Les critères non basés sur l'isolation cherchent à maximiser l'homogénéité de chaque groupe (on veut que les groupes soient les plus concentrés possible), en utilisant par exemple un critère de minimisation de la somme des distances internes au groupe :

$$D = \sum_i \sum_{j<i} d(\vec{x}_i, \vec{x}_j),\ i, j \in \alpha, \tag{4.10}$$

ou un critère de minimisation du diamètre du groupe :

$$D = \max d(\vec{x}_i, \vec{x}_j),\ i, j \in \alpha. \tag{4.11}$$

Les deux méthodes conduisent à des critères de partition différents à optimiser, et ont des performances dépendant beaucoup des données. La grande diversité de mesures possibles pour le critère à optimiser, combiné au nombre de possibilités pour le choix de la fonction de dissimilarité, empêchent de déterminer quelles mesures de classification sont objectivement meilleures que les autres. Néanmoins, un critère de partition, est employé très couramment, malgré les biais qu'il peut entraîner. Il s'agit de minimiser la somme des variances intra-classe :

$$D = \sum_{\alpha=1}^{c} \sum_{i \in \alpha} \left(d(\vec{x}_i, \vec{r}_\alpha)\right)^2, \tag{4.12}$$

où $\vec{r}_\alpha$ est le centre de masse de la classe α, à savoir $\vec{r}_\alpha = \frac{1}{\|\alpha\|} \sum_{i \in \alpha} \vec{x}_i$, avec $\|\alpha\|$ le nombre d'éléments de la classe.

Tous les critères de partition présentés ici utilisaient l'attribution de chaque élément à un seul groupe. Certains de ces critères peuvent se généraliser au cas de la classification floue, en sommant de façon probabiliste la valeur du critère pour chaque objet sur les différents groupes. Par exemple, la minimisation des variances intra-classes, pour un critère de classification floue, donne :

$$D = \sum_i \sum_{\alpha=1}^{c} p(\alpha|\vec{x}_i) \left(d(\vec{x}_i, \vec{r}_\alpha)\right)^2, \tag{4.13}$$

avec $p(\alpha|\vec{x}_i)$ la probabilité conditionnelle pour l'élément i d'appartenir au groupe α. Dans le cas de la classification floue, on ne cherche pas les partitions des données en c classes, mais les valeurs des probabilités conditionnelles d'appartenance aux groupes représentant le mieux les données. Ceci est un contexte parfait pour l'emploi à la fois de méthodes bayesiennes et de maximisation de la vraisemblance. Dans les deux cas, on cherchera à maximiser la vraisemblance des données au vu du modèle, en trouvant les valeurs des probabilités conditionnelles considérées comme des paramètres. Une explication plus détaillée de ce type de méthode est donnée dans la troisième partie de la thèse, pour le projet de classification des gènes en fonction de leur biais de codons.

Nous allons maintenant regarder de plus près les principaux algorithmes de classification employés couramment. Pour simplifier l'étude, nous nous limiterons aux algorithmes de classification dite dure.

4.2.2 Méthodes de réallocation dynamique

Les méthodes de réallocation dynamique sont employées de manière très courante pour la classification. Leur principal avantage est leur simplicité de mise en œuvre. Ces

algorithmes sont itératifs, et leur convergence dépend du niveau de précision requis sur le critère de partition à optimiser. Pour ces méthodes, le nombre de classes c doit être défini à l'avance. La méthode de base est la suivante :

1. On choisit au départ une partition aléatoire des données en c classes. Les centroïdes – autre nom des centres de masse – de ces classes sont calculés avec la distance choisie si nécessaire.
2. À chaque pas de temps, l'algorithme choisit un élément $\vec{x}_i$ dans une classe α, et vérifie si placer $\vec{x}_i$ dans une autre classe améliorerait la valeur courante du critère de partition à optimiser. Si oui, il le déplace dans la classe correspondante.
3. L'algorithme recalcule les centroïdes des c classes avec les éléments dans la nouvelle configuration, si nécessaire.
4. L'algorithme itère les pas 2 et 3 jusqu'à une condition d'arrêt précisée au départ.

De très nombreuses variantes de ces méthodes sont possibles, et ont été testées sans que l'une d'entre elles ne se démarque toujours des autres, selon les critères de jugement employés. Les méthodes varient, en plus du choix de la distance et du critère de partition, dont nous avons déjà parlé :

- Par le choix de la partition de départ. Elle peut être entièrement aléatoire, mais une technique plus souvent utilisée consiste à choisir aléatoirement c éléments au départ, puis à regrouper autour d'eux les autres $N - c$, en commençant par les plus proches. Là encore, les stratégies de choix de l'ordre dans lequel les éléments secondaires sont attribués aux classes correspondants aux c éléments de bases sont très diverses. Finalement la partition de départ peut tout à fait être une partition issue d'une première classification, par exemple pour tester la robustesse d'un résultat.
- Par les transformations effectuées à chaque pas de temps. L'élément $\vec{x}_i$ désigné peut être attribué peut être déplacé dans une classe si cela optimise le critère de partition, ou de façon probabiliste : dans ce cas l'algorithme est un recuit simulé, qui permet d'éviter d'être piégé dans un optimum local.
- Par le nombre de transformations simultanées autorisées à chaque calcul du critère de partition. L'idée est qu'en autorisant plusieurs transformations simultanées, et en calculant le critère de partition sur toutes ces transformations, on peut également éviter le piégeage dans des maxima locaux du critère de partition, en autorisant des changements de configuration plus importants.
- Finalement, quand le calcul des coordonnées des centroïdes est nécessaire, on peut choisir de réactualiser ce calcul après chaque déplacement ou après un certain nombre de transformations, ce qui a également des effets sur les solutions trouvées.

Il existe donc des variantes très nombreuses des méthodes de réallocation dynamique. La plus employée est très certainement la méthode des nuées dynamiques, plus connue sous son nom anglais de "k-means". Dans cette méthode, la distance est la norme euclidienne, le critère de partition est la minimisation de la variance intra-classe, et l'algorithme fonctionne en attribuant un élément au hasard dans la classe de laquelle il est le plus proche – la classe α telle que $d(\vec{x}_i, \vec{r}_\alpha)$ soit minimum – à chaque pas de temps, puis ils recalcule les centroïdes. Il est aisé de montrer que le critère converge toujours vers un minimum, et donc que l'algorithme également. De plus, cette version de l'algorithme converge en général très rapidement, ce qui explique pourquoi elle est employée si fréquemment.

4.2.3 Classification hiérarchique

Les méthodes de classification hiérarchique visent à classer les données non pas en un nombre c de groupes défini à l'avance, mais en une organisation hiérarchisée de groupes, que l'on peut représenter sous la forme d'un arbre ou dendrogramme. Cette classification permet, en plus de distinguer des objets de classes différentes, d'évaluer les différences et les liens entre les classes.

Formellement, une classification hiérarchique est un ensemble de N partitions $\mathcal{P}_i$, $i = 1..N$ des données. Ces partitions comprennent la partition triviale $\mathcal{P}_N$ dans laquelle chaque élément est dans une classe différente, et la partition $\mathcal{P}_1$, où tous les objets appartiennent à la même classe. Les partitions sont reliées entre elles par la relation d'inclusion suivante : si A est une classe de $\mathcal{P}_i$, alors soit $A \in \mathcal{P}_{i+1}$, soit A est l'union de toutes les classes de $\mathcal{P}_{i+1}$ qui n'appartiennent pas à $\mathcal{P}_i$. Chaque groupe d'une classification hiérarchique est donc soit un groupe à un seul élément, soit peut être divisé en plusieurs sous-groupes dans une partition d'ordre supérieur. Une figure aide beaucoup à visualiser ce type d'organisation des partitions (Fig. 4.2).

L'intérêt de ce type de classification réside dans le fait que le nombre de groupes n'est pas fixe, et qu'il est possible de déduire des relations entre les classes (par exemple, de quel autre groupe sont-ils des sous-ensembles). Un problème qui en découle est naturellement que si l'on veut une partition d'un ensemble de données, il faut en choisir une parmi les N proposées par la classification hiérarchique, ce qui ramène au problème du choix du nombre de groupes.

Algorithmes classiques de classification hiérarchique Deux grands types de méthodes permettent d'obtenir une classification hiérarchique. La première est l'utilisation d'algorithmes d'optimisation, qui visent à transformer directement la matrice de dissimilarité entre éléments en un ensemble de mesures ultramétriques. Dans ce cas, la mesure ultramétrique entre $\vec{x}_i$ et $\vec{x}_j$ représente la hauteur sur l'arbre à laquelle se rejoignent les branches partant de $\vec{x}_i$ et $\vec{x}_j$. Ces mesures u sont caractérisées par :

$$u(\vec{x}_i, \vec{x}_j) \leq \max\left(u(\vec{x}_i, \vec{x}_k), u(\vec{x}_k, \vec{x}_j)\right) \ \forall i, j, k \in [1..N]. \tag{4.14}$$

Cette inégalité est simplement la caractérisation mathématique du fait que la structure d'un arbre n'autorise qu'un seul chemin allant de la racine à un groupe donné.

On peut, à partir de la matrice de dissimilarité, déduire un ensemble de mesures ultramétriques entre objets, et ainsi trouver l'arbre reliant les classes. Ces méthodes sont numériquement lourdes, et n'ont que peu d'intérêt si on est moins intéressé par les distances ultramétriques trouvées que par la classification obtenue. Cependant, en phylogénie par exemple, les mesures ultramétriques peuvent représenter le temps depuis lequel deux espèces ont divergé, et ces méthodes sont donc très employées.

La seconde classe de méthode est celle des algorithmes de regroupement et de division. Tous deux fonctionnent itérativement en optimisant à chaque pas de temps le critère de partition. Dans les algorithmes de regroupement, les données sont classées au départ avec un seul point par groupe. On a donc N groupes distincts. Ensuite, itérativement, les groupes les plus semblables sont fusionnés, permettant de passer à chaque étape de k groupes à $k-1$, et ce jusqu'à avoir $k = 1$. Le choix des deux groupes qui vont être fusionnés est ici encore soumis à la définition de la distance entre deux groupes, et donc au critère de partition.Celui-ci peut être local (dépendant uniquement des éléments présents

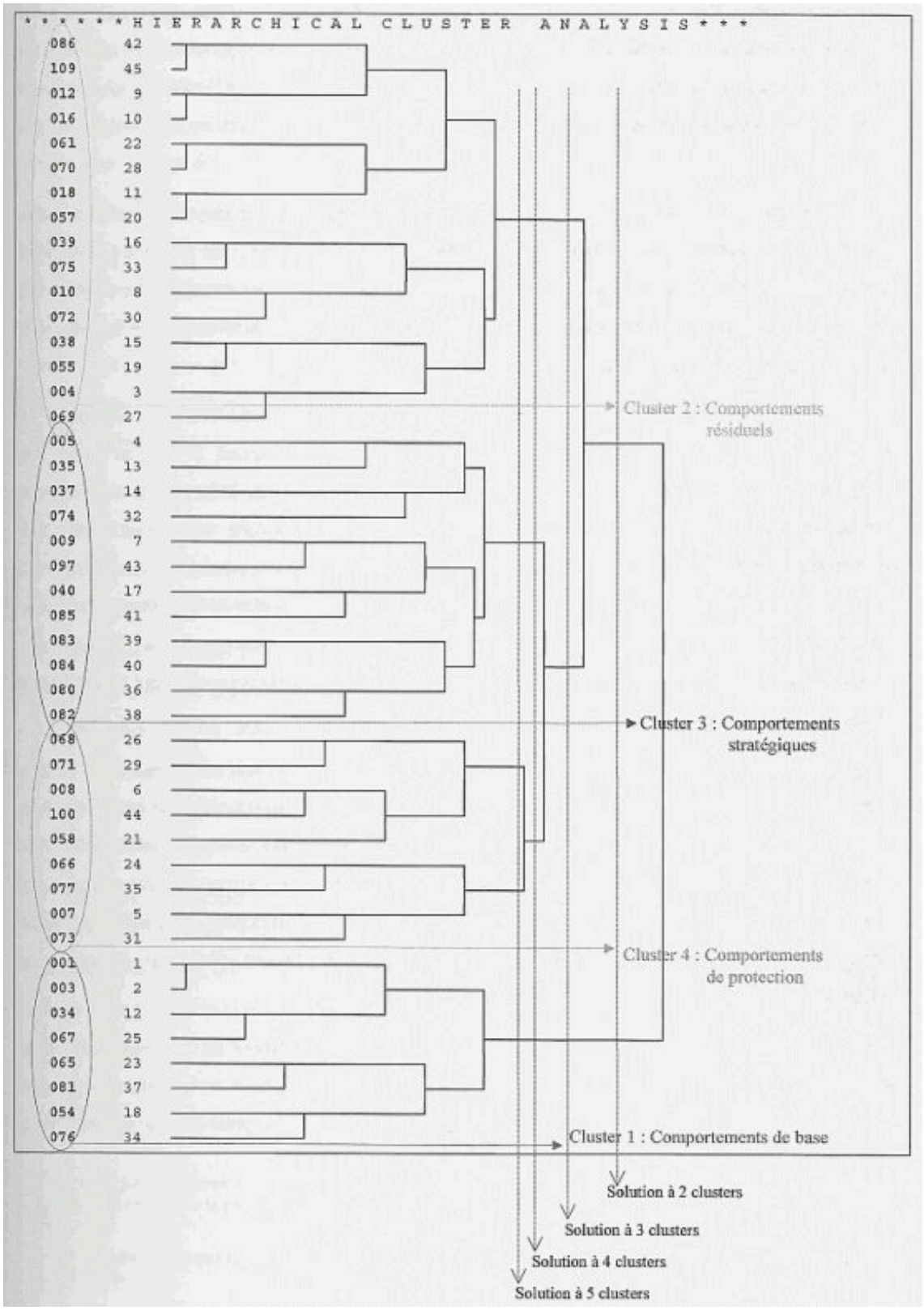


FIG. 4.2 – Exemple de dendrogramme, ou classification hiérarchique, appliquée aux stratégies de résistance au stress chez les adolescents. De droite à gauche on voit que le nombre de groupes augmente (les niveaux à 2, 3, 4 et 5 groupes sont tracés), augmentant la précision de la classification. Les grands groupes, qui représentent les principales stratégies, sont entourés.

entre les deux groupes) ou global (un indice calculé sur toute la partition, et dont la valeur doit varier de façon maximale à chaque itération).

Les algorithmes de division font exactement l'inverse : au départ, les N points appartiennent à un seul groupe. Ensuite, itérativement, un groupe va être divisé en deux sous-groupes, de telle sorte que la partition optimise un critère donné (local ou global). Ceci peut continuer jusqu'à avoir divisé les données en N singulets. L'avantage des méthodes de division réside dans le fait qu'en peu d'étapes, on peut obtenir une partition en un faible nombre de groupes, qui est souvent celle recherchée car offrant la plus grande compression des données. Leur défaut est que le nombre des partitions possibles d'un groupe de N objets en deux sous-ensembles étant extrêmement élevé, le critère de partition doit être bien choisi, sans quoi le problème de trouver la partition optimale en deux d'un groupe nécessite un temps exponentiel en fonction du nombre N d'objets à classer : il existe en effet 2^N bi-partitions d'un groupe de N éléments.

Un défaut général des algorithmes de classification hiérarchique est que, s'ils utilisent un critère de partition local sur le contenu des groupes, son optimisation ne se traduit pas forcément en une optimisation globale. Et l'usage de critères de partition globaux peut, quant à elle, conduire à des calculs très longs.

4.2.4 Méthodes d'apprentissage

Une toute autre classe de méthodes est celle des méthodes dites d'apprentissage. Ces méthodes diffèrent des précédentes dans le sens où elles sont heuristiques : elles ne donnent pas lieu à l'optimisation d'un critère de partition. La définition même de distance n'y est pas nécessaire. Ces méthodes, semblables à celles utilisées en intelligence artificielle, fonctionnent de la manière suivante :

- On commence avec un algorithme de classification plus ou moins aléatoire, ou biaisé en fonction de ce que l'on connaît déjà de la situation. Cet algorithme analyse les données de façon connue, mais ne correspondant pas forcément au problème, et la classification qu'il effectue est très mauvaise.
- Durant la phase d'apprentissage, on donne à l'algorithme des objets dont on connaît la classe réelle à classer. La connaissance de la classe réelle à laquelle ces objets appartiennent permet à l'algorithme se modifier (par exemple, en faisant varier les poids de certains de ses paramètres) de façon à classer dans le bon groupe les données. On itère alors jusqu'à ce que le modèle soit le plus précis possible, donc avec le plus grand jeu de données d'entraînement possible.
- Ensuite, on utilise l'algorithme classer un autre jeu de données, dont on ne connaît pas la vraie classification. L'idée est que l'algorithme, à forces de modifications, est passé d'un algorithme de classification aléatoire à un algorithme qui classe correctement les objets, sans que l'on ait eu à définir un critère particulier. L'algorithme classe les nouvelles données simplement comme il avait classé les anciennes ; il fonctionne en rassemblant naturellement les objets qui sont les plus semblables au vu de ses paramètres, qui ont été "entraînés" à trouver la bonne classe.

L'intérêt de ce type de procédures est double. Il permet tout d'abord de profiter de toutes les données déjà bien étudiées, pour lesquelles on connaît les classifications correspondantes avec un haut indice de confiance. De plus, il permet de trouver des classifications basées sur des critères non-linéaires complexes. Finalement, ce type d'algorithme peut se renforcer avec la présence de nouveaux jeux de données. Un exemple simple est d'imaginer

d'avoir à classer deux groupes bien distincts, mais séparés par une frontière extrêmement sinueuse, voir discontinue. Un algorithme classique comme les nuées dynamiques n'obtiendra pas dans ces cas d'aussi bon résultats qu'un algorithme d'apprentissage. Cependant, le défaut de ces méthodes est qu'elles nécessitent l'existence d'un jeu de données d'entraînement pour lesquelles on connaît avec certitude la classification désirée. Ceci est malheureusement rarement le cas en biologie, où les seules classification pré-existantes ne sont pas nécessairement sûres. Parmi ces algorithmes, on peut citer parmi les plus connus les réseaux de neurones ou les algorithmes du type "Support Vector Machines" (voir par exemple Cristianini and Shawe-Taylor (2000)).

Chapitre 5

L'apport de la théorie de l'information

5.1 Bases de théorie de l'information

La théorie de l'information a été développée par C. Shannon en 1948. Elle visait à modéliser les télécommunications, en se basant sur les techniques de codage. On peut y trouver de très nombreuses analogies avec la physique statistique et la notion d'entropie, et c'est sous cet aspect que les principaux résultats vont être présentés. On verra tout d'abord les notions de base de cette théorie, puis dans une deuxième partie on développera les liens entre classification, codage et compression, et comment par cette voie la théorie de l'information apporte un nouveau regard sur le problème de la classification.

5.1.1 Définitions

Sauf mention contraire, les définitions et théorèmes évoqués ici proviennent tous de Cover and Thomas (1991). Nous présentons les définitions de manière quelque peu formelle, sans les rattacher à un problème particulier pour le moment. Soit une distribution de densité de probabilité $p(x) = p(X = x)$ donnée, associée à une variable aléatoire X de support S. Par définition de la densité de probabilité on a la condition de normalisation :

$$\int_S p(x)dx = 1. \tag{5.1}$$

On peut quantifier l'entropie de la distribution p. De façon très similaire à l'entropie en physique statistique, que l'on associe généralement à une mesure du nombre de micro-états, l'entropie en théorie de l'information est une mesure de la non prédictibilité de la variable aléatoire X. Pour bien comprendre cela regardons deux cas limites. Si $p(x) = \delta(x - x_0)$, alors il n'y a aucune incertitude : toute réalisation de la variable aléatoire x aura pour résultat x_0. D'un autre côté, si on choisit la distribution uniforme $p(x) = 1/S$, avec S l'aire du support de la distribution (telle que $S = \int_S dx$), l'incertitude dans la prévision du résultat d'un tirage de la variable aléatoire x est maximale : tous les résultats sont équiprobables. L'entropie de la distribution est une mesure de cette incertitude, et on définit :

$$H(X) = -\int_S p(x) \ln(p(x))dx. \tag{5.2}$$

Un calcul élémentaire confirme que dans le cas d'une distribution de Dirac, on a $H(X) = 0$ et dans le cas d'une distribution uniforme, $H(X) = \ln(S)$. L'entropie est traditionnellement mesurée en bits, en utilisant des logarithmes de base 2 dans la formule précédente. Ici, on gardera la notation en logarithme népérien, sans se poser le problème de l'unité, car ce sont comparaisons d'entropie qui vont nous intéresser. On remarquera que l'entropie d'une distribution n'est rien d'autre, suivant cette définition, que l'espérance mathématique de $\ln(p(x))$.

Soient de deux variables aléatoires X et Y de distribution jointe $z(x, y) = q(y|x)p(x)$ et de supports respectifs S et T, où $q(y|x)$ est la probabilité conditionnelle d'observer un événement y sachant qu'un événement x a été observé. L'entropie des deux variables jointes $H(X, Y)$ s'écrit naturellement :

$$H(X, Y) = -\iint_{S,T} z(x, y) \ln(z(x, y))dxdy. \tag{5.3}$$

Cette notion peut être généralisée de la même façon à N variables.

Une dernière définition est celle de l'entropie conditionnelle de la variable aléatoire Y, sachant le résultat du tirage x de la variable aléatoire X. On voit que si X et Y sont très couplés, cette entropie conditionnelle doit être faible : l'information sur la valeur de x nous renseigne beaucoup sur les valeurs possibles de y. Au contraire, pour deux variables indépendantes, l'entropie de Y sachant X est la même que l'entropie de Y : l'information sur x ne nous apporte rien sur y. On voit tout de suite que l'inégalité $H(Y|X) \leq H(Y)$ doit toujours être respectée : le conditionnement apporte toujours de l'information, et la connaissance d'une variable ne peut pas diminuer la prédictibilité d'une autre. On définit l'entropie conditionnelle en termes des distributions de probabilités comme suit :

$$H(Y|X) = \iint_{S,T} z(y, x) \ln(q(y|x)). \tag{5.4}$$

5.1.2 Distance et information mutuelle

Une notion qui va nous intéresser à cause de son lien avec la notion de dissimilarité, est celle de distance entre deux distribution de probabilités p et q, associées à la même variable aléatoire X. Cette distance est une façon de mesurer l'écart entre les deux distributions, ou la différence attendue entre un tirage d'une série d'événements avec la distribution p. et un tirage fait avec la distribution q. Cette notion permet par exemple de quantifier l'erreur faite sur la valeur d'un tirage de X, dont la vraie distribution est p, si on l'estime avec q, et c'est pour cela qu'elle est très employée. On peut la calculer comme suit :

$$D(p\|q) = \int_S p(x) \ln\left(\frac{p(x)}{q(x)}\right) dx. \tag{5.5}$$

Cette distance, dite de Kullback-Leibler, n'est pas une métrique, car elle ne respecte pas l'inégalité triangulaire et n'est pas symétrique : on n'a pas forcément $D(p\|q) = D(q\|p)$, bien que cela puisse arriver en fonction des distributions choisies. Ce deuxième point peut être corrigé par symétrisation :

$$D^\star(p\|q) = \frac{1}{2}\left[D(p\|q) + D(q\|p)\right] = \frac{1}{2}\left[\int_S p(x) log\left(\frac{p(x)}{q(x)}\right) dx + \int_S q(x) log\left(\frac{q(x)}{p(x)}\right) dx\right]. \tag{5.6}$$

La distance $D^\star$ est la distance de Kullback-Leibler symétrisée. On voit que si $p(x) = q(x)$ presque partout (au sens mathématique du terme), on a bien $D^\star(p\|q) = 0$. Une généralisation de la distance de Kullback-Leibler symétrisée est (Lin, 1991) :

$$D_\lambda(p\|q) = \lambda D(p\|\lambda p + (1-\lambda)q) + (1-\lambda)D(q\|\lambda p + (1-\lambda)q). \tag{5.7}$$

Cette généralisation permet de voir que la distance de Kullback-Leibler n'est qu'un représentant d'une famille beaucoup plus large. L'un de ces représentant, la distance de Jensen-Shannon, correspondant à la valeur $\lambda = \frac{1}{2}$, est plus connu car il possède l'intéressante propriété d'être le carré d'une métrique, la métrique de Hellinger. Cette propriété, et d'une façon plus générale la famille des distances D_λ, quoique dépassant largement le cadre de notre étude, sont grandement utilisées dans les problèmes de théorie de la décision, à cause de ce que permettent de mesurer ces distances comme évoqué plus haut.

Ces différents points nous amènent à une définition importante, celle d'information mutuelle. Soient deux variables aléatoires X et Y de support S et T et de distributions respectives p et q. Notons comme avant $z(x, y)$ la distribution jointe de ces deux variables, avec $z(x, y) = q(y|x)p(x)$. L'information mutuelle représente le niveau de corrélation, ou de couplage, entre p et q. Cette corrélation est mesurée comme la différence entre la distribution z et la distribution produit $p \times q$. On a :

$$I(X, Y) = D\left(p(x, y)\|p(x)q(y)\right) = \iint_{S,T} z(x, y) \ln\left(\frac{z(x, y)}{p(x)q(y)}\right). \tag{5.8}$$

Dans le cas de distributions $p(x)$ et $q(y)$ indépendantes, cette information mutuelle vaut bien 0. L'information mutuelle représente en quelque sorte ce qui est déjà connu de la variable aléatoire Y, si l'on connaît X.

On peut également définir l'information mutuelle comme la quantité d'entropie présente dans la distribution q qui est déjà dans p. La notion d'information mutuelle est intimement reliée à l'entropie conditionnelle de Y, sachant X. On peut montrer que :

$$H(Y|X) = H(Y) - I(X, Y). \tag{5.9}$$

Cette relation montre comment l'incertitude sur la variable aléatoire Y diminue grâce à la connaissance du résultat de la variable aléatoire X, et pourquoi le conditionnement réduit l'incertitude. Elle permet également de déduire que $I(X, Y) \geq 0$. Essentiellement, l'entropie conditionnelle est celle de la distribution conditionnelle $q(y|x)$, comme on peut le voir en écrivant :

$$H(Y|X) = \iint_{S,T} z(x, y) \ln(q(y|x) = \int_S p(x) \int_T q(y|x) \ln(q(y|x)). \tag{5.10}$$

En écrivant la même relation pour X conditionnée par Y et en sommant avec l'équation 5.9, on obtient :

$$H(X, Y) = H(X) + H(Y) - I(X, Y). \tag{5.11}$$

Cette dernière relation rappelle l'égalité de théorie des probabilités $p(A \cup B) = p(A) + p(B) - p(A \cap B)$, pour deux ensemble d'événements A et B quelconques. Visuellement, il s'agit bien de la même chose : l'entropie de la distribution jointe de X et Y est la somme des deux entropies indépendantes, moins le facteur d'information mutuelle représentant ce que la connaissance d'une variable apporte sur l'autre.

5.1.3 Information transmise par un canal

La théorie de Shannon, originellement, visait à modéliser les télécommunications. Dans ce cadre il est naturel de visualiser le problème comme celui de la transmission, la plus fidèle possible, du résultat d'événements de la variable aléatoire X, sur un canal. Les notions de codage qui ont été introduites à l'époque vont nous être utiles pour comprendre l'approche informationnelle de la classification, nous allons donc les développer quelque peu ici. Le problème posé est le suivant : quelle est la longueur minimale L d'un message décrivant complètement un événement x de X ? Cela dépend évidemment de la "complexité" de l'événement x, ou plus précisément de la complexité d'un événement typique issu de X[1]. Si l'on veut parfaitement transmettre le signal, Shannon a montré que la réponse était :

$$H(X) \leq L < H(X) + 1. \tag{5.12}$$

Ici L est la longueur du message décrivant un événement x de X. Si le logarithme utilisé est de base 2, l'unité est le bit et le message est le codage de x en binaire. Dans le cas du logarithme népérien, on peut imaginer le même codage, mais la longueur doit être multipliée par une constante, $\ln(2)$, que l'on néglige car elle sera présente partout.

Le bit supplémentaire du terme de droite vient de la possibilité pour $H(X)$ d'être non-entier. On peut améliorer ce résultat en bornant le nombre L_n de bits nécessaire en moyenne par symbole pour décrire une séquence de n symboles x_i, $i = 1..n$ produits par la même source X, supposée stationnaire et sans mémoire. Dans ce cas le bit supplémentaire se partage sur les n événements, et on a :

$$H(X) \leq L_n < H(X) + \frac{1}{n}. \tag{5.13}$$

Pour $n \to \infty$, on voit qu'on peut dire que "l'entropie d'une source est la complexité des événements qu'elle produit", où la complexité est mesurée par la longueur nécessaire pour décrire fidèlement un événement. Il est important de remarquer que dans ce cas, la longueur du message envoyé est exactement la quantité d'information envoyée.

Ces résultats nous permettent donc de calculer quelle quantité d'information est nécessaire pour représenter fidèlement un événement. Mais un problème qui s'est très tôt posé en télécommunications, est celui de la transmission de l'information sur un canal bruité. Supposons qu'un événement x d'une variable aléatoire X soit transmis sur un canal bruité. À cause du bruit, le receveur ne pourra pas connaître x avec précision. Pour chaque message x envoyé, le receveur verra arriver y, qui dépend de x par un ensemble de probabilités conditionnelles $p(y|x)$ générées par le bruit. Dans ce cas, Shannon a montré que la quantité d'information reçue n'est plus en moyenne $L = H(X)$ par message envoyé, mais une quantité inférieure $L' = H(X) - H(X|Y)$, soit $L' = I(X, Y)$. Si on n'a pas de bruit, $y = x \ \forall x$ et $I(X, Y) = H(X)$: on se retrouve dans le cas précédent. Si le bruit est tellement fort que l'on peut considérer X et Y comme des variables indépendantes, on trouve $H(X|Y) = H(X)$: le fait de connaître la valeur y en sortie du canal ne donne aucune information sur la valeur x en entrée. Dans ce cas l'information transmise L' est

[1]La différence fondamentale entre la complexité moyenne de X et la complexité de chacun des événements x est prise en compte dans la théorie de la complexité algorithmique de Kolmogorov. Des développements récents pour appliquer cette théorie aux problèmes de classification et de compression ont eut lieu, mais ne seront pas abordés ici. Le lecteur intéressé pourra se référer à Li et al. (2003), Cilibrasi and Vitanyi (2005) ou encore Grunwald and Vitanyi (2004).

nulle, quelle que soit la longueur L du message envoyé. On voit donc que naturellement, le bruit fait perdre de l'information sur le signal envoyé, et que c'est l'information mutuelle $I(X, Y)$ entre le signal envoyé et le signal reçu qui mesure la quantité d'information transmise. Cette notion va être réemployée dans la section suivante sur la compression.

5.2 Codage, compression et fonction taux-distorsion

5.2.1 Principes de codage

Le bruit sur un canal réduit la quantité d'information que l'on peut transmettre par symbole, dans le cadre de communications où l'on cherche à envoyer un message représentant un événement x le plus fidèlement possible. Ce résultat va nous servir dans le cadre d'un approche différente, celle du codage et de la compression, dont on va voir le lien avec la classification dans la suite de ce chapitre.

Le codage consiste simplement à trouver un code permettant de décrire un message x avec des variables par exemple binaires. Les études de Shannon, de Fano et par la suite de Huffmann ont montré qu'il existait des codes permettant de décrire des événements x avec un message d'une longueur égale ou très légèrement supérieure à $H(X)$, calculée en bits. De tels codes ont été construits. Pour minimiser la longueur du message à envoyer en moyenne pour décrire un événement x, ils assignent à chaque événement x un mot-code de longueur inversement proportionnelle à $p(x)$. Par exemple, l'événement x ayant la plus grande probabilité d'arriver se voit attribuer le mot-code le plus court. Ainsi, dans la majorité des cas, c'est ce mot-code qui sera transmis, minimisant la longueur de la description.

5.2.2 Compression

Dans le cas de la compression l'objectif principal n'est pas de décrire exactement x, mais de le décrire de la manière la plus courte possible. Ici on ne détaillera pas la compression sans perte, utilisée en informatique de manière courante – qui n'est en fait qu'une forme de codage élaborée. Nous allons détailler le problème de la compression avec pertes, dite aussi non conservative. Ce problème, comme cela sera détaillé dans la section suivante, est formellement analogue à un problème de classification et de représentation des données.

Quand on compresse des données, on perd une partie de l'information sur elles pour pouvoir les décrire de façon plus succincte. La quantité d'information que l'on désire garder sur x est donc plus petite que $H(X)$. On veut, grâce à la compression, minimiser la longueur moyenne du message à transmettre pour décrire un événement x, de manière à ajuster la longueur du message à transmettre à la quantité d'information que l'on veut garder sur x. On se sert de la perte d'information due à la description trop courte de x, pour déterminer un code qui ne va décrire que la "partie de l'information" désirée, et donc qui nécessitera moins de $H(X)$ bits de longueur de description par événement. Plus la perte d'information acceptée sera importante, plus on pourra réduire la longueur de la description de x, puisque on pourra éliminer du message tout ce qui n'est pas intéressant dans notre nouvelle représentation.

Un exemple de compression de ce type serait de coder une séquence d'ADN en représentant chaque série de 10 bases par le nucléotide majoritaire dans ces 10 bases ; la description

finale de la séquence serait 10 fois plus courte que l'originale, tout en gardant (à un certain niveau) une partie de l'information originelle contenue dans la séquence.

Supposons que la variable aléatoire X produise des messages x avec les probabilités $p(x)$. On définit un codage probabiliste de x par des mots-code y, en associant aux mots-code y de $\mathcal{Y}$ des probabilités conditionnelles $p(y|x)$. Contrairement au codage, on ne veut pas avoir un mot-code y pour chaque événement x, mais on veut associer plusieurs événements x au même mot-code y avec les probabilités $p(y|x)$. Dans le cas de la compression, on peut donc utiliser un ensemble $\mathcal{Y}$ formé de moins d'éléments que le nombre des messages x possibles, puisque plusieurs x seront représentés par le même mot-code y.

On définit la distorsion entre un élément x de X et son représentant y comme le coût, ou la perte de significativité, à utiliser y pour représenter x. Cette distorsion est très similaire, comme on le verra, à la dissimilarité évoquée au chapitre précédent. Si cette fonction $d(x, y)$ est donnée pour chaque couple (x, y), on peut définir la distorsion observée lors de la transmission d'une séquence de n symboles x^n, si la séquence y^n est reçue :

$$d(x^n, y^n) = \sum_{i=1}^{n} d(x_i, y_i). \tag{5.14}$$

La distorsion D moyenne attendue lors de l'envoi d'une séquence x^n aléatoire est la moyenne sur les séquences x^n :

$$D = \sum_{x^n} \sum_{y^n} p(y^n|x^n) p(x^n) d(x^n, y^n). \tag{5.15}$$

D est la distorsion moyenne attendue. Elle ne dépend que des probabilités conditionnelles de codage $p(y|x)$. Le problème de la compression optimale est de trouver les distributions $p(y|x)$ et les mots-code y associés, de façon à minimiser la longueur moyenne des mots de code y employés, tout en minimisant en parallèle la distorsion D. Pour minimiser la longueur moyenne des messages, il faut choisir les mots-code les plus courts et donc les moins nombreux possibles, car plus le nombre de mots-code utilisé est grand, plus le nombre de mots-code dépassant une longueur donnée augmente. Cependant, des mots-code plus courts transmettront moins d'information sur l'événement initial x à décrire, et augmenteront la distorsion D.

La distorsion D représente la quantité d'information perdue lors du codage de x par y. Pour une distorsion fixée, on veut représenter les données dans le codage le plus économique, celui pour lequel la longueur du message L transmis par message x émis est minimale. Par analogie avec ce qui a été montré pour le canal bruité, on va voir que la longueur L, à D fixé, peut au minimum être :

$$L(D) = \min I(X, Y), \tag{5.16}$$

où le minimum est cherché à D fixé sur l'ensemble des distributions $p(y|x)$ telles que :

$$\sum_{(x,y)} p(x) p(y|x) d(x, y) \leq D. \tag{5.17}$$

Dans le cas du canal bruité, $I(X, Y)$ représentait la quantité d'information réellement transmise, et Y était la variable reçue. Ici, on voit que si on effectue une compression qui fait perdre de l'information sur X, on peut l'utiliser pour transmettre un message ne

décrivant que l'information qu'on a gardée sur X, donc plus court. Dans le cas du canal bruité, le bruit réduisait le rapport de la quantité d'information transmise à la longueur du message envoyé. Le résultat pour la compression nous montre que si on contrôle la perte d'information par une compression de X, on peut également réduire du même facteur la longueur du message nécessaire pour décrire un événement x. La compression est une perte d'information contrôlée qui permet de réduire la complexité de la description d'un objet x.

Le fait que ce soit le minimum de $I(X,Y)$ que l'on cherche peut sembler paradoxal, mais on peut le comprendre en étudiant deux cas limites. Si on utilise un codage aléatoire, avec x et y indépendants pour tout x, alors on a $I(X,Y) = 0$. Dans ce cas, le nombre de bits nécessaire pour représenter un événement x est nul : aucune information sur x ne peut être transmise par y, donc l'information transmise par un message est nulle, quel que soit sa longueur.

Au contraire, si y détermine x de façon unique, alors $I(X,Y) = H(X)$ et on retrouve le théorème précédent : on pourra transmettre toute l'information sur la variable X. Cela ne peut arriver que si la distorsion est nulle, ou suffisamment faible pour qu'à chaque élément y on puisse associer de manière déterministe un élément x. C'est le cas où l'ensemble des descriptions $\mathcal{Y}$ est au moins aussi grand que l'ensemble des messages possibles $\mathcal{X}$. Alors il est possible, si l'on connaît la fonction de codage, de revenir directement à x depuis y et donc de n'avoir perdu aucune information lors du codage. En contrepartie, la compression est nulle, puisque la longueur du message nécessaire pour décrire x est $H(X)$.

5.2.3 Application aux problèmes de classification

On a vu comment la théorie de l'information abordait le problème du codage, et celui de la compression de données : comment représenter des données issues de X par des descriptions Y les plus réduites possibles ?

Dans le cas de la compression de données, on code par une description courte un objet complexe. Cette approche est très similaire à certaines formes de classification : il s'agit, pour un ensemble de départ d'objets $\vec{x}_i$, $i = 1..N$, de trouver des représentants $\vec{y_j}$, $j = 1..c$ de façon à ce que chaque $\vec{y_j}$ décrive un certain nombre d'objets $\vec{x}_i$.

Il y a donc une analogie formelle entre le problème de la longueur de description de x minimale pour une distorsion donnée, et la classification avec un nombre minimal de représentants $\vec{y_j}$ d'un ensemble de vecteurs $\vec{x}_i$. Il est équivalent de chercher l'ensemble des mots-code y les plus courts permettant de coder les événements x issus d'une source X, à chercher un ensemble $\{\vec{y_j}\}$ contenant le moins grand nombre d'éléments et permettant de décrire les $\vec{x}_i$. On a donc analogie entre la recherche des mots-code et celle des représentants permettant de classifier les données. Dans les deux cas on est soumis à une contrainte sur la précision finale que l'on veut garder dans la description des x ou des $\vec{x}_i$. Cette contrainte est modélisée par la distorsion dans le problème de la compression, et par le critère de partition dans le problème de la classification.

La distorsion entre x et y peut être considérée comme l'analogue de la distance, définie au chapitre 4, entre un objet $\vec{x}_i$ et un représentant $\vec{y_j}$. On peut donc concevoir la classification comme le problème de trouver quelle représentation des données est la plus similaire aux données originelles, ce qui donne une condition sur la distance, tout en étant plus succincte, ce qui donne une condition sur le nombre de représentants, et donc le nombre de groupes utilisés.

Il est clair que le choix des représentants est l'étape cruciale dans ce problème. En effet, le choix des classes va conditionner la distorsion obtenue sur les données, mais également la longueur de description nécessaire pour représenter un des objets $\vec{x}_i$: une partition des données en 2 classes permet de décrire chaque événement par une seule variable binaire, mais cause probablement une grande distorsion dans les données.

Le théorème sur la valeur de la fonction taux-distorsion, qui relie le niveau de compression des données et la précision qu'elles gardent, peut être employé directement en classification. Il faut pour cela écrire la distance utilisée comme une distorsion. Il est évident que si les données sont classées en un petit nombre de groupes, la précision conservée sur chaque objet de départ est moins bonne ; la seule classification permettant de conserver toute l'information sur les objets initiaux est la classification triviale en N groupes.

Formellement, on a l'analogie suivante : on veut classer de façon probabiliste en un certain nombre de classes N éléments $\vec{x}_i$. Le nombre de classes n'est pas spécifié à l'avance dans ce cas, puisqu'il va dépendre de la distorsion qu'on s'autorise a avoir sur les données, tout comme le nombre de mots-code dans un codage diminue quand on s'autorise plus de distorsion. Si on représente chaque classe par un représentant $\vec{y}_j$, on peut définir la distance entre chaque élément et un représentant $d(\vec{x}_i, \vec{y}_j)$ comme on le faisait précédemment, en choisissant une fonction de distance.

Dans ce cas, le critère de partition qui va être optimisé est la distance moyenne entre les éléments et leurs représentants :

$$\langle d \rangle = \sum_{i=1}^{N} \sum_{j=1}^{c} q(\vec{y}_j|\vec{x}_i) p(\vec{x}_i) d(\vec{x}_i, \vec{y}_j), \tag{5.18}$$

où $q(\vec{y}_j|\vec{x}_i)$ est la probabilité d'employer le représentant $\vec{y}_j$ pour décrire $\vec{x}_i$, ou autrement dit la probabilité de classer $\vec{x}_i$ dans le groupe j si la classification est "dure".

Comme précédemment, le facteur régulant le nombre de groupes utilisés est ici aussi l'information mutuelle entre les objets originaux $\vec{x}_i$ et leurs représentations $\vec{y}_j$. La différence fondamentale avec le problème de compression est que l'on n'a pas de valeur D fixée à respecter pour la distorsion moyenne. On veut minimiser simultanément $\langle d \rangle$ et $I(X, Y)$, ce qui va nous conduire à un équilibre entre le nombre de groupes que l'on va obtenir (proportionnel a $I(X, Y)$) et la distorsion finale, car la minimisation du nombre de groupes tend à augmenter la distorsion.

Ceci nous montre directement que le formalisme de la théorie de l'information permet de poser le problème du nombre de groupes optimal pour décrire un ensemble. Cette écriture permet également de voir qu'il n'y a pas de nombre de groupes optimal au sens général, puisque quand le nombre de groupes augmente, la précision de la description va toujours augmenter en parallèle. Ceci explique pourquoi un critère extérieur est utilisé dans chaque technique de classification. Un argument récemment employé pour calculer le nombre optimal de groupes, est d'utiliser la limitation intrinsèque des données en précision (Still and Bialek, 2004). Celle-ci permet en effet de trouver une borne supérieure sur le nombre de groupes, au delà de laquelle on commence à séparer des objets sur la base de ce qui n'est peut être qu'une erreur due à l'imprécision de la mesure initiale. Son étude est décrite dans la section suivante.

5.2.4 Solution du problème de classification

Dans cette nouvelle formulation, résoudre le problème de classification est un problème d'optimisation classique qui consiste dans ce cas à trouver les distributions $q(\vec{y}_j|\vec{x}_i)$ qui minimisent à la fois $I(X,Y)$ et $\langle d\rangle$. Ce problème peut être résolu en appliquant la technique des multiplicateurs de Lagrange, donc en calculant l'expression :

$$\min(\langle d\rangle + TI(X,Y)). \tag{5.19}$$

Par analogie avec la mécanique statistique, on voit que la dissimilarité choisie au début du problème joue le rôle d'une énergie, tandis que l'information mutuelle entre Y et X correspond à l'opposé de l'entropie. La solution de ce problème d'optimisation est un ensemble de distributions de Boltzmann paramétrées par T :

$$q(\vec{y}_j|\vec{x}_i) \simeq \frac{q(\vec{y}_j)\exp\left(-\frac{d(\vec{x}_i,\vec{y}_j)}{T}\right)}{Z(\vec{x}_i,T)}, \tag{5.20}$$

avec la fonction de partition :

$$Z(\vec{x}_i,T) = \sum_{j=1}^{c} q(y)\exp\left(-\frac{d(\vec{x}_i,\vec{y}_j)}{T}\right). \tag{5.21}$$

Ces équations permettent d'obtenir une condition sur les représentants $\vec{y}_j$. Par dérivation de l'équation 5.19 par rapport à un changement infinitésimal de la position des représentants $\vec{y}_j$, supposé ne pas avoir d'effet sur les attributions des $\vec{x}_i$, on a :

$$\sum_{i=1}^{N} q(\vec{y}_j|\vec{x}_i)\frac{\partial d(\vec{x}_i,\vec{y}_j)}{\partial \vec{y}_j} = 0. \tag{5.22}$$

Il s'agit là d'une condition intuitive, disant que les représentants $\vec{y}_j$ doivent être au centre des objets qu'ils représentent. On voit que ici aussi la distance choisie joue un rôle primordial, déterminant la géométrie du système et les représentants employés.

Le nombre de groupes est déterminé par le nombre des probabilités $q(\vec{y}_j)$ différentes de 0. Ces probabilités sont calculées parallèlement avec les valeurs de $q(\vec{y}_j|\vec{x}_i)$, grâce à l'algorithme de Blahut-Arimoto (Blahut, 1972). Les $q(\vec{y}_j|\vec{x}_i)$, quant à eux, déterminent les attributions probabilistes des éléments dans les classes. Si T est donné, le problème de classification est donc résolu, au sens où on a une classification floue des données originelles autour de représentants. T est le multiplicateur de Lagrange. Sa valeur est donnée par la valeur de la distorsion $\langle d\rangle$ désirée. Ici comme en physique statistique, on peut assimiler T à une température : quand la température est grande, le minimum de l'équation 5.19 correspond à $I(X,Y) \to 0$. Alors tous les objets $\vec{x}_i$, $i = 1..N$ sont décrits de la même manière par les $\vec{y}_j$, à savoir que aucun représentant n'apporte d'information sur l'objet qu'il représente. La compression est maximale ; cela revient à considérer tous les objets comme identiques, et à les classer dans le même groupe. On peut préciser cette idée en remarquant que la limite $T \to \infty$ dans l'expression de $q(\vec{y}_j|\vec{x}_i)$ conduit à ce que tous les $q(\vec{y}_j|\vec{x}_i)|_{T\to\infty}$ soient égaux. Si c'est le cas, on n'a bien qu'un seul groupe, car rien dans $\vec{x}_i$ ne permet de l'attribuer à un groupe plutôt qu'à un autre. Si au contraire $T \to 0$, chaque point doit être représenté par lui-même pour minimiser la distorsion. En effet on

a $q(\vec{y}_j|\vec{x}_i) \to 0$ sauf si $d(\vec{x}_i, \vec{y}_j) = 0$, auquel cas $q(\vec{y}_j|\vec{x}_i) = 1$. Donc dans cette configuration chaque point est représenté par lui-même, et forme son propre groupe.

Le paramètre T régule l'importance donnée à la compression par rapport à la précision : à haute température, on ne différencie pas les points et la compression est maximale, à basse température, les particularités de chaque point sont importantes.

L'analyse d'un modèle de ce type, inspiré de la physique (Rose et al., 1990), a montré que l'on observe des transitions de phase entre nombre de groupes différents quand on fait varier T dans le modèle précédent. Ces transitions de phase représentent les changements de la précision de la classification obtenue. On obtient grâce à elles une structure hiérarchique de groupes, où l'augmentation de la précision dans la classification conduit à l'augmentation du nombre de groupes par une série de transitions de phase, de façon similaire aux transitions de phase ordre-désordre observées dans un modèle ferromagnétique. Cette analogie avec la physique statistique nous donne la même conclusion que précédemment concernant le problème du choix du nombre de groupes optimal : *Toute augmentation du nombre de groupes conduit à une augmentation de la précision de la description des données, et inversement ; il n'existe donc pas de nombre optimal de groupes dans une classification.*

5.3 "Information bottleneck" et dissimilarité

Des travaux récents (Still and Bialek, 2004; Still et al., 2004; Tishby et al., 1999) ont permis de développer un formalisme basé sur la théorie de l'information, qui prolonge ce qui a été vu à la section précédente, et permet de trouver une alternative théoriquement élégante au problème qui reste posé, le choix de la distance dans les problèmes de classification. En effet, l'interprétation informationnelle de la classification a le même problème que la classification classique : la fonction $d(x, y)$ à employer pour calculer la distorsion du point de vue de l'information doit être donnée, exactement comme l'était la mesure de dissimilarité dans les problèmes de classification.

On pose le problème ainsi (Tishby et al., 1999). On a au départ N objets $\vec{x}_i$, et on veut les compresser avec un code formé de c éléments y_j, $j = 1..c$. Ceci est bien équivalent à un problème de classification floue des N objets en c groupes, où l'on cherche à déterminer les probabilités conditionnelles $p(\vec{y}_j|\vec{x}_i)$. On a vu que, pour une fonction de distorsion $d(\vec{x}_i, \vec{y}_j)$ donnée, et une distorsion valant $\langle d \rangle$, les affectations optimales étaient :

$$q(\vec{y}_j|\vec{x}_i) \simeq \frac{q(\vec{y}_j) \exp\left(-\frac{d(\vec{x}_i,\vec{y}_j)}{T}\right)}{Z(\vec{x}_i, T)}. \tag{5.23}$$

La valeur de T dépend directement de la valeur de $\langle d \rangle$ par l'équation :

$$T = -\frac{\partial \langle d \rangle}{\partial I(X, Y)}. \tag{5.24}$$

Ceci nous montre bien que la classification obtenue à partir de ces critères dépend explicitement de la fonction d choisie, par le biais de la température T employée. Afin de lever cette dépendance, supposons que l'information qui nous intéresse dans la valeur de $\vec{x}_i$ à classer soit l'inférence qu'elle nous permet de faire sur un autre variable, $\vec{t}_i$. La meilleure compression dans ce cas est donc celle qui donne le plus de chances de retrouver

la bonne valeur de $\vec{t_i}$ à partir de la valeur du représentant $\vec{y_j}$. La distorsion $\langle d \rangle$ est donc remplacée par la perte d'information sur $\vec{t_i}$ par le codage de $\vec{x_i}$ en $\vec{y_j}$, qui est $-I(Y,T)$. On obtient donc à optimiser :

$$\min -I(T,Y) + TI(X,Y). \tag{5.25}$$

On peut résoudre ce problème de la même manière que précédemment. On obtient les probabilités conditionnelles optimales suivantes :

$$q(\vec{y_j}|\vec{x_i}) = \frac{q(y)}{Z(\vec{x_i},T)} \exp\left(\frac{-D\left(w(\vec{t_i}|\vec{x_i})\|w(\vec{t_i}|\vec{y_j})\right)}{T}\right), \tag{5.26}$$

où $w(\vec{t_i}|\vec{x_i})$ est la probabilité conditionnelle d'inférer $\vec{t_i}$ à partir de $\vec{x_i}$, et $w(\vec{t_i}|\vec{y_j})$ celle de l'inférer à partir du représentant $\vec{y_j}$ de $\vec{x_i}$. On voit que les représentants trouvés tendent à rendre aussi proche que possible ces deux distributions, et donc à minimiser la perte d'information sur $\vec{t_i}$. La distance qui émerge naturellement des calculs cette fois est la distance de Kullback-Leibler. Ce formalisme, appelé "information bottleneck", permet d'utiliser la distance de Kullback-Leibler comme mesure de la distorsion des données. Au vu de l'universalité de cette distance, cette méthode est très satisfaisante, car elle lève les ambiguïtés liées au choix de d. De plus, il a été montré que cette méthode était une généralisation de la méthode traditionnelle des k-means (Still et al., 2004), si la variable $\vec{t_i}$ importante était la position des vecteurs $\vec{x_i}$. Ceci donne une assise théorique importante à la méthode des k-means en la replaçant dans le cadre de la théorie de l'information, et valide en même temps l'intérêt théorique de la méthode d'information bottleneck, en montrant qu'il ne s'agit pas d'une méthode supplémentaire de classification, mais d'une généralisation de méthodes déjà connues et employées.

Cependant, cette méthode nécessite de déterminer la variable $\vec{t}$ que l'on considère comme importante, et de pouvoir calculer les probabilités conditionnelles $p(\vec{t_i}|\vec{x_i})$. Ce calcul lui même peut faire l'objet de l'application d'une méthode de classification, et rend difficile l'application de ce formalisme théorique en pratique.

Nombre de groupes maximal Cependant, un des intérêts de ce formalisme est qu'il permet de tenter de répondre à la question du nombre de groupes optimal permettant de classer les données. Comme on l'a vu, ce nombre n'existe pas *a priori*, car toute augmentation du nombre de groupes va augmenter la précision de la classification. Cependant, l'hypothèse a été émise que l'imprécision des données elles-mêmes devait être prise en compte dans le choix du nombre de groupes. En effet, des données très bruitées peuvent ne représenter que quelques groupes tout en étant toutes apparemment très différentes. Le nombre maximum de groupes utilisé dans un problème de classification devrait donc être inversement proportionnel au bruit supposé sur les données.

Par une approche perturbative, en utilisant le formalisme de l'information bottleneck, il a été montré qu'on pouvait effectivement inférer le nombre maximum de groupes en lesquelles les données devaient être partitionnées (Still and Bialek, 2004). Ceci n'est qu'une borne supérieure, au delà de laquelle toute partition supplémentaire a plus de chances d'être due au bruit et pas à la nature des données. Techniquement, on calcule la perte d'information sur la classification créée par le bruit, et on observe que l'information contenue dans la description par les représentants $\vec{y_j}$ a un maximum. De ceci on obtient la valeur

du paramètre de Lagrange $T^\star$ qui optimise la description. Cette valeur est bornée de la manière suivante :

$$\frac{1}{2N}2^{I(X,Y)} < T^\star < \frac{K}{2N}2^{I(X,Y)}. \tag{5.27}$$

Dans cet encadrement K est le nombre d'états obtenus de la variable $\vec{t}$ qui est la variable d'importance. La valeur de $T^\star$ peut ensuite être reliée au nombre de groupes de la classification, via le calcul des $q(\vec{y}_j)$.

La même méthode ne peut pas être appliquée à la théorie de taux-distorsion directement. En effet, dans ce cas, on ne peut pas supposer d'erreur systématique sur le calcul de $\langle d \rangle$, alors que dans le cas de l'information bottleneck, on pouvait aisément supposer que les probabilités conditionnelles $w(\vec{t}|\vec{x})$ étaient bruitées. Dans ce formalisme plus général, il n'est pas possible d'obtenir un optimum sur T, et donc sur le nombre de groupes. On doit alors utiliser les travaux précédemment décrits (Rose et al., 1990) pour estimer, pour chaque nombre de groupes, la valeur du paramètre T qui maximise l'information conservée sur les éléments $\vec{x}_i$. L'usage des critères extérieurs dans le choix du nombre de groupes est donc encore d'actualité, même si les approches décrites ici ont permis d'éclaircir les fondements théoriques qui le dirigent.

Troisième partie

Mes travaux

Chapitre 6

Classification de gènes par leur biais d'usage de codons

6.1 État de l'art

Comme on l'a vu dans la section "Mesures du biais de codons", page 70, classer les gènes ou les génomes les uns par rapport aux autres en fonction de leur usage du code est une activité qui a occupé de nombreux chercheurs. En effet, comparer le biais de codons de plusieurs gènes permet de faire de nombreuses inférences sur leurs fonctions et leurs origines ; de même, comparer les biais généraux que l'on trouve à l'échelle du génome entre différentes espèces permet d'évaluer certaines des pressions de sélection auxquelles elles sont soumises. On peut en déduire une partie de leur histoire évolutive, ou dans quel environnement elles se sont développées. Dans ce but de nombreux indicateurs numériques ont vu le jour, avec des résultats plus ou moins précis.

Il n'existe pas que ces indicateurs numériques pour évaluer le biais d'usage de codons. En effet ils ne reflètent que partiellement la distribution des fréquences dans un gène ou un génome, et leur emploi induit nécessairement une perte d'information. Une autre méthode a été très employée pour évaluer le biais de codons à l'intérieur des génomes : il s'agit de l'analyse factorielle des correspondances. En quelques mots, l'idée est de projeter les données, à l'origine dans un espace d-dimensionnel, sur les 2 (ou 3) axes de l'espace sur lesquels elles ont la plus grande variabilité. Pour l'analyse de l'usage des codons dans un génome, par exemple, on peut considérer chaque gène g_i comme un ensemble de 61 valeurs, les 61 fréquences des codons qu'il contient (ou 59 si l'on ne veut considérer que les acides aminés dégénérés, ce qui exclut la méthionine et le tryptophane, représentés par un seul codon). Alors on peut se représenter ce gène comme un point dans un espace à 61 dimensions – on peut se restreindre à 60 dimensions, la dernière coordonnée étant déterminée par normalisation. En fonction de leur usage de codons, différents gènes seront plus ou moins proches dans cet espace. Géométriquement, l'analyse des correspondances trouve le plan $\mathcal{P}$ à deux dimensions qui maximise l'étalement des projections des points g_i sur $\mathcal{P}$. Ensuite, on peut représenter sur $\mathcal{P}$ les projections des gènes, et visuellement – ou analytiquement – regrouper les gènes en fonction de leur proximité dans cette projection. Ce type d'analyses a été effectuée sur *E. coli* (Médigue et al., 1991) et *B. subtilis* (Moszer et al., 1999) ainsi que sur beaucoup d'autres espèces bactériennes (voir Perrière and Thioulouse (2002) et les références citées). Elles ont permis de distinguer trois groupes de gènes, relativement à leur usage du code :

- Les gènes fortement exprimés, codant pour les protéines ribosomales par exemple. Leur usage du code est dirigé par l'abondance relative des ARNt.
- Les gènes du métabolisme.
- Les gènes ayant une origine extérieure à l'organisme, dont l'usage du code est nettement différent de tous les autres.

La découverte et l'identification des gènes de ce troisième groupe ont été l'un des succès de l'analyse des correspondances dans ce domaine. Les gènes acquis récemment par un organisme, par transfert horizontal, n'ont pas eu le temps de s'adapter à son usage du code. Ils sont encore marqués par leur biais de codons effectif, qui dépend de leur histoire et des passages qu'ils ont pu faire dans différents organismes auparavant, et ceci permet de les détecter. Cependant, l'efficacité du transfert horizontal entre deux organismes semble nécessiter une certaine similarité de biais de codons, au moins à l'échelle des gènes transférés (Medrano-Soto et al., 2004).

L'analyse des correspondances n'a pas été employée que pour la classification de gènes : elle a également été employée à l'échelle des génomes, pour détecter précisément les pressions de sélection s'y appliquant (Carbone et al., 2005) ; ou pour classer en fonction de leur composition en acides aminés les gènes d'un génome (Lobry and Gautier, 1994; Pascal et al., 2005, 2006), ce qui a permis d'identifier les protéines membranaires comme utilisant préférentiellement certains acides aminés particulier ; ou encore pour classer les génomes entre eux en fonction de leur composition en acides aminés (Tekaia et al., 2002), ce qui a permis d'évaluer l'influence de l'environnement sur le contenu en acides aminés des génomes. Cependant, des biais méthodologiques associés à l'emploi de l'analyse des correspondances ont été mis à jour, en particuliers dûs à l'influence du contenu en acide aminés rares sur les classifications et sur les précautions à prendre lors de la normalisation des données (Perrière and Thioulouse, 2002).

6.2 Notre méthode de classification

Nous avons donc décidé d'aborder le problème de la classification des gènes en fonction de leur usage de codons d'un œil nouveau, inspiré de la théorie de l'information. Nous avons donc développé un algorithme de classification, qui répartit les gènes en fonction de leur usage de codons dans des groupes, et qui définit par un critère objectif le nombre de groupes et leurs frontières. Ceci permet d'outrepasser le défaut de certaines des analyses par correspondances, où le nombre de groupes – et donc leurs délimitations – est défini visuellement en étudiant la représentation des gènes sur le plan à 2 dimensions. De plus, notre algorithme présente d'autres avantages :

- Il n'y a pas de perte d'information par réduction dimensionnelle, car l'analyse a lieu dans l'espace à 59 dimensions directement, et pas sur le plan.
- Les comptes des codons sont explicitement utilisés, pas seulement leurs fréquences dans le gène. Ceci pose un problème en analyse des correspondances, car si des gènes de longueurs différentes peuvent être étudiés à partir des comptes, à condition de normaliser correctement chaque colonne de la matrice des données, ceci n'est pas toujours bien fait (Perrière and Thioulouse, 2002). Le même oubli – ou emploi simplifié – de l'analyse des correspondances peut advenir dans l'emploi des fréquences de codons comme variable. Ceci a pour conséquence une perte d'information sur la longueur des gènes, et modifie l'importance relative des gènes de longueurs différentes

sur la classification à cause de problème d'échantillonnage : les fluctuations non significatives d'usage de codons sont plus importantes sur les gènes courts, et un gène très court de 10 résidus formé à 10% d'un codon particulier doit moins biaiser la classification qu'un gène ayant la même fréquence d'usage de ce codon et long de 1000 triplets.
- Le choix du nombre de groupes est défini par un critère objectif basé sur la stabilité des groupes formés. Aucune information biologique n'est incluse *a priori* dans la classification : l'algorithme n'a pas d'idées préconçues sur ce qu'il s'attend à observer[1]. Ceci permet d'affirmer que les groupes observés ne sont formés que sur la base du biais d'usage de codons, et donc d'en déduire toutes les propriétés qui y sont reliées.
- Notre algorithme n'emploie pas de distance entre gènes, comme celle qui est implicitement utilisée en analyse des correspondances. Au lieu de cela, il optimise une propriété basée sur les groupes dans leur ensemble, et pas sur des relations par paires des gènes les composant.

Nous allons résumer ici comment fonctionne notre méthode de classification, et sur quelles bases théoriques elle se fonde. Nous ne présenterons ici que les principes, les calculs étant présentés dans l'article et détaillés dans l'annexe B, page 237.

6.2.1 Critère de partition

Notre méthode de classification se base sur la maximisation d'un critère, basé sur la répartition des gènes dans les différents groupes, et mesurant le gain d'information apporté par la classification en fonction du biais d'usage de codons, par rapport à une classification homogène. Nous nous plaçons dans un cadre bayesien, et calculons la distribution des probabilités d'usage de codons dans chaque groupe qui maximise les chances d'observer les gènes classés à l'intérieur. *A priori*, la distribution d'usage de chaque codon est supposée uniforme dans chaque groupe, car aucun gène n'est classé. Après la classification des gènes dans les groupes, cette distribution a changé, et s'est biaisée de façon à refléter au mieux les gènes que contient chaque groupe : c'est la distribution *a posteriori*. Le gain d'information est mesuré en comparant la distribution des probabilités d'usage des codons dans chaque groupe avant et après classification. La quantité maximisée est l'écart entre ces deux distributions, considérée comme la distance de Kullback-Leibler les séparant.

Ceci nous permet de trouver le critère à maximiser. Cependant, il est impossible de tester toutes les classifications et de trouver celle qui le maximise, le nombre de classifications de plus de 4000 gènes en S groupes étant trop nombreuses. Un algorithme spécial a donc été développé et testé, inspiré à la fois des algorithmes de classification hiérarchique et des algorithmes de type k-means. Cet algorithme a été préféré à des méthodes de recuit simulé pour des questions de vitesse de convergence, sans que cela ne modifie en rien les conclusions.

Notre algorithme fonctionne en deux temps. Tout d'abord, par une méthode de classification hiérarchique modifiée (voir ci-dessous), il construit une partition pour chaque nombre de groupe possible. Ensuite, un critère de stabilité est calculé sur chaque partition, permettant de désigner celle qui est la plus représentative de la réalité.

[1] Contrairement au chercheur, qui malgré toute son objectivité interprète toujours les résultats à sa manière.

6.2.2 Algorithme de classification

Notre algorithme de construction des classifications est hiérarchique : il construit successivement une série de partitions de moins en moins fines, en fusionnant à chaque étape les deux groupes dont la réunion maximise le critère défini précédemment. Ce critère étant basé sur les comptes des codons à l'intérieur de chaque groupe, le calcul de toutes les fusions possibles est aisé, et se fait simplement en mesurant la différence induite par l'enlèvement des deux groupes à fusionner et le gain apporté par leur fusion. Ceci permet à l'algorithme d'être rapide, surtout quand de nombreuses fusions sont possibles.

Afin que notre classification ne soit pas piégée dans un maximum local, une phase de réallocation dynamique itérative a lieu lorsque le nombre de groupes est suffisamment faible, à partir de 40 dans l'article. À cette étape, chaque gène est successivement déplacé dans chaque groupe – d'où l'intérêt de commencer cette procédure quand le nombre de groupes est faible – et la nouvelle valeur du critère de partition est calculée. Au final, le déplacement conservé est celui qui augmente la valeur du critère de partition, s'il existe. Cette phase a lieu itérativement à deux niveaux : tout d'abord chaque gène est testé comme décrit ci-dessus, puis, si au moins un mouvement a eu lieu, tous les gènes sont de nouveau testés. Ceci permet de replacer des gènes situés au niveau des "frontières" des groupes. Cette phase permet également de maximiser le critère d'information même sur les gènes dont l'usage du code est trop spécifique pour vraiment appartenir à un groupe, quand le nombre de groupes est faible.

6.2.3 Estimation du nombre de groupes

Finalement, la partition qui représente le "mieux" les données est estimée par une mesure simple de la stabilité des groupes. Techniquement, on mesure pour chaque gène la probabilité que son usage du code le fasse attribuer dans le groupe auquel il appartient vraiment. Cette valeur est moyennée arithmétiquement sur les gènes de chaque groupe, puis géométriquement sur les groupes, pour renforcer le poids d'éventuelles valeurs faibles à l'échelle d'un groupe. Ceci permet facilement de pointer, par exemple, deux groupes aux distributions de probabilités très proches : les valeurs des stabilités de chaque gène dans ce cas seront proches de 0.5, et le produit sur tous les gènes fera chuter la stabilité globale très rapidement.

Une précaution doit être prise : calculer simplement la stabilité comme expliqué au paragraphe précédent ne permet pas de trouver la partition des données ayant le nombre de groupes le plus fiable, car ce critère de stabilité est strictement décroissant en fonction du nombre de groupes. Pour résoudre ce problème, nous comparons les valeurs des stabilités obtenues à S groupes à celle obtenue sur une partition en S groupes d'un jeu de données non structuré. Pour cela, nous construisons tout d'abord ce jeu de données, à partir des probabilités moyennes d'usage de chaque codon dans le génome. Le fait de garder ces probabilités, ainsi que le nombre et les longueurs des gènes, identiques au jeu de données réel, permet de ne pas biaiser notre analyse. Ensuite, nous appliquons l'algorithme de classification sur ces données, et calculons la stabilité des groupes de ce jeu de données. Le maximum de la différence entre la stabilité réelle et la stabilité du jeu de données non structuré définit le nombre de groupes réel des données.

6.3 L'article

Voici l'article, tel qu'il a été publié dans la revue *PLoS Computational Biology* (Fig. 6.1). Ici il est suivi des figures et des tables présentées dans le journal comme matériels supplémentaires, ainsi que par l'addendum que nous avons écrit pour compléter notre analyse par la suite.

Les principales conclusions sont les suivantes :

- Notre méthode de classification permet de retrouver et d'améliorer en précision les résultats des précédentes analyses par correspondances faites sur *E. coli* K12 et *B, subtilis* (Médigue et al., 1991; Moszer et al., 1999), en trouvant 4 groupes de gènes chez la première et 5 chez la seconde.
- La similarité du biais de codons entre plusieurs gènes est très corrélée avec l'appartenance aux mêmes voies métaboliques, ou aux mêmes opérons.
- De plus, le biais d'usage de codons des gènes est corrélé à leurs fonctions. En plus de la traditionnelle association entre gènes fortement exprimés et biais de codons, que nous retrouvons, on observe que les gènes anaboliques – des gènes exprimés en situation de carence pour permettre la survie de l'organisme – de *B. subtilis* partagent le même biais. Ceci est à mettre en relation avec les travaux dont nous avons parlé à la section "Le paradoxe des codons rares", paragraphe 3.5.c, page 84 sur les codons sensibles ou insensibles, et la différence de contenu cellulaire en ARNt en fonction des conditions extérieures.
- Les gènes partageant le même biais de codons ont une tendance à être voisins sur le chromosome, aussi bien chez *E. coli* que chez *B. subtilis*. Ces corrélations spatiales s'étendent sur des distances en moyenne cinq fois plus grandes que la longueur typique des opérons. Ce phénomène que nous avons mis en avant pour la première fois pourrait être expliqué par une sélection pour une régulation de la traduction en fonction du contexte d'expression génétique.

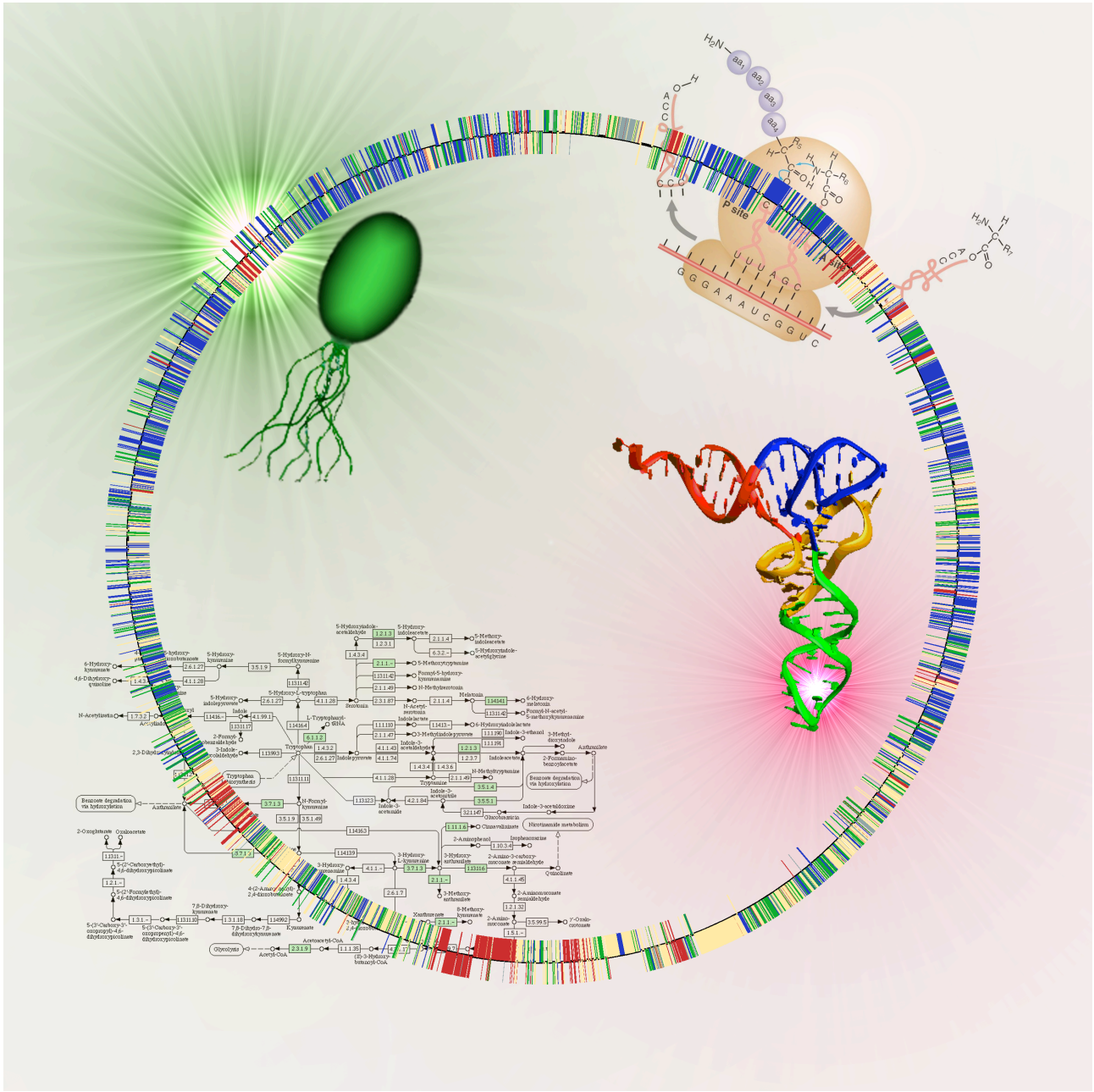


FIG. 6.1 – Image que j'ai préparée et proposée à la revue *PLoS Computational Biology* pour la couverture du numéro contenant notre article.

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Codon Usage Domains over Bacterial Chromosomes

Marc Bailly-Bechet[1], Antoine Danchin[2], Mudassar Iqbal[3,4], Matteo Marsili[3], Massimo Vergassola[1*]

1 CNRS URA 2171, Institute Pasteur, Unité Génétique in silico, Paris, France, **2** CNRS URA 2171, Institute Pasteur, Unité Génétique des Génomes Bactériens, Paris, France, **3** Abdus Salam International Center Theoretical Physics, Trieste, Italy, **4** Computing Laboratory, University of Kent, Canterbury, Kent, United Kingdom

The geography of codon bias distributions over prokaryotic genomes and its impact upon chromosomal organization are analyzed. To this aim, we introduce a clustering method based on information theory, specifically designed to cluster genes according to their codon usage and apply it to the coding sequences of *Escherichia coli* and *Bacillus subtilis*. One of the clusters identified in each of the organisms is found to be related to expression levels, as expected, but other groups feature an over-representation of genes belonging to different functional groups, namely horizontally transferred genes, motility, and intermediary metabolism. Furthermore, we show that genes with a similar bias tend to be close to each other on the chromosome and organized in coherent domains, more extended than operons, demonstrating a role of translation in structuring bacterial chromosomes. It is argued that a sizeable contribution to this effect comes from the dynamical compartimentalization induced by the recycling of tRNAs, leading to gene expression rates dependent on their genomic and expression context.

Citation: Bailly-Bechet M, Danchin A, Iqbal M, Marsili M, Vergassola M (2006) Codon usage domains over bacterial chromosomes. PLoS Comput Biol 2(4): e37. DOI: 10.1371/journal.pcbi.0020037

Introduction

The degeneracy of the genetic code entails that all amino acids except methionine and tryptophan are encoded by multiple synonymous codons. The usage of synonymous codons is far from neutral, though, and strong biases in their frequencies were observed in the first genomic sequences (see [1]). A general relation of proportionality between bias and tRNA abundance was early remarked both in *Escherichia coli* and *Saccharomyces cerevisiae* for highly expressed genes [2–4]. For this class of genes, the bias is thought to be driven by the rapidity of the translation process and is quantified by a Codon Adaptation Index (CAI), gauged on the frequencies observed in ribosomal proteins and some additional genes, highly expressed under exponential growth conditions [5]. Highly and lowly expressed genes are clearly separated in two different groups by multivariate cluster analysis [6].

Expression levels do not exhaust the possible sources of selective pressures on protein encodings. For example, proteins synthesized under conditions of starvation for certain amino acids obey rather different principles of selection. Mazel and Marlière [7] showed that, under conditions of sulphur limitation, the most abundant proteins of the cyanobacterium *Calothrix* are encoded so as to reduce their sulphur requests. More recently, Elf et al. [8] have shown that when the codon reading is part of a control loop that regulates synthesis of a starved amino acid the codon choice seems to be as sensitive as possible to starvation.

Furthermore, a possible role of the translation kinetics and codon usage for a proper folding of the nascent protein was proposed by Thanaraj and Argos [9,10]. Finally, a whole class of genes known to have a specific type of bias is composed of horizontally transferred genes, as shown using multivariate correspondence analysis [11,12]. This remark was subsequently used to trace back the evolutive origin of outer membrane genes in *E. coli* [13] and to identify biases in the functions of horizontally transferred genes [14]. While general properties of codon usage have been considered in great detail, little information is available on the global organization of the bias over the chromosomes. This is the issue broached in the present paper. The methodology that we employ is to cluster genes according to their codon bias and analyze the resulting groups. This procedure has a twofold advantage.

First, it allows identifying groups of genes sharing a similar codon usage and, looking at their composition, inferring the possible causes of the observed biases. Second, information on the codon usage of the various genes is condensed into their cluster membership, whose correlations and distribution over the chromosome are most conveniently analyzed. General-purpose multivariate methods for clustering genes according to their codon usage have been reviewed by Perrière and Thioulouse [15], who raised a list of relevant points on their limitations. In particular, the counts of the various codons for the different genes are highly variable and might be rather low for some amino acids.

Standard choices for the distance between couples of genes are therefore doomed to strongly fluctuate and possibly to lead to artifacts. Furthermore, no objective criterion is usually provided to choose the number of clusters. Those points motivated us to devise a new clustering method, specific to the problem of codon bias analysis. The procedure is presented in detail in the Materials and Methods section.

Editor: Martin Vingron, Max Planck Institute for Molecular Genetics, Germany

Received November 21, 2005; **Accepted** March 13, 2006; **Published** April 21, 2006

DOI: 10.1371/journal.pcbi.0020037

Abbreviations: CAI, Codon Adaptation Index

* To whom correspondence should be addressed. E-mail: massimo@pasteur.fr

Synopsis

Genomic sequencing projects are clearly showing that cellular components are not randomly encoded over bacterial chromosomes. Order arises for a variety of reasons. Bailly-Bechet and colleagues focused here on the role of translation in shaping bacterial chromosomes. Due to degeneracy of the genetic code, each amino acid can be encoded by multiple codons. Gene encoding is not random, though, and, depending on the genes, some codons are preferred to their synonyms. This is the so-called codon bias phenomenon. The authors analyzed the usage of synonymous codons for protein encoding and its geography over bacterial chromosomes. They found that genes sharing similar codon bias tend to be close to each other on the chromosome, in coherent patches more extended than transcriptional units. Their hypothesis is that those correlations in codon bias enable the cell to locally recycle tRNAs employed during translation, reducing stalling of the ribosomes due to rare tRNAs. This also entails a dependence of expression rates of a gene on its chromosomal context. Furthermore, their analysis made clear that genes involved in anabolic pathways, mainly active when the cell is starving, have a similar codon usage, and that they are encoded on the lagging strand of DNA. They hypothesize that this is due to relative translation efficiency of the lagging strand as compared with the leading one, illustrating the role of translation in creating structural evolutionary constraints.

The basic idea is to assign *all* coding sequences of a genome to S clusters and look for the best partition in terms of information content. Each cluster is characterized by its own distribution of codon usage, i.e., the probabilities of using a given codon to encode a given amino acid, and the distribution is supposed to be common to all the coding sequences composing the cluster. The number of clusters S is determined by a systematic criterion based on cluster stability.

The Results section presents the application of the new method to the coding sequences of the two most-studied representatives of gram-negative and gram-positive bacteria, *E. coli* and *Bacillus subtilis*. The analysis of the clusters so identified indicate that they are, both statistically and biologically, highly consistent and that our clustering method significantly improves over previous works. The biological significance and implications of the clusters are further investigated in the Discussion section, where we discuss the possible mechanisms yielding the strong and extended correlations in codon bias observed over the chromosomes and the implications for chromosomal organization.

Results

The clusters obtained by our new clustering method for *E. coli* and *B. subtilis,* and their geography over the chromosomes, will be presented in the following subsections.

Cluster Structures in *E. coli* and *B. subtilis*

The number of clusters identified for *E. coli* K12 and *B. subtilis* are four and five, respectively, as shown by the curves in Figure 1. In Figure 2, the posterior average probabilities of codon usage for phenylalanine, threonine, and valine are reported. These three amino acids are chosen as others are either more rare (C,H,Y), have their codons enriched in GC bases (A,G,P), are affected by deamination processes (N,Q), or have a biased distribution along the proteins (D,E,K) [16]. Probabilities of usage for all amino acids are reported in Tables S1 and S2. In Figure 3, we report the posterior probability distributions for three codons of the previously mentioned amino acids phenylalanine, threonine, and valine. The curves show that the clusters are indeed well-separated and that the separation arises by the combined effect of the various codons is not dominated by a single one. An important point is that the clustering is not due to trivial differences in GC content between genes, as the average GC content of the genes in the various clusters varies only from 49.28% to 49.32% in *E. coli,* and from 42.10% to 42.18% in *B. subtilis*.

Strong indications in favor of the biological significance of the clusters stem from three different statistics: the Codon Adaptation Index (CAI), the distribution of the cluster memberships among genes composing operons, and their distribution among genes coding for proteins intervening into a common metabolic pathway. As for the CAI [5], genes used to gauge the index are all highly expressed and share codon usages strongly biased toward the most abundant tRNA iso-acceptors expressed under exponential growth conditions [2,3]. Those genes are therefore expected to co-cluster. Indeed, we find that the great majority (32/59) of genes used to gauge the CAI index [17] for *B. subtilis* belongs to the first group in Figure 2 (the complete list of the cluster memberships for the CAI genes of *B. subtilis* is available in Table S5). The statistical significance of the event is very high (gathering 32 genes or more in the first cluster has a probability of 10^{-29} to occur by chance).

For *E. coli* K12, the co-clustering of its genes used to gauge the CAI index [17] is even stronger, as they all belong to the first group in Figure 2, and the event has a probability 10^{-44} to occur by chance. Genes belonging to operons are co-transcribed in a polycistronic mRNA molecule, and they are then expected to share similar pressures on the translation process. Exceptions and special cases ought to be expected for various reasons: genes transcribed from alternative promoters, different folding kinetics and expression levels, and differential regulation of the translation process among the various genes of the operon, etc. For example, genes within the *gal* operon of *E. coli* are involved in functions only partially overlapping and their polarity is regulated by the Spot42 noncoding RNA [18].

It is, however, expected that at least on a global statistical level, genes within a common operon should display a tendency to share a similar usage of codons, i.e., co-cluster. The same tendency is expected for genes belonging to common metabolic pathways, as their expression tends to be correlated, namely in time. Indeed, considering the list of known operons and metabolic pathways and comparing their cluster memberships to null models generated as described in Materials and Methods, we obtain the results shown in Figure 4. Genes belonging to common operons and/or metabolic pathways have a strong tendency to share the same cluster membership. The observed values of the z-scores (8.9, 15.7 for *E. coli,* and 15.6, 43.9 for *B. subtilis*) correspond to extremely low p-values (3×10^{-19}, 8×10^{-56}, 4×10^{-55}, and exp (–968.3), respectively), and our clustering method manifestly allows significant improvements over previous results obtained by general-purpose multivariate clustering methods [11,12].

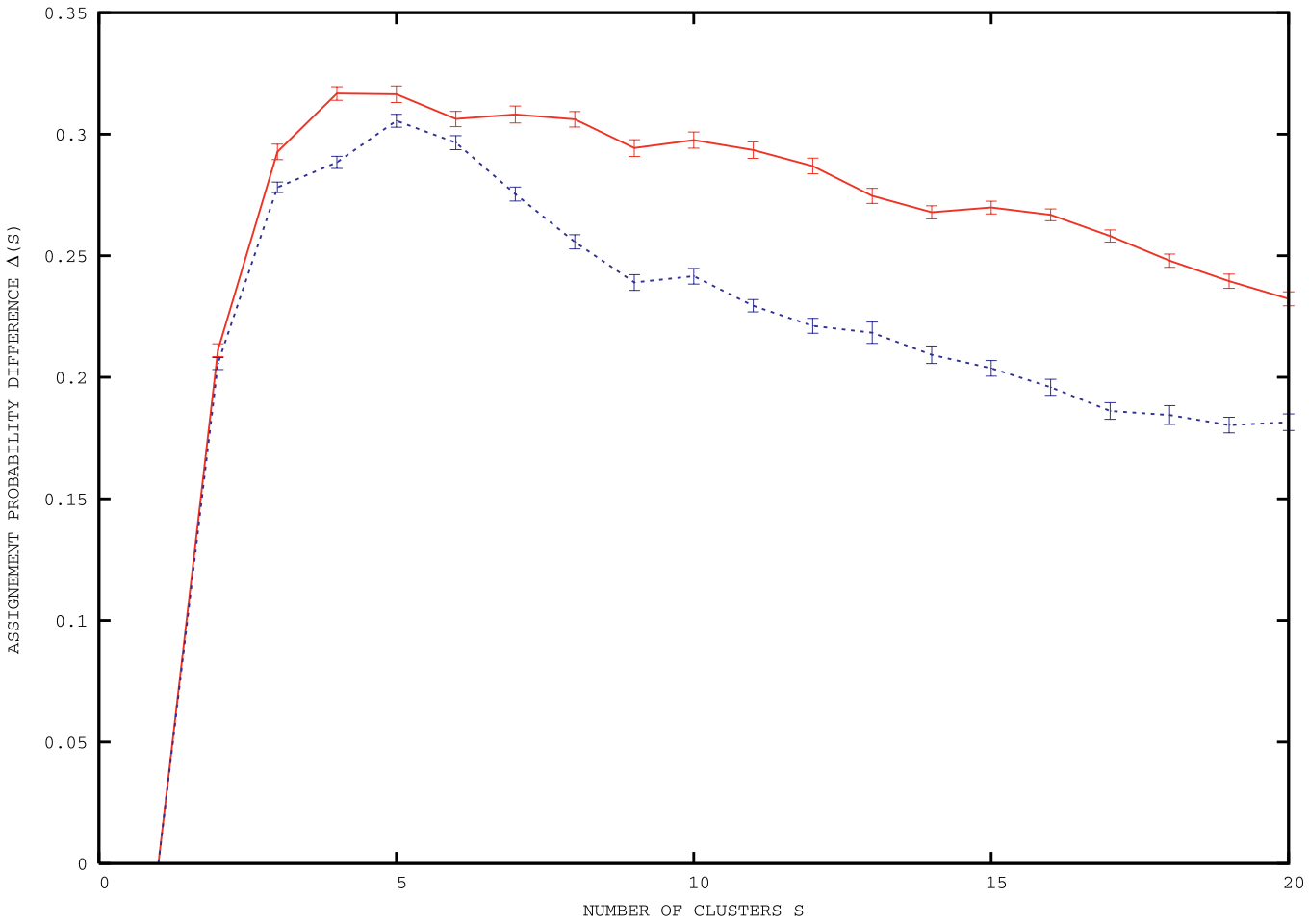


Figure 1. The Cluster Stability Curves, Quantified by the Difference $\Delta(S) = B(S) - B_{random}(S)$ of the Assignment Probabilities Defined in the Body of the Text, versus the Number of Clusters *S*

The curves are for *B. subtilis* (dashed blue) and *E. coli* K12 (solid red). The retained number of clusters corresponds to the maximum of the stability curve.
DOI: 10.1371/journal.pcbi.0020037.g001

Functional Properties and Distribution over the Strands of Genes in the Clusters

Clusters identified in the previous subsection have marked properties regarding the functional categories of their genes. As previously shown, the first groups in Figure 2 contain an overwhelming number of highly expressed genes involved in translation, ribosomal structure, and biogenesis. This was largely expected on the basis of known results [2,3,5]. More interestingly, other clusters, too, have quite specific properties in terms of the functional categories of their composing genes. A systematic analysis is performed using COG functional annotations [19] and looking at the composition of the various clusters. Deviations from the behavior expected by chance are assessed using artificial chromosomes generated as described in the Materials and Methods section. The results are reported in Tables S3 and S4.

A first class of genes whose distribution is highly non-homogeneous across groups is that of genes poorly characterized and/or of unknown function (COG classes –, R, and S). Indeed, a striking excess of those genes is found in the second groups of both *B. subtilis* and *E. coli* K12. A more detailed analysis reveals that a great deal of them are in prophage, mobile, and horizontally transferred regions. Furthermore, when the two previous groups are compared to the "horizontally transferred" groups previously found in [11,12], a large overlap is found. This confirms the special usage of codons by horizontally transferred genes and the possibility of detecting them by their codon bias.

Another class of genes which we find to be biased is composed of genes involved in the motility of the cell (COG class N). They also feature a peculiar usage of the codons, appearing preferentially in the fifth cluster of *B. subtilis*. A third, large class of genes with a special distribution among the clusters is composed of metabolic synthesis and transport genes. The third group in Figure 2 for *B. subtilis* features indeed a significant excess of genes belonging to the COG categories C (energy production and conversion), E (amino acid transport and metabolism), and F (nucleotide transport and metabolism).

The fourth group also contains an excess of genes involved in carbohydrate transport and metabolism (the COG G category). Metabolic genes in *E. coli* also tend to gather in the third group, with significant overabundances of genes belonging to the COG categories C, E, H (coenzyme transport and metabolism), and P (inorganic ion transport and metabolism). Deviations to the random values for those

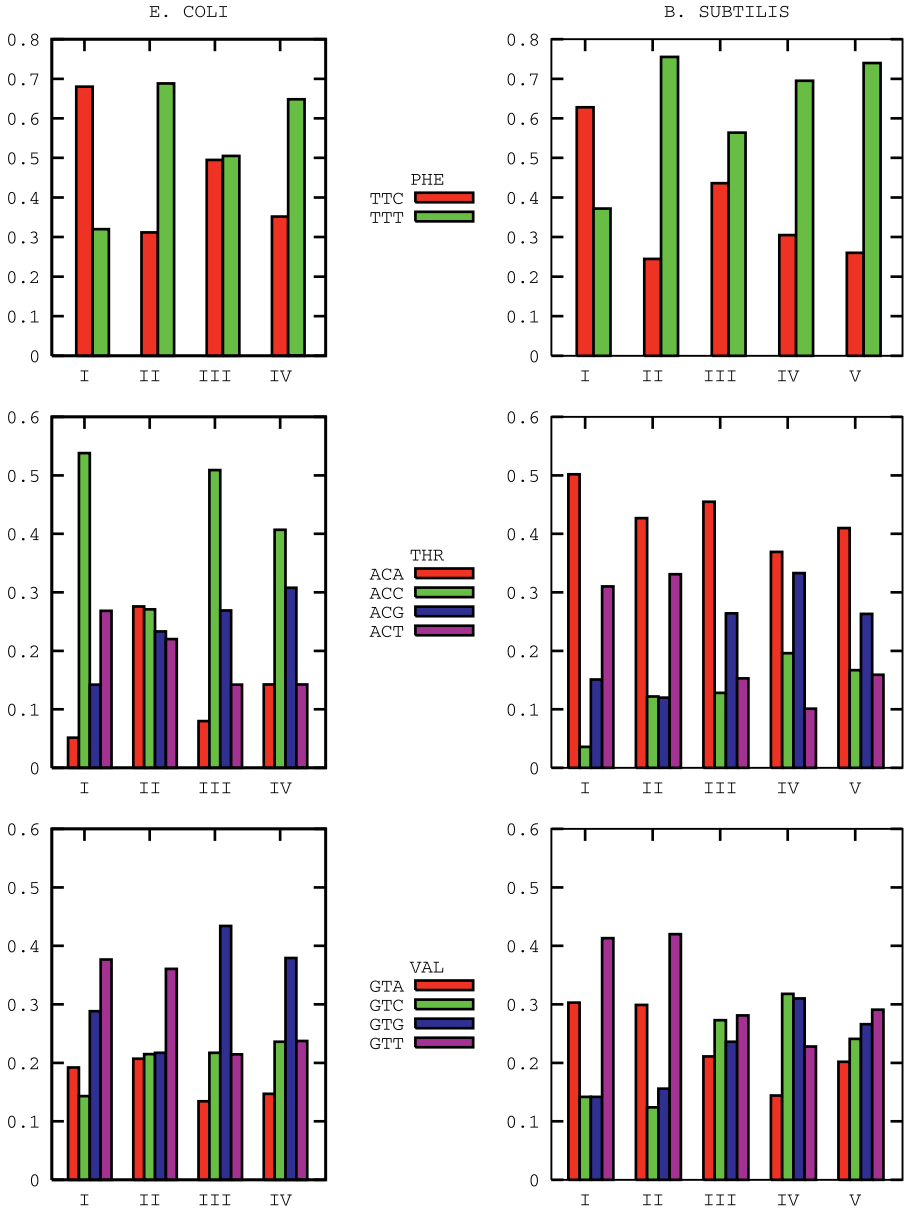


Figure 2. Average Posterior Probabilities of Usage for the Codons of Phenylalanine, Threonine, and Valine in the Clusters Identified for *E. coli* K12 and *B. subtilis*

E. coli K12, left column; *B. subtilis,* right column.
Clusters are identified by a roman number on the *x*-axis. The corresponding standard deviations are on the order of a few percent of the average values.
DOI: 10.1371/journal.pcbi.0020037.g002

classes are highly significant, with *z*-scores all larger than 3.4 and soaring up to 6.5.

In addition to genes coding for cytoplasmic metabolic genes, we find that many genes in this class code for transport systems. The corresponding proteins are associated with the bacterial envelope, a compartment that is significantly smaller in volume than the cytoplasm, asking for a consistently smaller number of individual proteins. Whether this quantitative feature is relevant to our observation remains to be seen. The functional properties just presented appear even more relevant if coupled with the analysis of the strand of the genes composing the clusters, i.e., their direction of transcription as compared with the direction of the replication fork. The distribution of genes over the two strands is a major feature of prokaryotic genomes, with a dramatic asymmetry in *B. subtilis,* where about 74% of the genes are transcribed in the same direction as the replication forks, i.e., located on the leading strand of the chromosome. The global effect is weaker in *E. coli* (about 55% of the genes are on the leading strand), but specific classes of genes are known to be strongly biased, e.g., essential genes on the leading strand [20].

While most clusters do not feature any significant preference for a particular strand, a few of them do, as shown in Figure 5. The most relevant biologically (see the discussion in the next section) is the strong overabundance of genes on the lagging strand found in the third cluster of *B. subtilis*. The strand asymmetry emerges also from the codon usage posterior probabilities (see Table S2). Indeed, leading

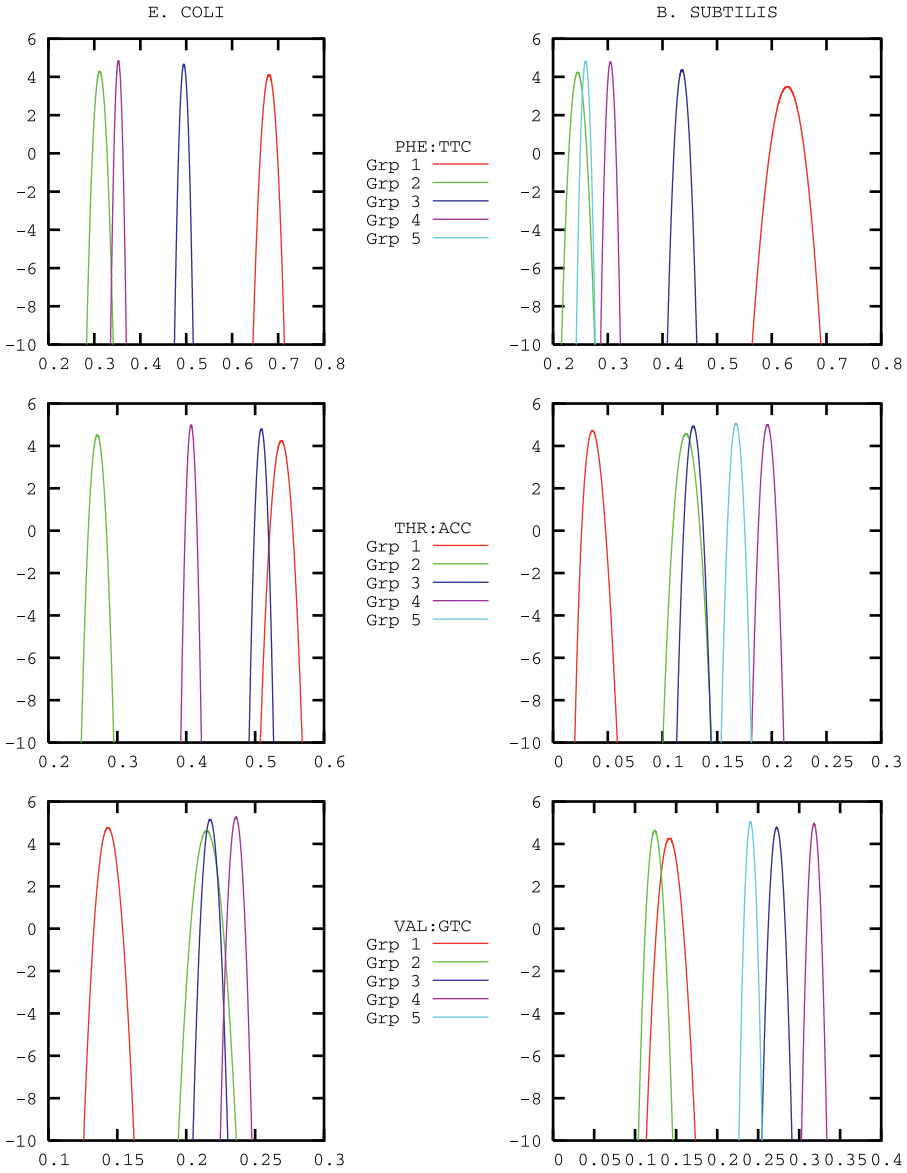


Figure 3. The Posterior Probability Distributions for Three Representative Codons: TTC (Phenylalanine), ACC (Threonine), and GTC (Valine) in the Clusters That We Identified for *E. coli* K12 and *B. subtilis*

E. coli K12, left column; *B. subtilis,* right column.
The curves are meant to show that the clusters are well separated by the combined information on the various codons.
DOI: 10.1371/journal.pcbi.0020037.g003

and lagging strands have a marked excess of guanines and cytosines, respectively, violating the naïve expectation of an equidistribution [21,22]. The reason is that the two DNA strands are exposed as single strands for quite different lags during replication, due to the kinetics of the formation and ligation of the Okazaki fragments. That induces different rates and dynamics in the mutation and repair processes, eventually leading to the observed G/C asymmetry (see [23,24] and references therein). In conclusion, the third cluster of *B. subtilis* is the same as previously shown to contain an excess of genes involved in energy production and transport and metabolism of nucleotides, carbohydrates, amino acids, and metabolites.

Correlations in Codon Usage over the Chromosomes

Let us now consider the spatial correlations of cluster memberships along the genomic sequence. The simplest relevant statistic to quantify them is the joint probability that two genes, g and $g + l$, belong to the same cluster ($s_g = s_{g+l}$):

$$P2(l) = \langle \delta | (sg, sg + l) \rangle - \sum_{i=1}^{s} f_i^2, \qquad (1)$$

where δ is the Kronecker delta function, S is the number of clusters and fi is the total fraction of genes belonging to the i-th cluster. The asymptotic value $\sum_{i=1}^{s} f_1^2$, corresponding to decorrelation between the two positions, is subtracted to ensure that the function in Equation 1 decays to zero at large distances, as shown in Figure 6. Note that genes are ranked in increasing order with respect to their translation start, so that l coincides with their spacing. In Figure 6 correlations are very extended, especially for *B. subtilis,* witnessing a similar usage of the code within rather wide domains. The most

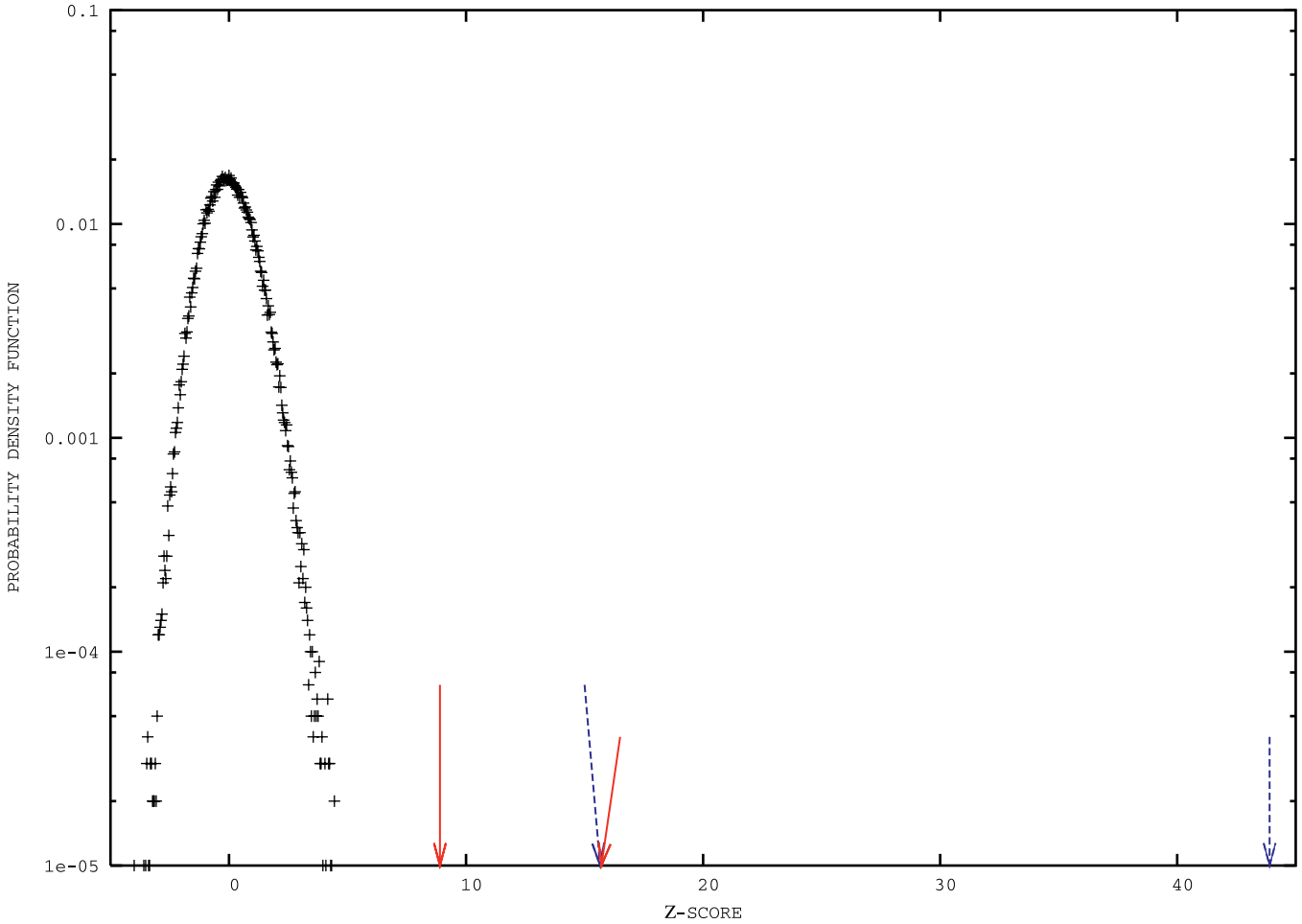


Figure 4. A Centered Gaussian Probability Distribution of Unit Variance (Black), Corresponding to the Random Distribution Obtained in the Null Models, and the Values Actually Observed in Our Clusters (Arrows)

Values reported on the abscissae are *z*-scores, i.e., the deviations to the mean normalized by the standard deviation.
Red solid and blue dashed arrows correspond to *E. coli* K12 and *B. subtilis,* respectively. Short arrows point to the values of the *z*-scores that we measure for the fraction of pairs of genes within a common operon and belonging to the same cluster.
Long arrows refer to the same quantities for pairs of genes within a common metabolic pathway.
Note that, as the Gaussian distribution is meant to show, our *z*-scores are highly significant, e.g., $z_{score} \geq 8 \mapsto$ probability $= 6 \times 10^{-16}$ to occur by chance. See also that values of the *z*-scores previously obtained, using general-purpose clustering methods, were much smaller: 5.30 and 3.29, for operons and metabolic pathways, respectively.
DOI: 10.1371/journal.pcbi.0020037.g004

immediate possible explanation is that correlations might be simply due to constraints imposed by operons. This is, however, not the case, as shown in Figure 7. Lengths of the operons are way too short to account for the correlation lengths observed in Figure 6. Even in the case of *E. coli*, correlations extend to lengths five times larger than the average length of the operons. Alternative arguments leading to the same conclusion are presented in Figures S1 and S2. Another natural thought is that the extended correlations in Figure 6 might reflect the G/C skewed distribution. We have, however, previously remarked that variations in the GC content of the clusters are very tiny, ruling out this simple possibility. Even leaving statistics aside, a direct inspection reveals that cluster memberships are organized in coherent domains, often extending beyond the limits of known operons. Prophages and horizontally transferred regions contribute to the trend, but the coherence is not restricted to those cases and does not seem to be associated with any particular functional class or regions of the chromosome. A possible explanation of the phenomenon will be proposed in the Discussion section.

Discussion

Two results obtained that were explained in the section above seem particularly relevant to the organization of bacterial chromosomes and will be discussed here in a more extended way. The first is the extent of codon bias correlations observed in Figure 6, much longer than what could be accounted for by operons. Theoretically, the existence of long-range correlations among individual nucleotides is well-known (see [25–28]). At a higher level of organization, sequence domains of order higher than operons, dubbed über-operons or super-operons, have been evidenced in the literature [29,30]. It has been noted by Rogozin et al. [30] that sizeable minorities in super-operons do not have any obvious functional relationship to the rest of the neighborhood, but seem to "car pool" with it.

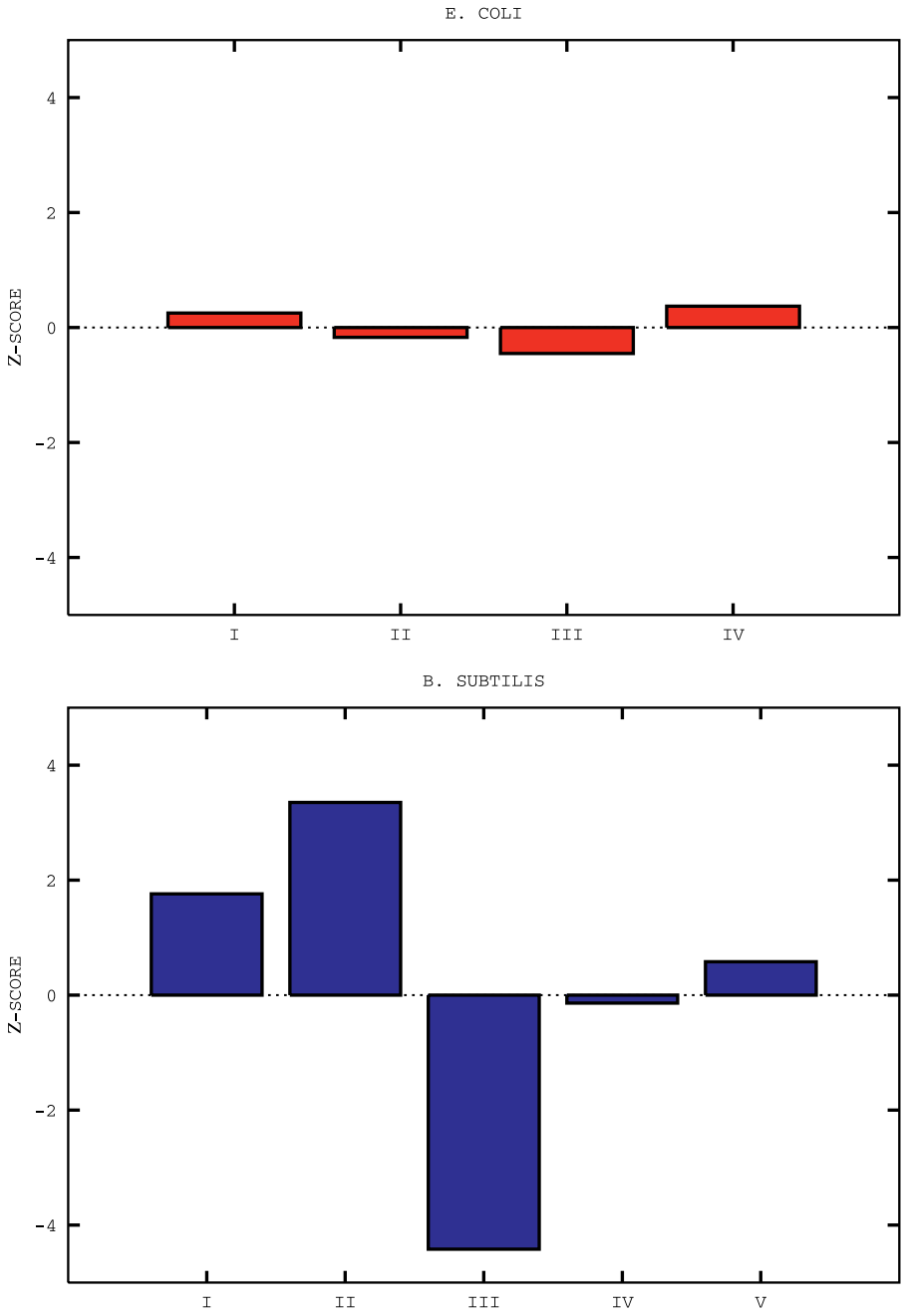


Figure 5. The Distribution of the Number of Genes on the Leading Strand for the Clusters of *E. coli* K12 and *B. subtilis*

E. coli is shown on the top graph, and *B. subtilis* is shown on the lower graph. Clusters are identified by a roman number on the *x*-axis, and *z*-scores relative to null models are indicated on the *y*-axis.

Note the depletion of leading strand genes in the third cluster of *B. subtilis*.

DOI: 10.1371/journal.pcbi.0020037.g005

Experimentally, recent data demonstrate that bacterial chromosomes have a definite spatial arrangement and are organized in macrodomains [31]. Macrodomains are playing a major role in the nucleoid organization and have strong practical implications for tentatively minimizing the size of artificial genomes [32]. A relation between these structural macrodomains and the sequence domains discussed here is plausible but remains to be demonstrated. Our results go in the direction of domains of order higher than operons. The novel point brought by our analysis is the explicit connection made between these structures and the translation process. Indeed, Figure 6 demonstrates that neighbouring genes tend to have a similar bias in their codon usage and suggests that the corresponding mRNAs reciprocally affect their translation processes. In other words, efficiency and rates of translation of mRNAs might not be a function of the mRNA only, but be quite sensitive to its genomic and expression context, too.

A sense of the relevance of these context effects might be drawn from a few simple estimates. Their goal is to assess the

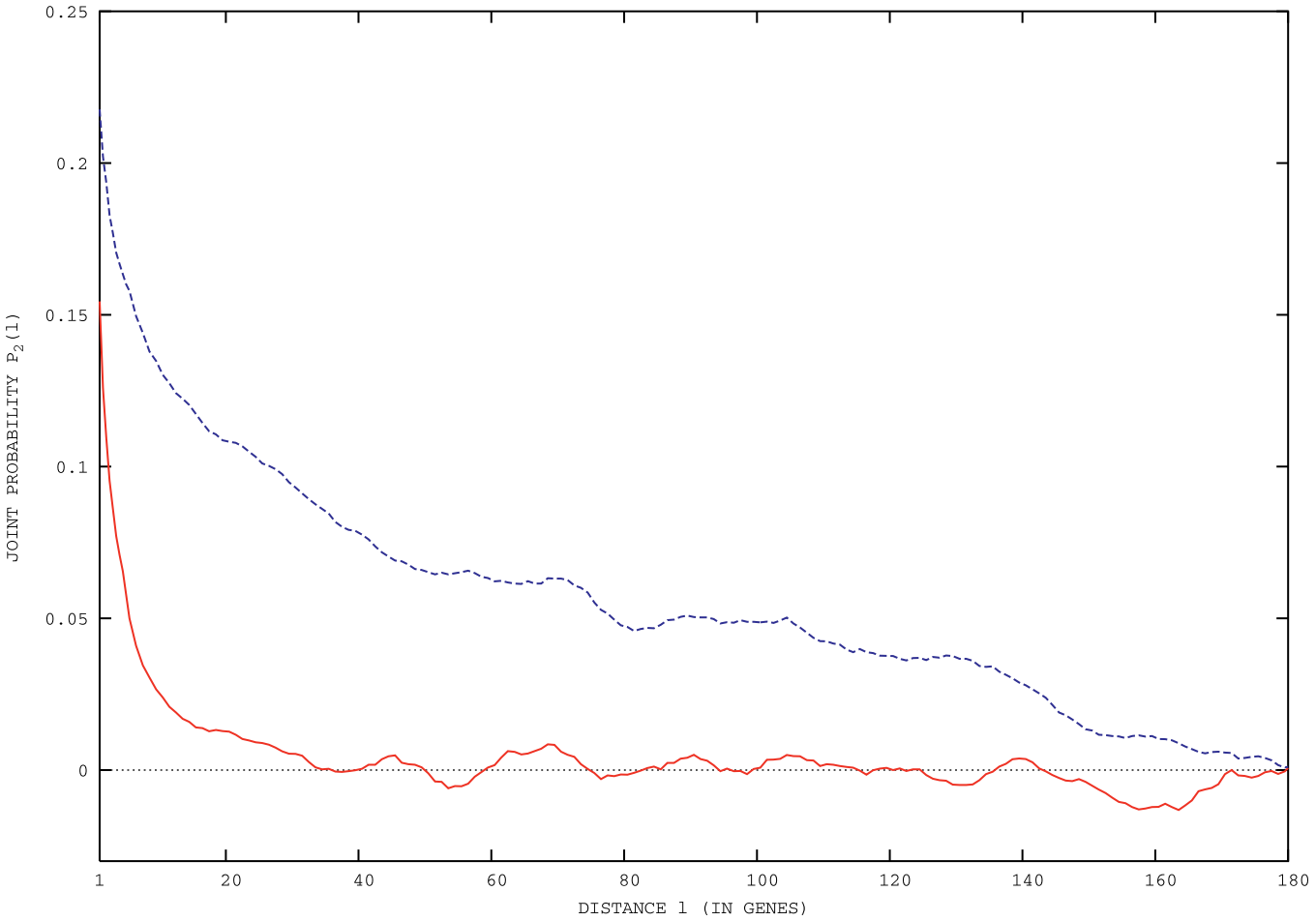


Figure 6. The Correlation Function (1) of Cluster Memberships versus the Distance among Genes for *B. subtilis* and *E.coli* K12
Blue dashed lines are for *B. subtilis,* and red solid lines are for *E. coli.*
DOI: 10.1371/journal.pcbi.0020037.g006

importance of tRNA recycling effects and the *rationale* is as follows: if the concentration of tRNAs turned out to be limiting, it would be sensible to propose that neighbouring genes tend to use codons similarly, so as to maximize their reciprocal recycling of tRNAs; conversely, if tRNAs turned out to be very abundant, it would be hard to imagine that such effects might be of any relevance. We shall suppose that tRNAs diffuse within the cell. No specific value for their diffusivity will be needed and, even though the hypothesis is likely to be an oversimplification, it should allow capturing the right orders of magnitude. The size of the cell is taken as *Scell* $\simeq$ 1 μm and the number of ribosomes $N_{ribo} \simeq 20{,}000 - 60{,}000$. The number of copies n of the various species of tRNAs in *E. coli* have been measured by Dong et al. [33] and vary from a few hundreds to several thousands. The typical distance between synonymous tRNAs is simply estimated as $lc \simeq S_{cell}/n^{1/3}$.

Let us now consider a tRNA that has just been employed somewhere in the elongation of a polypeptide chain and estimate the distance it will travel before being caught again by another ribosome. This is a classical calculation of diffusion-limited cross section, already employed in the biophysical literature to estimate the time for a transcription factor to find its target over DNA (see, e.g., [34] for a recent review). The result that we shall need is Smoluchowski's probability, $1 - 4\pi b/r$, that a particle at an initial distance r from a target of size b diffuses away from it without being caught. In our case, the targets are the ribosomes and their number will grow with r as $N_{ribo}\ (r/S_{cell})^3$. The recycling length l_{recy}, i.e., the distance r such that it is practically certain that the tRNA will be caught again by a ribosome, is obtained from the relation $N_{ribo}\ (r/S_{cell})^3 \times 4\pi b/r \simeq 0(1)$. Conservatively assuming the target size b to be 1/10 of the size of the ribosomes ($\simeq$25 nm), we come up with an estimate of $l_{recy} \simeq 0(0.1\mu m)$, comparable to the typical distance l_c for tRNAs having a thousand copies in the cell. The upshot is that the recycling of tRNAs is of importance for many of them, namely those rare and moderately abundant.

Notwithstanding the crudeness of previous estimates, there are biological indications supporting the conclusion that rare tRNAs might indeed be limiting in the translation process. Early experiments by Varenne and co-workers showed significant pauses at codons associated with rare tRNAs [35]. Another suggestive indication is the high concentration of tmRNAs, the surrogate tRNAs that append a peptide tag to nascent polypeptides and "rescue" stalled ribosomes, promoting rapid degradation of tagged proteins. Their number of copies in the cell is abundant, on the order of 13,000 [36],

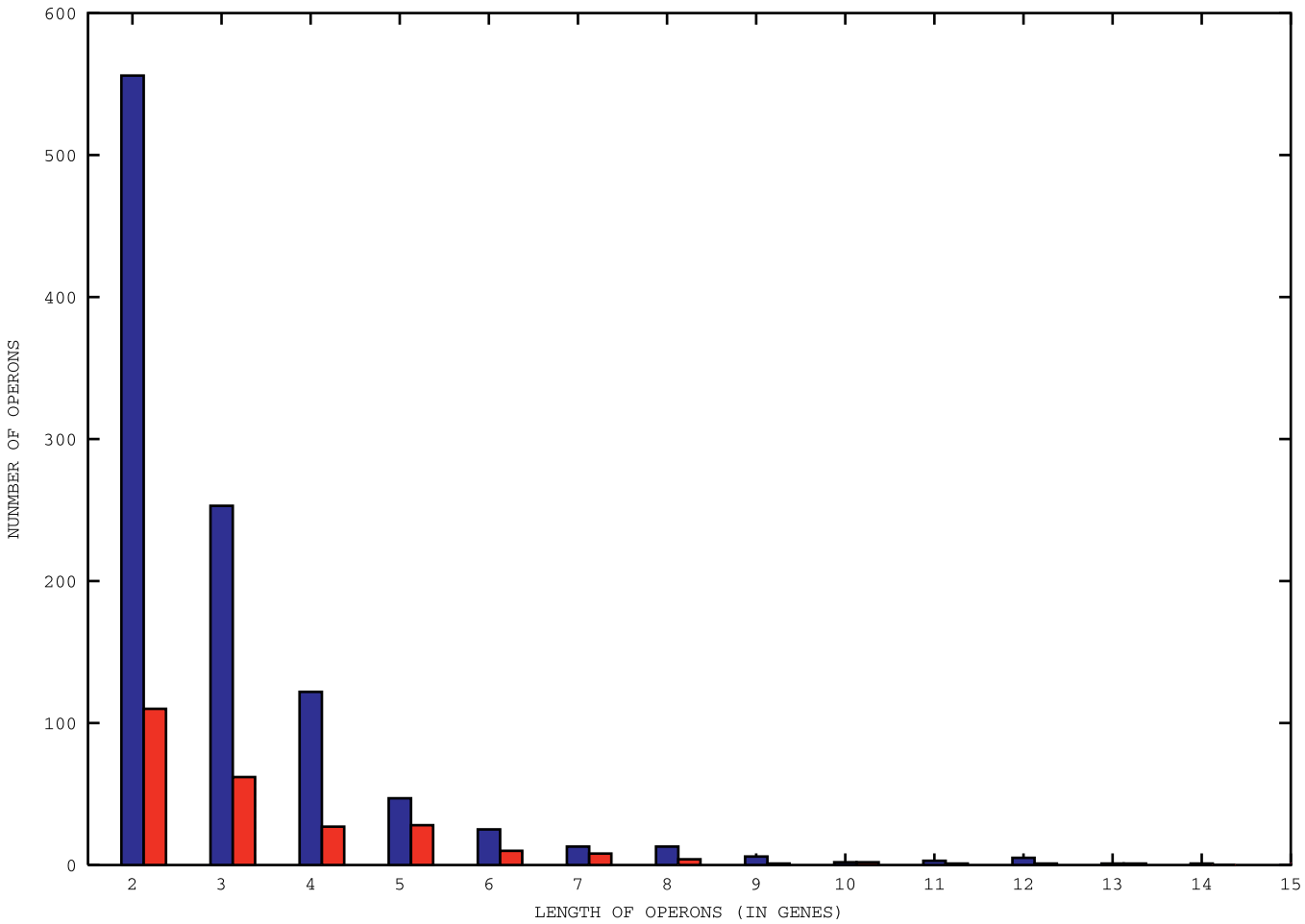


Figure 7. The Histograms of Lengths of the Known Operons for *B. subtilis* and *E. coli* K12
Blue boxes are for *B. subtilis,* and red boxes are for *E. coli* K12.
DOI: 10.1371/journal.pcbi.0020037.g007

and it was recently shown that those concentrations are safely well above saturation [37]. This witnesses the importance of ribosome stalling events, e.g., due to delays in the recruitment of rare tRNAs. The concentration of tmRNAs in the cell is in fact strikingly higher as compared with that of rare tRNAs. This suggests that some recycling of rare tRNAs ought to be at work to create higher transient local aggregations of tRNAs, compensating for their much lower average value over the whole cell. The experimental observations reported in [38], of channeling and slowing-down of the diffusion of macromolecular components of the translation apparatus, might be relevant in that respect.

The combination of all previous arguments leads us to propose a role for the codon bias domains over bacterial chromosomes that we have found, viz., that they allow a coordinated control of the expression levels of nearby genes and increase their reciprocal tRNA recyclings, so as to alleviate stalling effects. A very interesting experiment to test these ideas, yet quite difficult to design, would consist in reliably measuring possible dependencies of mRNA translation rates on their genomic and expression context. The second intriguing result presented here is the fact that anabolic genes in *B. subtilis* tend to aggregate in a single cluster and that this cluster features an excess of genes over the lagging strand. Specifically, genes in the aforementioned cluster belong to the functional classes of transport and metabolism of amino acids, carbohydrates, and nucleotides.

We shall argue that these observations are in fact strongly related and driven by the following biological mechanisms. First, genes of the previous functional classes are likely to be mostly expressed and employed in poor media, where the bacterium cannot easily import its essential metabolites from the external medium and is obliged to finely scavenge its environment and/or to synthesize them. These processes of synthesis will induce a long lag between two successive replications, in sharp contrast to the case of a rich medium. There, generation times are so rapid that bacteria are essentially always replicating, and several replicative forks are progressing at the same time over the chromosome.

Second, head-on collisions between transcriptional and replicative machineries are known to be deleterious to the proper functioning of the cell. The dynamics of the interaction between DNA and RNA polymerases have been thoroughly investigated [39–41]. Replication elongation is found to be weakly affected by co-directional transcription, whilst head-on collisions induce a severe inhibition of the replicative fork progression. It is therefore quite sensible that a strong selective pressure is at work in prokaryotic genomes

to reduce deleterious effects of head-on collisions. Those are the major cause for the strand asymmetry observed in prokaryotic genomes and, in particular, of essential genes [20]. Pressure to avoid head-on collisions seems particularly cogent in *B. subtilis,* where about 74% of all genes are found on the leading strand.

Combining the two previous remarks provides a clue to the observed preferential positioning of anabolic genes on the lagging strand: due to longer replication times in poor media than in rich media, genes expressed in the former will be subject to a relatively lower pressure to be on the leading strand as compared with genes active in rich media. Furthermore, transport proteins are located in the membranes or the periplasm, compartments that are significantly smaller than the cytoplasm, asking therefore for a significantly lower number of individual proteins of that type. There is even a strong selection pressure against too high expression of membrane proteins as reflected by the toxicity of overexpression of the corresponding genes (see [42] for a review of significant data in the domain). The resulting differential selective pressures might then contribute to the observed strand asymmetry.

This hypothesis can be directly tested by measuring the expression levels of genes, e.g., in a transcriptome experiment. The only caveat and precaution to be taken is that bacteria in the cultures should be synchronized with respect to their cell cycle, and the expression levels not be averaged out over the cell cycle, as in standard in experiments. Averaging is clearly inappropriate for genes whose expression levels strongly depend on the cycle of the cell, e.g., for the classical example of *ftsZ* [43]. Tracking the expression of those genes requires working with synchronous cultures and specific methods to meet this goal (see [44] for a review). Novel possibilities recently advanced [45] appear particularly promising and appropriate for an experimental test of the hypothesis suggested by our results, namely that genes encoded in the lagging strand direction are preferentially expressed in inter-replicative phases.

Materials and Methods

Given a set of G genes, all supposed to be translated according to the standard genetic code, our aim is to find their best partition into S clusters. More precisely, each cluster is supposed to have a common distribution of codon usage, i.e., nucleotide sequences of genes belonging to the same cluster are all supposed to be encoded with that common distribution. Our goal is to determine the cluster partition that best describes the observed counts of codons. Note that the number S of clusters is unknown and ought to be found, too. As shown in the following subsections, we shall weight the various cluster configurations by the information that they encode on the codon usage probability distributions. We shall first derive the expression of the cluster information content in terms of the codon counts. Next, we shall describe how the configurations yielding the maximum information are sought numerically and how the method here compares with methods previously employed in the literature. Finally, we shall analyze the stability of the clusters so identified and provide a quantitative criterion for choosing the number of clusters. The last subsection is a brief description of the procedures to generate random artificial chromosomes as null models.

Gathering information on codon usage distributions. The distribution of codon usage for the s-th cluster C_s is parameterized by the set of probabilities $\{p_a^{(s)}(c)\}$ that codon c be used to encode amino acid a. The degree of degeneracy for the a-th amino acid is denoted by qa, e.g., the index c runs from 1 to $q_a = 4$ for glycine and $q_a = 2$ for phenylalanine. The amino acids to be clustered are those admitting multiple encodings, so that methionine and tryptophan can be excluded without any loss of generality. The index a then runs up to $A = 18$. A priori, the only information available is that amino acids might be encoded by any one of their synonymous codons. This state of ignorance is best described by a uniform prior distribution:

$$P^{(0)}(\{p_a^{(s)}(c)\}) = \prod_{a-1}^{A} \Gamma(q_a)\delta\left(\sum_{c=1}^{q_a} p_a^{(s)}(c) - 1\right). \quad (2)$$

Dirac's δ function in Equation 2 imposes the constraint that, for each amino acid, the sum of the probabilities over synonymous codons is normalized to unity. Euler's Γ function ensures the normalization of the probability distribution, as can be easily checked using the general formula (see, e.g., [46]):

$$\int \prod_{i=1}^{K} p_i^{\alpha_i}\delta(\sum p_i - 1)dp_i = \frac{\prod_{i=1}^{K}\Gamma(\alpha_i + 1)}{\Gamma\left(K + \sum_{i=1}^{K}\alpha_i\right)}. \quad (3)$$

The uniform prior Equation 2 appears more appropriate to our situation than a prior uniform in the logarithms of the probabilities (see, e.g., [47]) as we know from the genetic code that synonymous codons are a priori all possible. Choosing a log-uniform prior would not, at any rate, modify substantially the results presented in the sequel. A posteriori, observing the codon counts of the genes assigned to the s-th cluster C_s, we can infer its posterior distribution of codon usage as:

$$p^{(post)}(\{p_a^{(s)}(c)\}) = \prod_{a=1}^{A} \frac{\Gamma(N_a^s + q_a)}{\prod_{c=1}^{qa}\Gamma(N_a^s(c) + 1)}\delta\left(\sum_{c=1}^{qa} p_a^{(s)}(c) - 1\right)\prod_{c=1}^{qa} p_a^{(s)}(c)^{N_a^s(c)}. \quad (4)$$

Here, $n_a^{(g)}(c)$ is the number of times codons of type c are used to code for amino acid a in gene $g \in C_s$ and we have used the shortcut notations: $N_a^{(s)}(c) \equiv \sum_{g\in c_s} n_a^{(g)}(c)$ for the total number of times codons of type c are used for amino acids of type a in the s-th cluster and $N_a^{(s)} \equiv \sum_{c=1}^{qa} N_a^{(s)}(c)$ for the total number of occurrence of amino acid a in cluster s. Equation 4 is an instance of Bayes theorem: the prior is given by Equation 2 and the likelihood that codon counts of gene g be generated with the probability distribution of cluster s is a product of multinomials of order q_a:

$$\mathcal{L}(g \in C_s) = \prod_{a=1}^{A}\left[\frac{\Gamma\left(\sum_c n_a^{(g)}(c) + 1\right)}{\prod_{c=1}^{q_a}\Gamma\left(n_a^{(g)}(c) + 1\right)}\prod_{c=1}^{q^a} p_a^{(s)}(c)^{n_a^{(g)}(c)}\right]. \quad (5)$$

Information acquired on the codon usage distributions of the clusters is defined in terms of the classical Kullback-Leibler relative entropy between the posterior and the prior distribution (see, e.g., [48]) as:

$$I = \sum_{s=1}^{S}[\langle\log(P^{(post)}/P^{(0)})\rangle_{post} + \langle\log(P^{(0)}/P^{(post)})\rangle_0] \quad (6)$$

where the symbols P_0 and P_{post} denote the averages with respect to the prior and the posterior distributions Equations 2 and 4, respectively. The information in Equation 6 can be calculated analytically and expressed in a simple form as a function of codon counts. To that purpose, it is sufficient to use the identity: $<\log f> = \lim_{n\to 0}\frac{\langle f^n\rangle - 1}{n}$ and Equation 3 to compute the resulting averages. The final expression is:

$$I = \sum_{s=1}^{S}\sum_{a=1}^{A}\left\{\left[\sum_{c=1}^{q_a} N_a^{(s)}(c)\Psi(1 + N_a^{(s)}(c))\right] - N_a^{(s)}\Psi(q_a + N_a^{(s)})\right\} \quad (7)$$

where we have omitted for simplicity constant terms, i.e., those which do not depend on the cluster configurations. The logarithmic derivatives Ψ of the Euler Γ function are calculated using the well-known formula [49]: $\Psi(n) = \sum_{k=1}^{n}\frac{1}{k} - \gamma - \frac{1}{n}$, with $\gamma = 0.5772\ldots$ being Euler's constant. For each number of clusters S, we aim at identifying that assignment of the G genes to the S clusters that maximizes the information in Equation 7. It is worth noting that optimizing an entropy function is quite natural for our problem. Indeed, for $G \gg 1$ and clusters sufficiently populated, posterior probability distributions are inferred from very long sequences of symbols, whose alphabet is defined by the set of synonymous codons. Since the empirical frequencies of codon usage are the types of the resulting

sequences and their fluctuations are controlled by large deviation asymptotics (see chapter 12 in [48]), the entropy of the underlying probability distributions appears indeed as an appropriate quantity to consider.

Numerical implementation and comparison to other methods. We tried several methods to optimize the information in Equation 7, and the upshot is that its landscape in biological applications considered here is not particularly rough. This permits using a simple and rapid iterative method, based on a combination of hierarchical clustering and *k*-means [50,51]. The hierarchical clustering algorithm starts from clusters composed of individual genes and iteratively proceeds upward to generate optimal configurations for each possible S number of clusters. Iterations are based on the two following steps: 1) pairs of clusters are merged so as to get the maximal I in Equation 7; 2) the resulting configuration is taken as the initial condition for a *k*-means iteration (with $k = S - 1$). Elementary moves consist of changes in the cluster of assignment for each pair of genes. Moves increasing the score in Equation 7 are accepted and the procedure is repeated until the composition of the clusters does not change anymore. We have explicitly verified that other optimization methods, e.g., simulated annealing, are more time-consuming and do not modify the results in any substantial way. Let us conclude this subsection with a brief discussion on the choice and the comparison of our clustering method with previous works. As we have just discussed, the numerical method of optimization relies on the combination of two standard and commonly employed methods (*k*-means and hierarchical clustering). Conversely, the choice of the quantity to be optimized in Equation 7 is less usual. A more standard procedure would be to define a distance among pairs of genes and then minimize the sum of the intracluster distances. If counts of events are involved, as in codon bias clustering, classical choices for the pair-wise distance are the Euclidean distance between synonymous codon usage values or between percentage codon usage values [6] and the χ^2 metrics employed in [11,12]. Our motivation for going through the derivation leading to Equation 7 is that the counts of the codons feature a large variability over the various genes and that they can be rather low for some of the amino acids. The former implies that statistics such as percentage usages do capture average effects but are not quite rigorous in their accounting for the fluctuations: the same difference in percentage usage between two genes might indeed be highly significant or not, depending on the total number of counts involved. As for the A_2 metrics, its general relevance relies on the limit of a large number of counts, a hypothesis which is not verified for all amino acids in some of the genes. Possible consequences of enforcing χ^2 metrics with a low number of counts are described in [15], showing that the presence/absence of rare amino acids might dominate the clustering. Those problems might be fixed of course by restrictions on the length of the proteins, discarding rare amino acids, and, generally speaking, expert pre-and post-processing. This labor is reduced by maximizing Equation 7 and having a systematic criterion for the choice of the number of clusters (see the next section), even though the price to pay is a lengthier derivation. That was our reasoning in the choice of the clustering method and our motivation for favoring Equation 7 and the criterion presented in the next section for the number of clusters.

Choosing the number of clusters. The problem of how many clusters provide an appropriate description of the data is a classical issue in clustering [52,53]. A general perspective is given in [54] where the problem is reformulated in terms of an energy-versus-entropy competition. That elegantly demonstrates that the choice of the number of clusters is bound to depend on our level of description, condensed in [54] in the temperature of the system. The same fact is concretely indicated by Monte Carlo simulations by van Nimwegen et al. [55] for the clustering of transcription factor binding sites to predict regulons. When the space of possible configurations is sampled by Monte Carlo dynamics, clusters typically evaporate, drift, and fuse, and none of them lives forever, which makes a precise cluster membership identification quite problematic. A large variety of criteria for the choice of the number of clusters have been put forward in different problems [55–63]. In our case, since we shall be looking at functional categories of the genes composing the clusters, it is important to have a very reliable assignment of genes to clusters. We are therefore interested in imposing a criterion on the quality of the assignment and the stability of the clusters under reassignment. To this purpose, we shall employ a heuristic self-consistency criterion which has the advantage of being simple and rid of free parameters. A measure of the self-consistency in assigning gene g to cluster s is provided by the quantity: $b_g^s = \mathcal{L}(g \in C_s) / \mathcal{L}(g \in C_s) + \sum_{s' \neq s} \mathcal{L}(g \in C_{s'})$. $\mathcal{L}$ is the likelihood, defined in Equation 5, that the codon counts of gene g be generated with the probability distribution of cluster s. A value of $b_g^{(s)}$ close to unity implies that the gene matches uniquely well the usage of cluster s, and we can then be confident that its assignment is meaningful.

Let us then consider a configuration of S clusters, identified as described in previous sections. The quality of the corresponding assignments is quantified by the geometric average $B(S) = \prod_{s=1}^{S} \left(\sum_{g \in Cs} b_g^s / \sum_{g \in Cs} 1 \right)^{\frac{1}{S}}$. Taking the arithmetic mean inside each cluster ensures that this measure is not dominated by individual genes, while the geometric mean across clusters ensures that none of them has poor assignments if $B(S)$ is sufficiently close to unity. Rather than fixing an ad hoc threshold on $B(S)$, we have found it more effective to compare the stability of clusters obtained for real data to those in null models. Specifically, we calculate the posterior probability distribution of the real dataset for a unique cluster, comprising all genes. This single-cluster probability distribution is then used to generate an artificial dataset: each gene has the same length and number of amino acids as in the real genome, but amino acids are randomly encoded with the previous single-cluster distribution. This procedure guarantees that the overall statistics of codon usage is preserved, yet no cluster structure is by definition present in artificial data. Artificial data are then clustered as previously described and the average $Brandom(S)$ for these random data is computed over a sufficient number of realizations. The number of clusters retained is the one corresponding to the maximal difference $\Delta(S) = B(S) - Brandom(S)$, as shown in Figure 1 for *B. subtilis* and *E. coli*. Note that the assignment probabilities $B(S)$ for the number of clusters corresponding to the maxima in Figure 1 are 0.9 and 0.94, witnessing a strong consistency and statistical significance of the clusters identified. We experimented on various datasets generated with a prescribed distribution of codon usage and found that the method just described efficiently recovers the correct structure of the clusters and their distributions of codon usage.

Artificial chromosomes and null models. Given G genes and the numbers G_s ($s = 1, \ldots, S$) of genes in the S clusters, random chromosomes were generated as follows. Initially, one has G_1 cluster labels of the first type, $G2$ of the second type and so on ($\sum s\ Gs = G$), and a label is picked randomly and attached to any one of the G genes. One then iterates the procedure, randomly attaching the remaining labels to yet unlabeled genes. This ensures that all finite-size effects and the size of the clusters are correctly taken into account. Null statistics were obtained measuring the quantity of interest, e.g., COG distributions, over artificial chromosomes and accumulating statistics over an ensemble of 100,000 realizations. The resulting distributions are close to Gaussian by the central limit theorem. It was therefore appropriate to weight the significance of the deviations between real data and random cases by the corresponding z-scores, i.e., the deviation of the observed value to the mean of the random case, normalized by its standard deviation.

Data sources. We downloaded the complete annotated genomes from the NCBI microbial genome database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria). The list of genes used to gauge the Codon Adapatation Index (CAI) is downloaded from the EMGLib [17] Web site (http://pbil.univ-lyon1.fr/emglib/codon.html). The list of characterized transcripts for *E. coli* and *B. subtilis* is from [64, 65], while their metabolic pathways were taken from the KEGG: Kyoto Encyclopedia of Genes and Genomes database (http://www.genome.jp/kegg). The list of COG functional categories is discussed in [19] and is available at http://www.ncbi.nlm.nih.gov/COG.

Supporting Information

Figure S1. Figure Correlation for Randomized Chromosome *E. coli*

Solid line is the correlation function of cluster memberships as in Figure 6, for *E. coli*. Dashed line is the correlation function obtained for a randomized genome where the intra-operon contributions to P2(l) are retained but those stemming from different operons are depleted. Specifically, the randomization procedure is realised as follows. Labels are randomly permuted within the operons, yet keeping the fractions of genes fixed. For example, an operon with three genes belonging to the cluster α and two to β is randomized into one with three genes belonging to cluster γ and two to δ, with γ and δ randomly chosen. The genes composing the operon will then give the same contribution to P_2(l). However, since random permutations are independent among different operons, the inter-operon correlations will be depleted.

Found at DOI: 10.1371/journal.pcbi.0020037.sg001 (25 KB EPS).

Figure S2. Figure Correlation for Randomized Chromosome *B. subtilis*

The same curves as in Figure S1, for *B. subtilis*. Note that the correlation length is strongly reduced in randomized genomes, witnessing the fact that constraints imposed by operons are not sufficient to account for the extended correlations observed in Figure 6.

Found at DOI: 10.1371/journal.pcbi.0020037.sg002 (26 KB EPS).

Table S1. The Average Posterior Probabilities of Usage of the Synonymous Codons for the Four Clusters Identified in *E. coli* K12

Found at DOI: 10.1371/journal.pcbi.0020037.st001 (3 KB TEX).

Table S2. The Average Posterior Probabilities of Usage of the Synonymous Codons for the Five Clusters Identified in *B. subtilis*

Found at DOI: 10.1371/journal.pcbi.0020037.st002 (3 KB TEX).

Table S3. The Distribution of Genes among the Functional COG Classes for the Clusters Identified in *E. coli*

For each of the COG categories, the first line is the measured number of genes for that COG category, while the second line is the corresponding *z*-score (deviation to the number expected by chance, normalized by the standard deviation).

Found at DOI: 10.1371/journal.pcbi.0020037.st003 (2 KB TEX).

Table S4. The Distribution of Genes among the Functional COG Classes for the Clusters Identified in *B. subtilis*

For each of the COG categories, the first line is the measured number of genes for that COG category, while the second line is the corresponding *z*-score (deviation to the number expected by chance, normalized by the standard deviation).

Found at DOI: 10.1371/journal.pcbi.0020037.st004 (3 KB TEX).

Table S5. Repartition of the Genes Employed to Gauge the Codon Adaptation among the Clusters Identified in *B. subtilis*

Note the highly significant concentration in the first cluster. Genes used to gauge the CAI index for *E. coli* are all concentrated in the first cluster.

Found at DOI: 10.1371/journal.pcbi.0020037.st005 (28 KB PDF).

Acknowledgments

Useful discussions with E. D. Siggia are gratefully acknowledged.

Funding. The authors received no specific funding for this study.

Competing interests. The authors have declared that no competing interests exist. ■

References

1. Post L, Nomura M (1980) DNA sequences from the *str* operon of *Escherichia coli*. J Biol Chem 255: 4660–4666. Available: http://www.jbc.org/cgi/content/abstract/255/10/4660. Accessed 20 March 2006.
2. Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9: r43–r74.
3. Ikemura T (1981) Correlation between the abundance of *Escherichia coli* tRNAs and the occurrence of the respective codons in its protein genes. J Mol Biol 146: 1–21.
4. Ikemura T (1982) Correlation between the abundance of yeast tRNAs and the occurrence of the respective codons in protein genes: Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting tRNAs. J Mol Biol 158: 573–597.
5. Sharp P, Li W (1987) The codon adaptation index—A measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15: 1281–1295.
6. Sharp P, Tuohy T, Mosurski K (1986) Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res 14: 5125–5143.
7. Mazel D, Marlière P (1989) Adaptive eradication of methionine and cysteine from cyanobacterial light-harvesting proteins. Nature 341: 245–248.
8. Elf J, Nilsson D, Tenson T, Ehrenberg M (2003) Selective charging of tRNA isoacceptors explains patterns of codon usage. Science 300: 1718–1722.
9. Thanaraj TA, Argos P (1996) Protein secondary structural types are differentially coded on messenger RNA. Protein Sci 5: 1973–1983. Available: http://www.proteinscience.org/cgi/content/abstract/5/10/1973. Accessed 20 March 2006.
10. Thanaraj TA, Argos P (1996) Ribosome-mediated translational pause and protein domain organization. Protein Sci 5: 1594–1612. Available: http://www.proteinscience.org/cgi/content/abstract/5/8/1594. Accessed 20 March 2006.
11. Médigue C, Rouxel T, Vigier P, Hénaut A, Danchin A (1991) Evidence for horizontal gene transfer in *Escherichia coli* speciation. J Mol Biol 222: 851–856.
12. Moszer I, Rocha EP, Danchin A (1999) Codon usage and lateral gene transfer in *Bacillus subtilis*. Curr Opin Microbiol 2: 524–528.
13. Guerdoux-Jamet P, Hénaut A, Nitschke P, Danchin A (1997) Is the Escherichia coli outer membrane a patchwork of products from different genomes? DNA Res 4: 257–265. Available: http://dx.doi.org/10.1093/dnares/4.4.257. Accessed 20 March 2006.
14. Wang H, Badger J, Kearney P, Li M (2001) Analysis of codon usage patterns of bacterial genomes using the self-organizing map. Mol Biol Evol 18: 792–800.
15. Perrière G, Thioulouse J (2002) Use and misuse of correspondence analysis in codon usage studies. Nucleic Acids Res 30: 4548–4555. Available: http://nar.oxfordjournals.org/cgi/content/abstract/30/20/4548. Accessed 20 March 2006.
16. Pascal G, Médigue C, Danchin A (2005) Universal biases in protein composition of model prokaryotes. Proteins 60: 27–35.
17. Perrière G, Bessières P, Labedan B (2000) EMGlib: The enhanced microbial genomes library (update 2000). Nucleic Acids Res 28: 68–71. Available: http://nar.oxfordjournals.org/cgi/content/abstract/28/1/68. Accessed 20 March 2006.
18. Moller T, Franch T, Udesen C, Gerdes K, Valentin-Hansen P (2002) Spot 42 RNA mediates discoordinate expression of the *E. coli* galactose operon. Genes Dev 16: 1696–1706. Available: http://www.genesdev.org/cgi/content/abstract/16/13/1696. Accessed 20 March 2006.
19. Tatusov R, Fedorova N, Jackson J, Jacobs A, Kiryutin B, et al. (2003) The COG database: An updated version includes eukaryotes. BMC Bioinformatics 4: 41. Available: http://www.biomedcentral.com/1471–2105/4/41. Accessed 20 March 2006.
20. Rocha EPC, Danchin A (2003) Essentiality, not expressiveness, drives gene-strand bias in bacteria. Nat Genet 34: 377–378. Available: http://dx.doi.org/10.1038/ng1209. Accessed 20 March 2006.
21. Sueoka N (1995) Intrastrand parity rules of DNA base composition andusage biases of synonymous codons. J Mol Evol 40: 318–325.
22. Lobry JR (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 13: 660–665. Available: http://mbe.oxfordjournals.org/cgi/content/abstract/13/5/660. Accessed 20 March 2006.
23. Wu C, Maede N (1987) Inequality in mutation rates of the two strands of DNA. Nature 327: 169–170.
24. Dudkiewicz M, Mackiewicz P, Mackiewicz D, Kowalczuk M, Nowicka A, et al. (2005) Higher mutation rate helps to rescue genes from the elimination by selection. Biosystems 80: 193–199.
25. Peng C, Buldyrev S, Goldberger A, Havlin S, Sciortino F, et al. (1992) Long-range correlations in nucleotide sequences. Nature 356: 168–170.
26. Li W (1997) The study of correlation structures of DNA sequences: A critical review. Comput Chem 21: 257–272.
27. Bernaola-Galván P, Carpena P, Román-Roldán R, Oliver J (2002) Study of statistical correlations in DNA sequences. Gene 300: 105–115.
28. Audit B, Ouzounis C (2003) From genes to genomes: Universal scale-invariant properties of microbial chromosome organization. J Mol Biol 332: 617–633.
29. Lathe WC III, Snel B, Bork P (2000) Gene context conservation of a higher order than operons. Trends Biochem Sci 25: 474–479.
30. Rogozin IB, Makarova KS, Murvai J, Czabarka E, Wolf YI, et al. (2002) Connected gene neighborhoods in prokaryotic genomes. Nucleic Acids Res 30: 2212–2223. Available: http://nar.oxfordjournals.org/cgi/content/abstract/30/10/2212. Accessed 20 March 2006.
31. Boccard F, Esnault E, Valens M (2005) Spatial arrangement and macrodomain organization of bacterial chromosomes. Mol Microbiol 57: 9–16. Available: http://dx.doi.org/10.1111/j.1365–2958.2005.04651.x. Accessed 20 March 2006.
32. Hashimoto M, Ichimura T, Mizoguchi H, Tanaka K, Fujimitsu K, et al. (2005) Cell size and nucleoid organization of engineered *Escherichia coli* cells with a reduced genome. Mol Microbiol 55: 137–149.
33. Dong H, Nilsson L, Kurland CG (1996) Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. J Mol Biol 260: 649–663.
34. Halford SE, Marko JF (2004) . Halford SE, Marko JF (2004) How do site-specific DNA-binding proteins find their targets? Nucleic Acids Res 32: 3040–3052. Available: http://nar.oxfordjournals.org/cgi/content/abstract/32/10/3040. Accessed 20 March 2006.
35. Varenne S, Knibiehler M, Cavard D, Morlon J, Lazdunski C (1982) Variable rate of polypeptide chain elongation for colicins A, E2 and E3. J Mol Biol 159: 57–70.
36. Altuvia S, Weinstein-Fischer D, Zhang A, Postow L, Storz G (1997) A small, stable RNA induced by oxidative stress: Role as a pleiotropic regulator and antimutator. Cell 90: 43–53.
37. Moore SD, Sauer RT (2005) Ribosome rescue: tmRNA tagging activity and capacity in *Escherichia coli*. Mol Microbiol 58: 456–466.

38. Negrutskii B, Stapulionis R, Deutscher M (1994) Supramolecular organization of the mammalian translation system. Proc Nat Acad Sci U S A 91: 964–968. Avalable: http://www.pnas.org/cgi/content/abstract/91/3/964. Accessed 20 March 2006.
39. Brewer BJ (1988) When polymerases collide: Replication and the transcriptional organization of the *E. coli* chromosome. Cell 53: 679–686.
40. French S (1992) Consequences of replication fork movement through transcription units in vivo. Science 258: 1362–1365.
41. Mirkin EV, Mirkin SM (2005) Mechanisms of transcription–replication collisions in bacteria. Mol Cell Biol 25: 888–895. Available: http://mcb.asm.org/cgi/content/full/25/3/888. Accessed 20 March 2006.
42. Kunji R, Slotbooma DJ, Poolman B (2003) *Lactococcus lactis* as host for overproduction of functional membrane proteins. Biochim Biophys Acta 1610: 97–108.
43. Garrido T, Sanchez M, Palacios P, Aldea M, Vicente M (1993) Transcription of *ftsZ* oscillates during the cell cycle of *Escherichia coli*. EMBO J 12: 3957–3965.
44. Helmstetter CE, Thornton M, Grover NB (2001) Cell-cycle research with synchronous cultures: An evaluation. Biochimie 83: 83–89.
45. Bates D, Epstein J, Boye E, Fahrner K, Berg H, et al. (2005) The *Escherichia coli* baby cell column: A novel cell synchronization method provides new insight into the bacterial cell cycle. Mol Microbiol 57: 380–391.
46. Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis. Cambridge: Cambridge University Press. 356 p.
47. Jaynes E (1967) Prior probabilities. IEEE Trans Syst Sci Cybernet 4: 227–241.
48. Cover TM, Thomas JA (1991) Elements of Information Theory. New York: J. Wiley & Sons, Inc. 542 p.
49. Jeffrey A, Zwillinger D, editors (2000) Gradshteyn and Ryzhik's Table of integrals, series, and products. 6th edition. San Diego: Academic Press. 1163 p.
50. Lloyd S (1957) Least squares quantization in PCM. Murray Hill (New Jersey): Bell Telephone Laboratories.
51. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: LeCam L, Neyman J, editors. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press. Volume 1, pp. 281–297.
52. Bock HH (1996) Clustering and classification, chapter probability models and hypotheses testing in partitioning cluster analysis. Singapore: World Scientific. pp. 378–453.
53. Gordon AD (1999) Classification. London: Chapman and Hall/CRC Press. 256 p.
54. Rose K, Gurewitz E, Fox GC (1990) Statistical mechanics and phase transitions in clustering. Phys Rev Lett 65: 945.
55. van Nimwegen E, Zavolan M, Rajewsky N, Siggia ED (2002) Probabilistic clustering of sequences: Inferring new bacterial regulons by comparative genomics. Proc Nat Acad Sci U S A 99: 7323–7328. Available: http://www.pnas.org/cgi/content/abstract/99/11/7323. Accessed 20 March 2006.
56. Stone M (1974) Cross-validatory choice and assessment of statistical predictions. J R Stat Soc 36: 111.
57. Blatt M, Wiseman S, Domany E (1996) Superparamagnetic clustering of data. Phys Rev Lett 76: 3251–3254.
58. Balasubriaman V (1997) Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. Neural Comput 9: 349–368.
59. Smyth P (2000) Model selection for probabilistic clustering using cross-validated likelihood. Stat Comput 10: 63.
60. Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a dataset via the Gap statistic. J R Stat Soc B 63: 411.
61. Fraley C, Raftery A (2002) Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 97: 611.
62. Giada L, Marsili M (2002) Algorithms of maximum likelihood data clustering with applications. Physica A 315: 650–664.
63. Roth V, Lange T, Braun M, Buhmann J (2004) Stability-based validation of clustering solutions. Neural Comput 16: 1299–1323.
64. Itoh T, Takemoto K, Mori H, Gojobori T (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. Mol Biol Evol 16: 332–346. Available: http://mbe.oxfordjournals.org/cgi/content/abstract/16/3/332. Accessed 20 March 2006.
65. De Hoon M, Imoto S, Kobayashi K, Ogasawara N, Miyano S (2004) Predicting the operon structure of *Bacillus subtilis* using operon length, intergene distance, and gene expression information. Pac Symp Biocomput 9: 276–287.

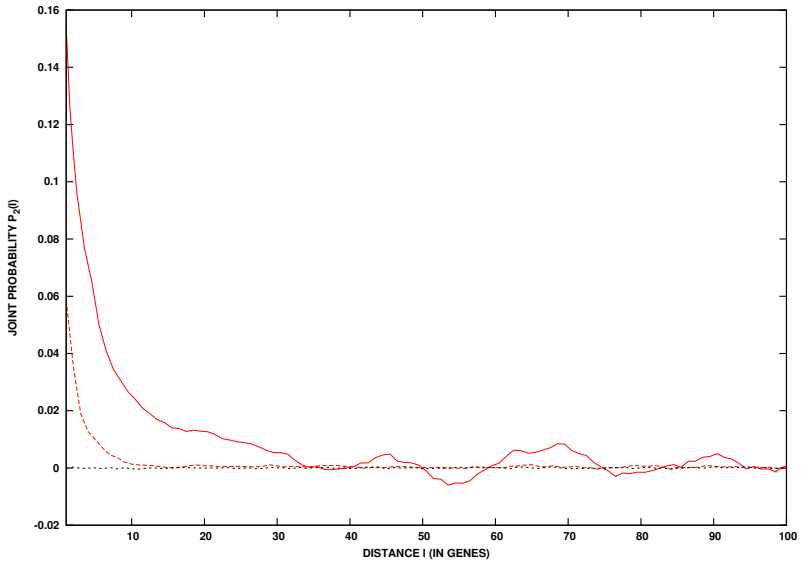


Figure S1: Solid, the correlation function of cluster memberships as in Fig. 5, for *E. coli*. Dashed, the correlation function obtained for a randomized genome where the intra-operon contributions to $C_2(l)$ are retained but those stemming from different operons are depleted. Specifically, the randomization procedure is realised as follows. Labels are randomly permuted within the operons, yet keeping the fractions of genes fixed. For example, an operon with 3 genes belonging to the cluster α and 2 to β is randomized into one with 3 genes belonging to cluster γ and 2 to δ, with γ and δ randomly chosen. The genes composing the operon will then give the same contribution to (7). However, since random permutations are independent among different operons, the inter-operon correlations will be depleted.

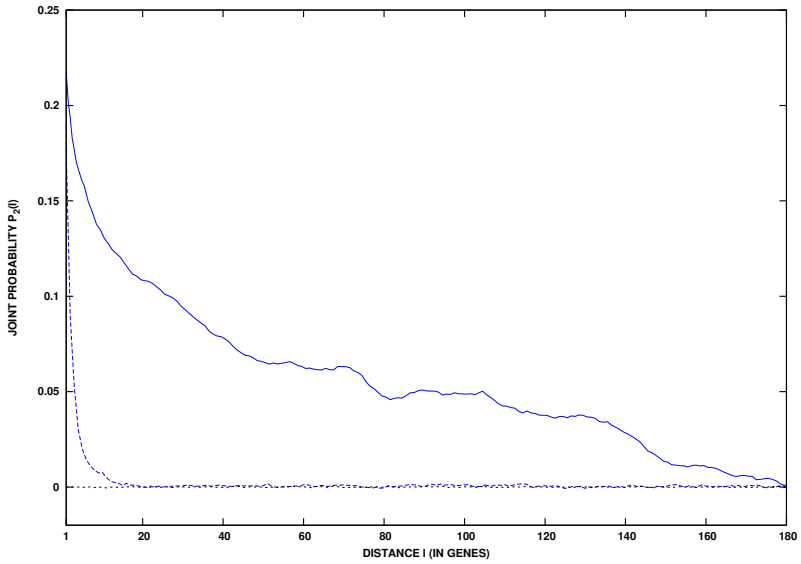


Figure S2: The same curves as in previous figure for *B. subtilis*. Note that the correlation length is strongly reduced in randomized genomes, witnessing the fact that constraints inposed by operons are not suficient to account for the extended correlations observed in Fig. 5.

Ala	1	2	3	4
GCA	.239	.308	.176	.210
GCC	.179	.226	.291	.287
GCG	.327	.228	.406	.358
GCT	.255	.237	.127	.145
Arg				
AGA	.006	.155	.012	.034
AGG	.003	.076	.007	.023
CGA	.009	.129	.033	.081
CGC	.347	.227	.468	.405
CGG	.005	.120	.067	.133
CGT	.629	.293	.414	.324
Asn				
AAC	.800	.357	.653	.484
AAT	.200	.643	.347	.515
Asp				
GAC	.514	.278	.405	.336
GAT	.486	.722	.595	.664
Cys				
TGC	.612	.443	.617	.544
TGT	.388	.557	.383	.456
Gln				
CAA	.197	.470	.294	.377
CAG	.803	.530	.706	.623
Glu				
GAA	.764	.690	.686	.669
GAG	.236	.310	.314	.331
Gly				
GGA	.026	.215	.070	.134
GGC	.428	.264	.456	.392
GGG	.048	.176	.144	.177
GGT	.498	.345	.330	.297
His				
CAC	.677	.317	.487	.376
CAT	.323	.683	.513	.624
Ile				
ATA	.006	.215	.026	.076
ATC	.628	.272	.472	.381
ATT	.366	.513	.502	.542

Leu	1	2	3	4
CTA	.041	.253	.083	.148
CTC	.058	.153	.110	.146
CTG	.010	.066	.025	.042
CTT	.082	.098	.110	.106
TTA	.749	.257	.590	.449
TTG	.060	.173	.082	.108
Lys				
AAA	.792	.741	.775	.757
AAG	.208	.259	.225	.243
Phe				
TTC	.680	.312	.495	.352
TTT	.320	.688	.505	.648
Pro				
CCA	.159	.280	.167	.192
CCC	.017	.172	.086	.162
CCG	.715	.265	.631	.477
CCT	.108	.284	.116	.168
Ser				
AGC	.268	.184	.305	.291
AGT	.055	.206	.113	.178
TCA	.061	.213	.096	.125
TCC	.244	.111	.170	.128
TCG	.075	.114	.184	.163
TCT	.296	.172	.131	.116
Thr				
ACA	.051	.276	.080	.143
ACC	.538	.271	.509	.407
ACG	.142	.233	.269	.308
ACT	.268	.220	.142	.143
Tyr				
TAC	.612	.304	.490	.386
TAT	.389	.696	.510	.614
Val				
GTA	.192	.207	.134	.147
GTC	.143	.215	.217	.236
GTG	.288	.217	.434	.379
GTT	.377	.361	.214	.237

Table S1: The average posterior probabilities of usage of the synonymous codons for the four clusters that we identified in *E. coli* K12.

Ala	1	2	3	4	5
GCA	.321	.380	.293	.236	.298
GCC	.080	.133	.195	.246	.207
GCG	.233	.111	.277	.317	.236
GCT	.366	.376	.235	.201	.259
Arg					
AGA	.118	.408	.263	.222	.281
AGG	.003	.136	.028	.108	.118
CGA	.024	.142	.077	.093	.116
CGC	.356	.075	.282	.228	.157
CGG	.007	.075	.101	.209	.171
CGT	.493	.164	.249	.140	.156
Asn					
AAC	.710	.313	.523	.451	.380
AAT	.290	.687	.477	.549	.620
Asp					
GAC	.446	.247	.398	.385	.337
GAT	.554	.753	.602	.615	.663
Cys					
TGC	.562	.346	.587	.616	.521
TGT	.438	.654	.413	.384	.479
Gln					
CAA	.706	.706	.516	.433	.518
CAG	.294	.294	.484	.567	.482
Glu					
GAA	.773	.716	.726	.629	.677
GAG	.227	.284	.274	.371	.323
Gly					
GGA	.293	.367	.309	.288	.335
GGC	.339	.173	.386	.388	.298
GGG	.043	.163	.115	.186	.186
GGT	.325	.298	.189	.137	.181
His					
CAC	.549	.251	.389	.334	.287
CAT	.451	.749	.611	.666	.713
Ile					
ATA	.016	.267	.056	.097	.169
ATC	.548	.204	.408	.413	.330
ATT	.436	.529	.536	.490	.500

Leu	1	2	3	4	5
TTA	.209	.377	.198	.142	.208
TTG	.107	.163	.133	.152	.172
CTA	.054	.112	.044	.037	.053
CTC	.060	.061	.114	.141	.103
CTG	.139	.082	.232	.303	.235
CTT	.431	.205	.277	.225	.229
Lys					
AAA	.832	.717	.749	.664	.684
AAG	.168	.283	.251	.336	.316
Phe					
TTC	.628	.245	.436	.305	.260
TTT	.372	.755	.564	.695	.740
Pro					
CCA	.273	.345	.193	.139	.198
CCC	.012	.097	.043	.105	.106
CCG	.307	.162	.465	.523	.401
CCT	.409	.395	.300	.233	.294
Ser					
AGC	.202	.132	.243	.257	.217
AGT	.078	.207	.083	.082	.112
TCA	.229	.263	.258	.213	.238
TCC	.075	.087	.129	.154	.120
TCG	.017	.069	.071	.127	.109
TCT	.399	.243	.215	.167	.204
Thr					
ACA	.502	.427	.455	.369	.410
ACC	.036	.122	.128	.196	.167
ACG	.151	.120	.264	.333	.263
ACT	.310	.331	.153	.101	.159
Tyr					
TAC	.604	.276	.411	.358	.303
TAT	.396	.724	.589	.642	.697
Val					
GTA	.303	.299	.211	.144	.202
GTC	.142	.124	.273	.318	.241
GTG	.142	.156	.236	.310	.266
GTT	.413	.420	.281	.228	.291

Table S2: The average posterior probabilities of usage of the synonymous codons for the five clusters that we identified in *B. subtilis*.

COG	1	2	3	4
-				
#	14	236	61	187
z-score	-5.2	16.7	-6.3	-5.5
B				
#	0	0	0	0
z-score	0	0	0	0
C				
#	24	4	45	55
z-score	4.4	-5.2	3.6	-1.2
D				
#	1	3	6	10
z-score	-0.6	-0.7	0.8	0.2
E				
#	14	4	64	64
z-score	0.6	-5.7	6.5	-1.1
F				
#	10	1	6	13
z-score	5.	-2.4	-0.3	-0.5
G				
#	13	6	29	51
z-score	1.8	-3.8	1.7	0.7
H				
#	1	1	28	29
z-score	-1.9	-3.7	4.7	0.2
I				
#	7	3	10	20
z-score	2.1	-2.1	0.4	0.2
J				
#	23	3	12	17
z-score	9.0	-2.9	-0.1	-2.6
K				
#	4	34	14	54
z-score	-1.8	2.9	-2.3	0.6

COG	1	2	3	4
L				
#	1	24	18	69
z-score	-2.9	0.1	-1.7	2.9
M	1	2	3	4
#	14	7	36	45
z-score	2.0	-3.6	3.2	-0.8
N				
#	1	14	7	29
z-score	-1.7	1.1	-1.5	1.3
O				
#	14	2	10	24
z-score	5.1	-3.	-0.4	-.02
P				
#	5	6	31	40
z-score	-0.8	-3.1	3.4	0.1
Q				
#	1	6	9	22
z-score	-1.3	-0.8	0.2	1.2
R				
#	3	25	18	71
z-score	-2.4	0.1	-1.9	2.8
S				
#	6	18	14	87
z-score	-1.5	-1.9	-3.1	5.
T				
#	1	12	8	34
z-score	-1.8	0.1	-1.4	2.1
U				
#	6	0	1	5
z-score	5.2	-1.8	-1.2	-0.5
V				
#	0	1	8	10
z-score	-1.3	-1.7	2.1	0.4

Table S3: The distribution of genes among the functional COG classes for the clusters identified in *E. coli*. For each of the COG categories, the first line is the measured number of genes for that COG category, while the second line is the corresponding z-score (deviation to the number expected by chance, normalized by the standard deviation).

COG	1	2	3	4	5
-					
#	20	340	146	221	429
z-score	-5.2	19.1	-6.0	-9.0	2.0
B					
#	0	0	0	1	0
z-score	-0.2	-0.4	-0.5	1.6	-0.7
C					
#	21	6	57	53	27
z-score	5.4	-3.7	5.5	0.9	-5
D					
#	0	2	9	8	14
z-score	-1.2	-1.2	1.3	-0.6	0.9
E					
#	8	4	76	101	88
z-score	-1.2	-6.	4.	2.7	-1.1
F					
#	4	4	36	13	20
z-score	0.4	-2.1	6.5	-2.4	-1.6
G					
#	8	20	47	120	79
z-score	-1.2	-3.0	-0.6	5.5	-2.1
H					
#	2	3	24	29	42
z-score	-1.2	-3.1	1.4	-0.1	1.6
I					
#	4	2	14	29	30
z-score	0.3	-2.9	-0.2	1.5	0.6
J					
#	61	6	44	20	20
z-score	22.2	-3.4	3.5	-4.4	-5.7
K					
#	7	34	30	99	113
z-score	-1.6	-0.6	-3.5	2.2	1.9

COG	1	2	3	4	5
L					
#	2	26	15	47	39
z-score	-1.6	2.3	-2.	1.8	-1.1
M					
#	4	15	31	66	57
z-score	-1.3	-1.8	-0.2	2.6	-0.5
N					
#	3	4	8	3	34
z-score	0.5	-1.2	-0.6	-3.8	4.7
O					
#	16	9	24	16	29
z-score	6.1	-1.1	1.8	-2.6	-0.8
P					
#	5	5	38	61	47
z-score	-0.7	-3.8	1.9	2.8	-1.2
Q					
#	0	6	16	29	28
z-score	-1.9	-1.5	0.4	1.5	0.1
R					
#	2	22	64	111	137
z-score	-3.5	-3.8	0.3	1.6	2.5
S					
#	6	27	52	87	120
z-score	-2.	-2.1	-0.3	0.2	2.4
T					
#	2	3	14	50	49
z-score	-1.4	-3.5	-1.9	3.2	1.6
U					
#	2	1	9	7	6
z-score	0.9	-1.4	2.3	-0.1	-1.1
V					
#	1	6	4	29	17
z-score	-1.	-0.6	-2.2	3.6	-0.8

Table S4: The distribution of genes among the functional COG classes for the clusters identified in *B. subtilis*. For each of the COG categories, the first line is the measured number of genes for that COG category, while the second line is the corresponding z-score (deviation to the number expected by chance, normalized by the standard deviation).

Gene	Group	Gene	Group	Gene	Group	Gene	Group
abrB	2	rpsE	2	yflH	4	pdhC	2
rplJ	2	rpmD	2	yfhD	4	ylaI	3
rplL	2	rplO	2	cspB	2	ylaJ	5
rplC	2	rplM	2	yhcN	1	divIVA	4
rplD	2	rpsI	2	prsA	2	rpmB	2
rplB	2	ybfQ	4	yhfD	2	fliJ	1
rplP	2	ycdA	4	cotW	4	cotE	5
rpsH	2	ycnE	2	ykwD	2	ynzH	1
rplF	2	ydbN	3	yknT	5	ynzC	4
rplR	2	ydcN	1	ykzG	1	yonK	1
ypzA	3	eno	2	rpsT	2	cotG	2
yokF	1	yttA	4	cspD	2	ytlB	3
ypjD	5	yvcE	2	fer	2	yvzB	2
ysnF	5	ywhB	4	rpmI	2	yxeE	4
yscB	5	ahpC	2	rpsD	2		

Table S5: Repartition of the genes employed to gauge the Codon Adaptation Index [5], as in [17], among the clusters identified in *B. subtilis*. Note the highly significant concentration in the second cluster. Genes used to gauge the CAI index for *E. coli* are all concentrated in the first cluster.

Addendum to the paper "Codon usage domains over bacterial chromosomes" PLoS Computational Biology Vol. 2, No. 4, e37

Marc Bailly-Bechet, Antoine Danchin, Mudassar Iqbal
Matteo Marsili, Massimo Vergassola.

April 20, 2007

An issue left unexplained in the paper [1] is the striking quantitative difference between *E. coli* and *B. subtilis*. This is clearly visible in Fig. 6 of [1], where it is shown the probability that two genes at distance ℓ belong to the same cluster of codon usage. Clusters are characterized by a similar codon bias and were identified using a novel information-based clustering method. While both curves decay on distances sizably longer than what could be accounted by operons, *B. subtilis* is manifestly correlated on much longer distances. It is hard to develop a biologically well-founded explanation for such a striking difference between the two organisms. This observation and discussions with Dr. Morten Kloster (Princeton Univ.) spurred us to reconsider the issue and further pursue our analysis of the clusters. The purpose of this addendum is to describe this analysis, which allows to point out an incorrect statement made in the paper and provide an explanation for the aforementioned difference between the two organisms.

Contrary to what was previously stated, the GC content of the various clusters is not quite homogeneous and the correct values are reported in Table 1.

Cluster	1	2	3	4
GC%	.527	.443	.541	.522

Cluster	1	2	3	4	5
GC%	.439	.358	.450	.470	.436

Table A1: The GC percentage for the four and the five clusters of *E. coli* and *B. subtilis*, which were identified on the basis of their codon usage.

The demonstration given in [1] that clusters are biologically significant still holds and does not depend on the GC content. In particular, the third cluster of *B. subtilis*, which was shown to feature an over-representation of anabolic genes and lagging-strand

transcriptional orientation, does not show any particular GC content. The clusters which most significantly deviate from the average are the second ones, enriched in AT. Both clusters have been shown in [1] to be enriched in horizontally transferred genes. The higher AT% shown in Table 1 is in agreement with the well-known observation that horizontally transferred genes tend to be AT rich (see [2] and references therein).

The GC percentage resolves the aforementioned observation of the different correlation lengths in Fig. 6 of [1] for *E. coli* and *B. subtilis*. To demonstrate this, we considered the same correlation functions plotted in Fig. 6, but for each individual cluster, to highlight the contribution of the various groups. Specifically, we measured the probability that two genes, g and $g+\ell$, belong to the same cluster ($s_g = s_{g+\ell}$), with the additional constraint that $s_g = S$ ($S = 1, \ldots, 4$ for *E. coli* and $S = 1, \ldots, 5$ for *B. subtilis*):

$$\mathcal{P}_2^{(S)}(\ell) = \frac{1}{N_S} \sum_g \delta(s_g\,, s_{g+\ell}) \delta(s_g, S)\,. \tag{1}$$

Here, δ is the Kronecker delta-function and the function $P_2^{(S)}$ is normalized by the total number of genes N_S belonging to the S-th cluster. The function can also be interpreted as the histogram of the distances among genes belonging the same cluster. The resulting curves for the various clusters are shown in Fig. A1 for *E. coli* and Fig. A2 for *B. subtilis*, with the value at large distances subtracted for more clarity.

A first observation is that the curves are more noisy than in Fig. 6 of [1]. This is quite natural as each group contains less genes and was our reason for grouping all the clusters together to produce Fig. 6. Some statistically robust informative behaviors are still clearly discernible, though. In particular, it is quite evident that the cluster of *B. subtilis* having the longest correlations is the fourth one. The correlation length of the cluster is clearly dominant over all the others and is comparable to the decay length observed in Fig. 6 of [1]. Since the fourth cluster is GC enriched it is quite sensible to ascribe the dominant contribution to its anomalous decay length in Fig. 2 to the strong correlations in the GC content present over the genome of *B. subtilis*. It is important, though, to remark that groups not biased in their GC content also feature extended correlations, longer than what could be accounted by operons, and the effects are now comparable in *E. coli* and *B. subtilis*. A contribution to those correlations might be driven by the advantage of recycling rare tRNAs to tame stallings in the translation process and ensure a coordinated expression of a set of neighboring genes, as discussed in the conclusions of [1]. The importance of pauses is translation is also highlighted by the large number of tmRNAs typically present in the cell [3, 4].

References

[1] Bailly-Bechet M, Danchin A, Iqbal M, Marsili M, Vergassola M (2006) Codon usage domains over bacterial chromosomes. PLoS Computational Biology, Vol. 2, No. 4, e37.

[2] Rocha EPC, Danchin A (2002) Base composition bias might result from competition for metabolic resources. Trends in Genetics 18(6):291–294.

[3] Altuvia S, Weinstein-Fischer D, Zhang A, Postow L, Storz G (1997) A small, stable RNA induced by oxidative stress: role as a pleiotropic regulator and antimutator. Cell 90:43–53.

[4] Moore S D, Sauer R T (2005) Ribosome rescue: tmRNA tagging activity and capacity in *Escherichia coli*. Molecular Microbiology 58(2):456–466.

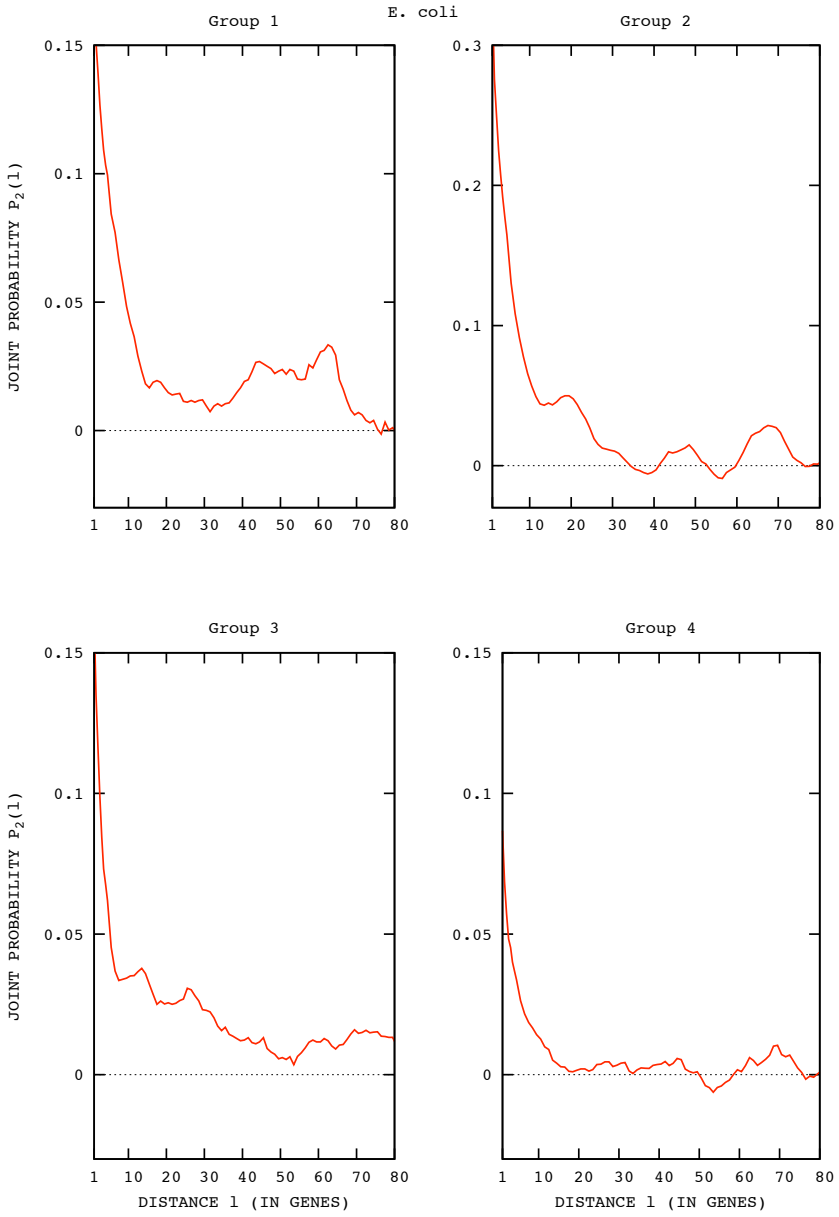


Figure 1: The probability distribution (1) of cluster membership for the four clusters identified in *E. coli*.

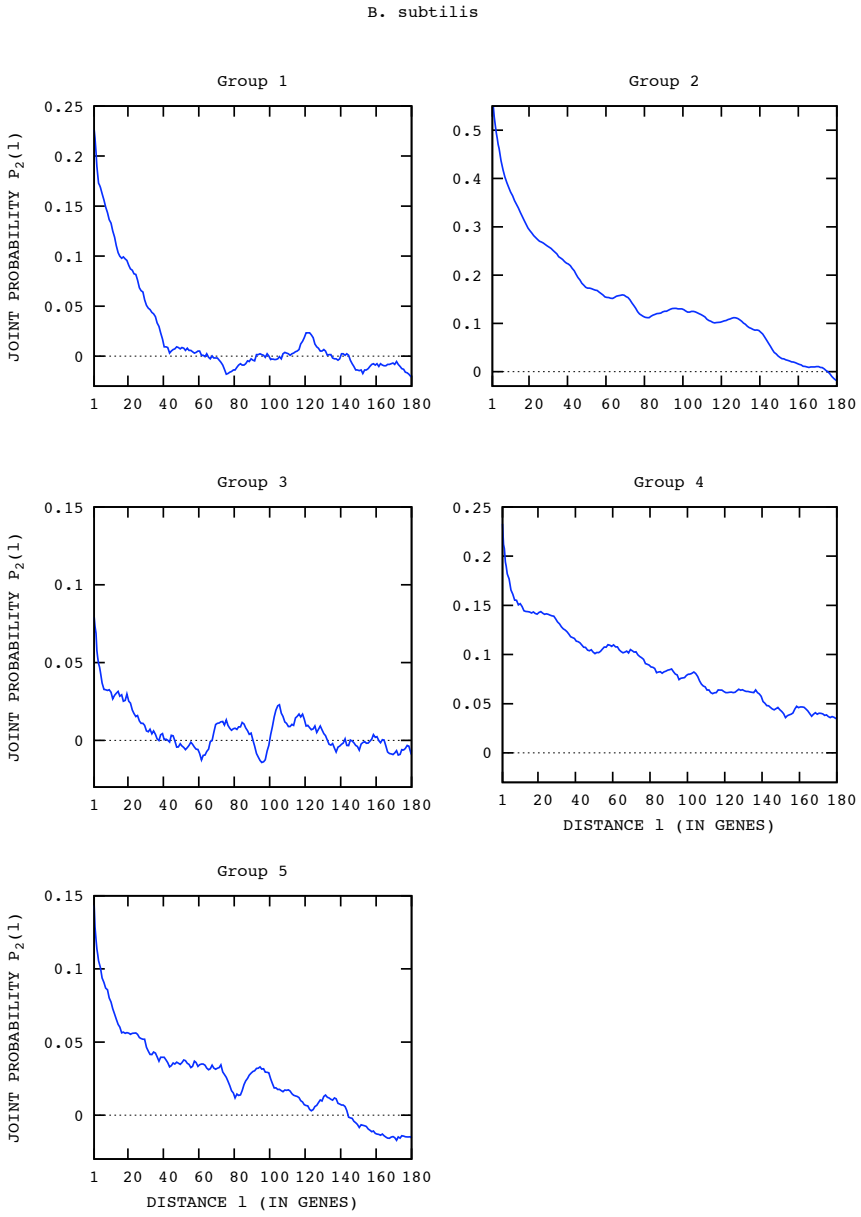


Figure 2: The probability distribution (1) of cluster membership for the five clusters identified in *B. subtilis*.

6.4 Perspectives

Les possibilités d'étude offertes à partir de ce point sont très vastes. En effet, les résultats biologiques que l'on peut inférer à partir de l'étude précise du biais d'usage de codons sont nombreux, comme on l'a remarqué plus haut. Plusieurs thèmes de recherche peuvent être envisagés en reprenant les méthodes que nous avons développées :

- Le premier qui vient à l'esprit est l'étude comparative de génomes bactériens. Cette étude peut être fait à plusieurs échelles : sur des organismes proches, elle peut être comparée aux homologies connues et déceler des différences sur l'emploi de certains codons ou acides aminés, qui autoriseraient à faire des hypothèse sur les causes environnementales ou sélectives qui façonnent les génomes. À grande échelle, la comparaison des classifications obtenues pour des génomes très différents permettrait de mettre en avant des caractéristiques globales communes.
- Notre algorithme pourrait être appliqué sur des génomes eucaryotes. Le problème du biais d'usage de codons est moins clair chez ces organismes, et l'emploi de méthodes de classification permettrait peut-être de mettre à jour une organisation inconnue des génomes eucaryotes.
- Une autre thématique qui est envisagée est d'utiliser la méthode de classification pour aider à l'annotation de génomes. Les résultats de la classification peuvent en effet être utilisés pour inférer la fonction d'un gène, ou son niveau d'expression. De plus, le format sous lequel ils apparaissent permet une automatisation facile du traitement des données.
- Notre analyse peut être restreinte sans difficultés à un groupe de gènes ou d'acides aminés d'intérêt particulier, autorisant son emploi dans des problèmes autres que l'analyse des génomes entiers.
- Un autre axe d'expansion de notre travail est la modélisation spatiale des processus de traduction de gènes proches, ayant ou non le même biais d'usage de codons. Ceci permettrait d'explorer en détails les effets de recyclage des ARNt, qui permettrait de réguler la traduction chez les bactéries en profitant du fait que les gènes proches sur le chromosome sont traduits par des ribosomes proches. Cette analyse peut s'envisager aussi bien analytiquement que numériquement.
- Finalement, un champ d'application de notre méthode est une discipline en pleine expansion, la métagénomique. Les quantités de données génétiques obtenues lors de l'analyse d'échantillons naturels d'eau ou de sol contiennent un grand nombre de séquences non homologues à celles des génomes connus, et forment un terrain de choix pour l'utilisation de la classification et les prédictions qu'elle permet d'accomplir.

Chapitre 7

Recrutement d'ARN de transfert par les bactériophages

Nous changeons ensuite de sujet et nous intéressons aux virus des bactéries, les bactériophages, du point de vue du biais d'usage de codons et son lien avec le système de traduction. En effet, une observation intrigante est que certains phages contiennent des ARNt, à l'exception de toute autre gène codant pour des composants du système de traduction. Nous nous sommes donc intéressés à ce problème, et avons mis en relation le contenu en ARNt des phages de leurs hôtes avec leurs biais d'usage de codons respectifs. Cette étude a permis de mieux comprendre les mécanismes évolutifs des phages, dont l'étude est en grande expansion, à cause des possibilités de thérapie médicale qu'ils offrent et de leur rôle dans les processus de transfert horizontal.

7.1 Historique

Il est connu depuis longtemps que certains phages ont un usage biaisé de codons, et possèdent des gènes codant pour des ARNt. Si l'existence d'un biais de codons peut sembler naturelle chez les phages quand on sait l'importance de maximiser la vitesse de la traduction dans leur cycle de vie, la présence d'ARNt est plus difficile à expliquer. En effet, les phages ne codent habituellement pour aucune composante du système de traduction, et doivent entièrement subvertir celui de leur hôte pour synthétiser les protéines nécessaires à la génération suivante. De plus la présence de gènes "inutiles" chez les phages est certainement contre-sélectionnée, car elle rallonge leur temps de réplication.

Le biais d'usage de codons des phages a été étudié et mis en relation avec celui de ses hôtes. En effet, si le biais de codon de l'hôte correspond à une adaptation de l'usage du code pour optimiser une ou plusieurs étapes de la traduction, il est naturel de penser que les pressions sélectives qui s'appliquent sur les gènes de l'hôte à ce niveau s'appliquent également sur ses phages, puisque leurs gènes sont traduits par la machinerie cellulaire de l'hôte. Dans une hypothèse où un ajustement du biais de codons permet de maximiser la vitesse de la traduction, on voit facilement qu'un phage ayant le même biais de codons que son hôte sera traduit plus vite qu'un phage n'étant pas biaisé de la même manière. Si le taux de réplication n'est pas limitant, ceci va entraîner une reproduction plus rapide du phage biaisé dans le même "sens" que son hôte, lui donnant un clair avantage sélectif.

Le biais d'usage de codons de nombreux phages a été étudié. Il a été montré que

le phage T7 avait un usage de codons corrélé à l'abondance des ARNt de son hôte, en particulier pour ses gènes fortement exprimés (Sharp et al., 1985). Ce résultats a été généralisé (Kunisawa et al., 1998) à de nombreux phages d'*E. coli*, mais il a été également observé qu'il ne semblait s'appliquer que pour les phages ne codant pas pour leur propre ADN polymérase, suggérant que les biais mutationnels durant la réplication pouvaient jouer un rôle important sur le biais d'usage de codons. Cette hypothèse a été renforcée par l'analyse récente du biais de codon de Mimivirus, un virus des eucaryotes (Sau et al., 2006). La transcription a également été évoquée comme un facteur modelant l'usage de codons des phages (Kano-Sueoka et al., 1999). Sur d'autres espèces, les résultats sont également mitigés : les phages de *Staphylococcus aureus* ont un biais d'usage de codons significatif, mais qui semble varier d'une espèce à l'autre (Sau et al., 2005), ce qui ne peut s'expliquer qu'au prix d'hypothèses supplémentaires dans une théorie basée sur des causes sélectives pour la traduction.

La présence d'ARNt dans les phages a également fait l'objet de nombreuses études, mais leurs résultats n'ont pas permis de proposer une théorie unificatrice pour expliquer leur existence. Les travaux sur le phage T2 ont montré que le contenu cellulaire d'*E. coli* en ARNt changeait après infection par T2, impliquant un rôle du phage dans la modification du système de traduction de son hôte (Kan et al., 1968; Kano-Sueoka and Sueoka, 1969; Sueoka and Kano-Sueoka, 1964). Chez le phage T4, la présence de 8 ARNt a tout d'abord été expliquée comme corrélée au biais de codons de ses gènes tardifs (Cowe and Sharp, 1991), puis au contraire comme compensant les déficiences de son hôte en terme d'ARNt lors de l'expression des gènes faiblement exprimés du phage (Kunisawa, 1992). Finalement leur rôle fut considéré comme "incertain" (Miller et al., 2003). Il fut ensuite montré que les ARNt du phage D29 de *M. tuberculosis* correspondaient aux codons majeurs du phage plutôt que de son hôte. Une autre fonction potentielle des ARNt dans les génomes de phage a également été supposée, en plus de pouvoir aider la traduction des protéines phagiques : il s'agirait de faciliter l'insertion des phages tempérés dans les génomes de leurs hôtes (Kropinski and Sibbald, 1999), grâce aux séquences palindromiques typiques des ARNt.

On voit que les résultats de ses études, s'ils ont souvent mis en avant un lien entre système de traduction et biais d'usage de codons des phages, n'ont pas permis d'identifier clairement leur rôle chez les phages. C'est à cette tâche que nous nous sommes attelés, en nous basant sur le grand nombre de génomes de phages séquencés au cours des dernières années, pour déchiffrer l'usage du code et des ARNt chez les bactériophages.

7.2 L'article

Nous présentons ici notre article, actuellement en cours de relecture avant publication dans la revue *Genome Research*. Les matériels supplémentaires sont ajoutés à la fin. Les principales conclusions sont :

- Le biais d'usage de codons des phages est corrélé à celui de leur hôte. Une différence significative est observée entre les phages tempérés (très corrélés) et les phages lytiques (moins corrélés).
- Les phages lytiques contiennent plus d'ARNt. Ceci nous permet d'argumenter en faveur d'une pression de sélection pour garder les ARNt basée sur la compensation des différences d'usage du code entre le phage son hôte.

– La modélisation de cette hypothèse, par une méthode d'équation maîtresse supposant un recrutement aléatoire des ARNt des phages chez leurs hôtes, et une perte sélective en fonction de la différence d'usage de codons entre eux, la confirme. Ceci est vrai aussi bien pour les phages lytiques que pour les phages tempérés, et est plus vraisemblable au vu des données qu'une sélection basée uniquement sur le biais de codons du phage ou de l'hôte.

Causes for the intriguing presence of tRNAs in phages

Marc Bailly-Bechet[1,4], Massimo Vergassola[1], Eduardo Rocha[2,3]

[1] CNRS URA 2171, Institut Pasteur, Unité Génétique in silico,
25 rue du Dr. Roux, F-75724 Paris Cedex 15, France

[2] Atelier de Bioinformatique, Université Pierre et Marie Curie-Paris 6,
12 rue Cuvier, 75005 Paris, France

[3] CNRS URA 2171, Institut Pasteur, Unité Génétique des Génomes Bactériens,
28 rue du Dr. Roux, F-75724 Paris Cedex 15, France

[4] Corresponding author.
E-mail mbailly@pasteur.fr; fax 33-1-40613927

Running title: Causes for the presence of tRNAs in phages
Keywords: bacteriophage, tRNA, codon usage

Abstract

Phages have highly compact genomes with sizes reflecting their capacity to exploit the host resources. Here, we investigate the reasons for tRNAs being the only translation-associated genes frequently found in phages. We were able to unravel the selective processes shaping the tRNA distribution in phages by analyzing their genomes and those of their hosts. We found ample evidence against tRNAs being selected to facilitate phage integration in the prokaryotic chromosomes. Conversely, there is a significant association between tRNA distribution and codon usage. We support this observation by introducing a master equation model, where tRNAs are randomly gained from their hosts and then lost either neutrally or according to a set of different selection mechanisms. Those tRNAs present in phages tend to correspond to codons which are simultaneously highly used by the phage genes while rare in the host genome. Accordingly, we propose that a selective recruitment of tRNAs compensates for the compositional differences between the phage and the host genomes. To further understand the importance of these results in phage biology, we analyzed the differences between temperate and virulent phages. Virulent phages contain more tRNAs than temperate ones, higher codon usage biases and more important compositional differences with respect to the host genome. These differences are thus in perfect agreement with the results of our master equation model and further suggest that tRNA acquisition may contribute to higher virulence. Thus, even though phages use most of the cell's translation machinery, they can complement it with their own genetic information to attain higher fitness. These results suggest that similar selection pressures may act upon other cellular essential genes that are being found in the recently uncovered large viruses.

A table containing accession numbers of all genomes analyzed is given in supplementary material.

Introduction

Parasites face numerous problems when colonizing their hosts. First among these is the optimisation of host exploitation, which is a particularly important problem for obligatory lethal pathogens because they have to get by with the available host resources. Host exploitation is often difficult to study because it is linked with complex traits. This is one of the reasons why the antagonistic associations between bacteria and their phages have recently resurged as interesting models to understand host-pathogen interactions and resulting life history-traits (De Paepe and Taddei, 2006, Turner and Chao, 1999). Phages are also important shuttles of horizontal gene transfer and thus major elements in the dynamics of bacterial evolution (Canchaya et al., 2004, Casjens, 2003, Daubin and Ochman, 2004). Among the genes carried by phages, toxins are particularly important for bacterial pathogenicity (Waldor and Mekalanos, 1996). Thus, phages and bacteria can transiently establish mutualistic interactions to antagonize eukaryotic hosts. Since phages are the most abundant life form on earth (Suttle, 2005), the consequences of their ecological interactions are most relevant for both the global ecosystem and human health.

The genomes of phages are typically small, providing them with few tools to divert resources from their environment. Accordingly, they have no proper metabolism and rely on the host cell's materials for their reproduction (Weinbauer, 2004). Phages also rely on the host machinery to reproduce and while some code for their own RNA and DNA polymerases (Knopf, 1998), they require most of the cell's translation apparatus for protein synthesis. Accordingly, the biases operating in the host sequences towards translation optimization may also operate in the phage genome. It has even been supposed (Krakauer and Jansen, 2002) that these biases could be a major force of both phage and bacterial genomes co-evolution.

Selection for optimal codon usage plays a major role in shaping bacterial genomes (Andersson and Kurland, 1990), especially in fast-growing bacteria (Rocha, 2004, Sharp et al., 2005) and among highly expressed genes (Grantham et al., 1981). In cellular organisms, the optimal codon usage is typically the one fitting best the abundance of tRNAs in the cell under exponential growth conditions (Ikemura, 1981). Thus, there is a co-evolution of tRNA abundance and codon usage bias that shapes the abundance of the different codons in gene sequences through long periods of time. This trait is under selection but is counteracted by the action of random mutations

that tend to make codon usage bias a reflect of the extant mutational biases (Muto and Osawa, 1987). Codon usage bias is thus said to be under selection-mutation-drift balance (Bulmer, 1991), and the result is that the bias is more intense i) in the genes under stronger selection, which are often, but not always (Bailly-Bechet et al., 2006, Elf et al., 2003), the most highly expressed (Gouy and Gautier, 1982) and ii) in those moments when there is more selection for the trait, e.g. exponential growth for most highly expressed genes (Dong et al., 1996). When a phage propagates in a bacterium cell, it is convenient to have a codon usage bias compatible with the one of the bacteria as that will facilitate the expensive and laborious mechanism of protein synthesis. Yet, this may be impossible to achieve because phages share with other parasitic DNA a tendency to be A+T rich relative to their bacterial hosts (Rocha and Danchin, 2002). The gap between phage and host genome compositions makes it difficult for the phage to perfectly fit the host codon usage.

Nearly 40 years ago, it was found that T4 phages carry some tRNA genes (Weiss et al., 1968). Although deletion of these genes leads to lower burst sizes and rates of protein synthesis (Wilson, 1973), the reasons why some phages contain tRNAs have remained enigmatic. Early work, also on T4, suggested that its tRNA gene content corresponded to the codon usage of some lowly expressed genes in the phage, for which the corresponding host tRNAs were too rare (Cowe and Sharp, 1991, Kunisawa, 1992). These studies also found that, while highly expressed genes in T4 tend to have a codon usage bias close to the host, *E. coli*, lowly expressed and late genes use codons for which some of the 8 T4 tRNAs could be useful. Yet, lack of data did not allow at that time to understand if the observation was a peculiarity of T4 and *E. coli* or a general feature. The finding that phages closely related to T4 showed extensive polymorphisms in the number and type of tRNA genes contributed to the near abandon of the work on this hypothesis (Miller et al., 2003). Meanwhile, the availability of hundreds of bacterial genomes highlighted the role of phages in bacterial evolution as vectors of horizontal gene transfer. About half of the sequenced genomes contain prophage sequences (Canchaya et al., 2004) and these may constitute up to 16% of the genome (Ohnishi et al., 1999). Importantly, prophage integration occurs at a tRNA gene for phages carrying lambda and P4-like integrases (Campbell, 1992). These phages do not carry tRNAs, but only a small part of a tRNA that compensates for the disruption in the host tRNA. In this context it has been proposed that tRNA presence in phages could be a by-product of

imprecise excision of prophages. This would not strictly require a positive effect of tRNAs on phage fitness, although one might suppose that their presence could be selected to compensate for insertions inactivating the host's tRNA (Canchaya et al., 2004). A problem with this hypothesis is that it fails to explain the presence of tRNAs in non-temperate phages. A possible explanation would be the capture of bacterial DNA by the phage during host chromosome degradation, before encapsulation of all the phage genetic material, as this process liberates large quantities of DNA (Weinbauer, 2004). Virulent phages could also acquire tRNAs by recombination with temperate phages. As genomic data showed that tRNAs provide integration points for phages, plasmids and pathogenicity islands, other putative roles for phage-encoded tRNAs have been neglected. Yet, this issue gains a special relevance at the light of the role of phages in bacterial evolution and given the recent discovery of large eukaryotic viruses containing many other elements of the translation machinery, such as elongation factors and tRNA synthetases (Raoult et al., 2004).

We have thus decided to investigate the relationship between tRNA copy number and codon frequency in bacteriophage genomes relative to their hosts. This is now possible because hundreds of phage and bacterial genomes have been sequenced. The study of codon usage in bacteriophage genomes presents additional constraints relative to the equivalent study in bacteria. Firstly, the process of selection of tRNAs must take into account that tRNAs have probably been taken from the host genome. Secondly, the codon bias of phages is modified by a general compositional bias towards higher A+T content than the host genome, which could blur the simple picture that arises in bacteria for the relationship between tRNA content and codon usage. Finally, the low number of tRNAs present in phage genomes implies the usage of careful statistics and sophisticated models. We have identified tRNAs in phage genomes and in their hosts and investigated the correlation between the tRNAs of the phage and its codon bias. Then, we developed a master equation model to simulate the acquisition and loss of tRNAs by phages, and, using likelihood comparisons, we found the most relevant selective processes that could drive it. We finally placed these results in the framework of phage ecology and evolution.

Results

Associations between tRNAs and codon usage in phages and their hosts

We collected from GenBank the complete genomes of phages and their hosts. We then identified tRNAs in both groups of genomes removing some elements which could complicate our analysis, such as pseudogenes (see Methods). We eliminated from further analyses all phages which were not annotated, which lacked tRNAs and those without a completely sequenced bacterial host. The final data set contains 15 host bacterial genomes and 37 genomes of phages. The genomes of these phages contain a total of 169 tRNAs, thus showing an average of ~ 4 tRNAs per phage. However, such an average value is somewhat misleading since the number of tRNAs inside each phage is very variable (Fig. 1). Most phages contain only one or two tRNAs, while a few contain more than 20 such sequences, which is nearly as many tRNAs as one can find in bacteria with minimal genomes (Rocha, 2004). All tRNA-containing phages are dsDNA phages. The main difference between the phages with tRNAs and those without any tRNAs stands in genome length: phages containing tRNAs are significantly longer than those without (average lengths are respectively 74 kb and 32 kb, $p = 10^{-6}$). We then separated phages into virulent and temperate, according to the published information on their ecology. In some cases both annotation files and published literature lacked information allowing such a classification and those phages were thus gathered in a third group. We could classify 21 phages as temperate and 12 as virulent. The abundance of tRNAs in their genomes is very different (Fig. 1), with no temperate phage containing more than 4 tRNAs. We shall get back to this issue in a subsequent section as it is relevant to understanding the role of tRNAs in phages.

We started our analyses by testing the hypothesis that the tRNA gene content for each anticodon is positively correlated with the complementary codon frequency in the genome of phages. This would be a situation similar to the one of bacteria, where tRNA gene content is highly correlated both with tRNA cellular content (Dong et al., 1996, Ikemura, 1985) and codon frequency (Ikemura, 1981). The test performed in each phage independently shows weak statistical power due to the few tRNAs they contain. Hence, we took all the 37 phages genomes into account and computed the probability distribution of their codon frequencies, i.e. the probability $p(f)$ that any

codon inside a phage will be used at a given frequency f. One can see that this distribution does not differ between the phages and their hosts (Fig. 2). This means that on average few codons are highly used, while most are rare and that the trends are similar in phages and their hosts. Then, we computed the same probability distribution restricted to the codons for which at least 1 cognate tRNA is found in the bacteriophage genome. The curve is rougher, because of the lower number of points, but it is clearly different from the previous one (Kolmogorov-Smirnov test, $p < 10^{-3}$). The peak around f=0.03 is due to the presence of numerous tRNA^{Met} inside phage genomes, in which methionine appears to be often used around this frequency. In the inset of figure 2, we plot the cumulated probability distribution of the codon frequency, i.e. the probability of finding codons with frequencies superior to the value given on the x-axis, for phage codons with and without the complementary tRNA. The lag between the two curves shows that there is a high proportion of codons having a matching tRNA among the high frequency codons. It is thus clear that there is a positive association between the frequency of a codon and the presence of the cognate tRNA in the phage genome.

We then computed the correlations of codon frequencies between each phage and its host. As previously noted for some of these phages (Kropinski and Sibbald, 1999, Kunisawa et al., 1998, Sau et al., 2005, Sharp et al., 1985), we find highly correlated codon frequencies: the average of the Pearson's coefficient R on all 37 couples phage-host is 0.78±0.04 (standard error), and 36 out of the 37 associated p-values are inferior to 0.05. This value has to be compared to the value found by computing the average correlation coefficient, 0.38±0.02, between a phage and a random host (computed between the phage and a random bacterial genome from the 356 we found available on GenBank). We conclude that the frequencies of codon usage are very correlated between a phage and its host. A similar pattern of correlations is observed among phages containing no tRNAs.

If phages and their hosts had exactly the same codon usage then a trivial explanation of the results of the previous paragraph would be that phages pick tRNAs randomly from their hosts and that the correlation between phage codon usage and tRNA gene content simply reflects the association between tRNA abundance and codon usage in the host. No hypothesis about selection on tRNAs in phages would then be required. Yet, a coefficient of correlation of 0.78 only allows explaining about half of the variance, leaving ample room for an autonomous selection strategy of tRNA acquisition in

phages. We shall show that the no-selection hypothesis does not fit the data as adequately as some models featuring selection.

tRNAs may be randomly recruited from the hosts but they are selectively kept

One can explain the previous results in at least two different ways. A purely neutral hypothesis is that tRNAs are drawn at random from the host genome. A selective refinement of this hypothesis is that tRNAs are drawn at random and kept because they help phage integration. In this case tRNAs should be kept in the phage at the same frequency as in the hosts, and their distribution would differ from a random uptake with no selection only by a greater magnitude of the rate of tRNA acquisition.

A second alternative is that, after tRNA recruitment, there is selection for keeping some tRNAs but not others. One would expect this to be correlated with the phage codon usage or the difference in codon usage between the phage and the host. In the first case, the tRNAs would render the phage less dependent on the host to translate its own proteins. The second case implies an evolutionary strategy of compensation of codon usage differences, where the phage keeps those tRNAs which are rare in the host and whose cognate codons are frequent in the phage genome.

We first assessed whether it is reasonable to assume that tRNAs are randomly picked from bacterial hosts. When analyzing the anticodons of tRNA genes we only observed 8 tRNAs, out of 177, that were present in the phage and not in its host. In 3 cases the Cove score of these tRNAs, as given by tRNAscan_SE, is less than 30, which is at the borderline of significance and suggests that they are false positives. For the 5 remaining tRNAs, we checked if some putative other host of the same phage had the corresponding tRNAs (i.e. a tRNA with the same anticodon). As discussed in Methods we only used in these comparisons one randomly chosen host for the phage, even when we knew several (e.g. in multiple sequenced strains of a species). In all the 5 cases cases we did find another host genome containing the phage tRNAs. Thus, all reliable tRNAs we observe are at least present in a given known host, in accordance with our recruitment hypothesis. The phylogenetic signal is erased very quickly. Yet, we aligned the tRNA sequences of the phages with those of all sequenced genomes of the host genus, and computed the average similarity between these sequences. The average over all 169 sequences is

70.74%. Unfortunately, it is impossible to make phylogenetic analyses of the tRNA genes found in phages to check the acquisition from the host because tRNA genes are very small and once acquired by phages they mutate 100 to 1000 fold faster than in the hosts per generation (Drake, 1991), which is further enhanced by the very high growth rates of phages.

We computed for each each phage the sum of its codon frequencies weighted by the number of exact cognate tRNAs it contains, and compared it to the value it would take if the same number of tRNAs were drawn at random from the host genome. We observed no significant association between the phage tRNA gene content and its codon usage bias (Kolmogorov test, $p = 0.68$).

We then computed the sum of the differences in codon usage between phage and host, weighted in the same way. This showed a highly significant association between tRNA presence and codon usage difference: the probability that the observed difference in codon usage arise via random tRNA uptake from the hosts is 4.10^{-3}. We conclude that the tRNA gene content in a bacteriophage is not due to a simple random drawing from the host tRNA distribution.

One usually assumes that tRNAs are integrated in the phage genomes either at the time of chromosome degradation or when there is imprecise excision of prophages. If tRNAs are recruited from the host genome without further selection, then one would expect temperate phages to contain more tRNAs than virulent phages (if we suppose their rates of acquisition of free DNA sequences to be equal) because they might employ both mechanisms. Yet, virulent phages have more, not less, tRNA genes than temperate phages. Since tRNA genes are not a random sample of the host tRNA genes we tried to identify the underlying selective process.

Modeling the acquisition and selective loss of tRNAs inside phage genomes.

We used a master equation approach to model the putative selective processes involved in the modulation of the phages tRNA gene content (see Methods). More precisely, we modeled the evolution of the probability $\mathcal{P}_{\alpha\beta,\bar{x}}(n)$ of phage α (whose host is β) to have n tRNAs of anticodon $\bar{x}$. We initially modeled only the random processes of acquisition and loss, i.e. our null model involves no selection upon the tRNA gene content. We defined the rate of acquisition, r, by normalizing the rate of loss to 1 (see Methods). We estimated the

value of r by maximum likelihood and found $r = 0.063$. This low value is in accordance with the fact that most phages have no or few tRNAs, as it coincides with the relative frequency of tRNAs in phages relative to their hosts.

Absence of selection of tRNAs for phage or host codon usage.

We then proceeded to introduce in the model different processes to identify the one which brought significant information. We specifically explored the relevance of three different selection processes to explain tRNA composition in phages. We started by considering the hypothesis that tRNAs are selected to match the most abundant codons in the phage genome. For this, we specified a rate of tRNA loss decreasing in $f_{\alpha,x}$ (the frequency of codon usage of phage α for codon x) and controlled by a selection coefficient s. We solved the new master equation, and found r and s by the maximum of the log–likelihood $\mathcal{L}$. The significance of a non-zero value of s can be computed by comparing the log–likelihood $\mathcal{L}_r$ for the null model with $s = 0$, i.e. no selection, and the log–likelihood $\mathcal{L}_{r,s}$ for the model with selection. The likelihood ratio method indicated that our estimation of s is not significantly different from 0 ($p = 0.15$), suggesting little, if any, tRNA selection based on the phage codon usage bias.

We then tested if phage tRNAs could be selected to match those codons that are rarer in the host. The estimated s value for this process was significant ($p = 0.018$), but only before applying the Bonferroni correction for our multiple statistical tests ($p = 0.072$). This selection process was therefore also discarded.

Our initial exploration of tRNA composition in phages showed very few cases of duplicated tRNA genes, since we only found 6 pairs and one triplet of similar tRNA species in a phage. Nevertheless, we tested explicitly the hypothesis that tRNA gene amplifications may contribute to explaining tRNA distribution in phages by simulating a process where tRNAs can be duplicated at a rate c. We considered too unlikely the hypothesis that a duplicated tRNA could mutate to become another species, because this could lead to a tRNA present on the phage and not in its host, which is not observed. The duplication rate, c was then estimated by maximum likelihood, but we inferred $c = 0$ as the most probable value. Hence, the most likely scenario for the multiple acquisition of similar tRNA species is the one of independent rounds of tRNA acquisition, not of gene duplication.

Selection for differences in codon usage between phage and host.

We finally tested the hypothesis that selection is based on the *difference* between the frequency of codons in the phage and its host. Our rationale was that selection could favor phages with tRNAs corresponding to codons that are abundant in the phage but rare in the host. These codons are expected to be poorly translated by the host machinery and lead to slow phage growth if not compensated by the phage own tRNAs. Solving the new model by maximum likelihood we found s to be significantly different from zero even with the correction for multiple tests ($p < 2.10^{-7}$). This strongly suggests that the selection process acting on tRNAs is based on the difference of codon usage between host and phage.

To further validate this conclusion we made three additional tests. Firstly, we solved the master equation model with a randomized dataset, generated by taking the observed values of the tRNA counts and associating them to random codon frequencies of the phage and hosts, taken among those observed. This randomization deleted the internal correlation between host and phage codon frequencies, and phage tRNA counts. As expected, the model gave non-significant values of the selection coefficient s when applied on this dataset.

Secondly, since most phages contain very few tRNAs, we used a binary model where $\mathcal{P}_{\alpha\beta,\bar{x}}(n)$ can only have the value 0, for no tRNA of a given type in the phage, and 1, for at least one such tRNA in the phage (see Methods). This binary model leads to some loss of information but is expected to be more robust. We found a similar p-value for the rejection of the hypothesis that s is equal to 0 ($p < 5.10^{-9}$). Thus, results seem robust and give strong support to the hypothesis that tRNAs are selected in phages to compensate for differences in codon usage between the phage and the host.

Finally, we tested if our results are robust to changes in the arbitrary selection of one among several known hosts. We made the same analysis by assuming each phage to infect all strains of a same species. This was done by creating an average genome representing the species genome (see Methods). Both the statistical analysis and the master equation modeling gave qualitatively and quantitatively the same results, highlighting their robustness.

Higher abundance of tRNAs in virulent phages

Virulent and temperate phages have very different ecologies. Thus, we investigated if both groups used tRNAs to compensate differences in codon usage with respect to the host. For this, we fitted the master equation model (without the non-significant gene duplication term) to the two sets of phages separately.

Splitting the phages leads to groups with low effectives. Yet, we still found a significant effect of selection for tRNA genes caused by the difference of codon usage between host and phage $\Delta f_{\alpha\beta,x}$, both for virulent and temperate phages (resp. $p < 5.10^{-7}$ and $p < 5.10^{-4}$). Thus, both types of phages contain tRNAs corresponding to their mid- to high-frequency codons (Fig. 3), which are also those showing the largest difference to the host codon usage.

Even if selection is present in both types of phages, the significance is higher for virulent phages, in spite of their smaller sample size. This is corroborated by three other observations. Firstly, in our sample virulent phages have an average of 7.9 tRNAs, whereas temperate phages only have 2 tRNAs (Fig. 1, significant difference, $p < 3.10^{-3}$). Secondly, the codon usage of the hosts correlates much better with the one of temperate phages (0.83 ± 0.03) than with virulent phages (0.61 ± 0.11). Thirdly, virulent phages tend to have stronger codon usage bias. To quantify this assumption, we used a measure of deviation of codon usage from a uniform distribution that accounts for the nucleotide composition of the genome, $\hat{N}_c{}'$ (Novembre, 2002). We computed its value for each phage and found that virulent phages are significantly more biased than temperate (Wilcoxon test, $p < 5.10^{-3}$).

Temperate and virulent phages are distinct in one major aspect. Temperate phages replicate both through lytic cycles and in the lysogenic state. While in the latter, temperate phages share the same mutational biases as the host. As a result, they tend to have a genome composition much closer to the one of the host than virulent phages, which share with other parasitic DNA a bias towards A+T richness relative to the host (Rocha and Danchin, 2002). This could lead temperate phages to show a codon usage bias closer to the one of the hosts, as observed. Since there are more differences in codon usage between the host and virulent phages, one would expect the latter to contain more tRNAs allowing to compensate for this difference.

Discussion

Bacteriophages have highly compact genomes that tend to lack any translation associated genes, with the notable exception of tRNAs. Our investigation on the reasons motivating this exception have shown that many genomes lack or have few tRNAs, whereas some genomes contain nearly as many tRNAs as some bacteria. There is a positive association between the size of the phage genome and the number of tRNA genes it contains. This suggests that tRNA genes are part of the phages accessory genome probably arising from multiple recruitment events and only being kept when selection for their presence is strong enough. We only found tRNAs in the genomes of dsDNA phages. The other phages may miss tRNAs either because they are much more compact, thus excluding non-essential information, and/or because the folding of tRNAs may pose problems in the organization of the chromosomes of RNA or ssDNA phages. The presence of tRNAs among dsDNA phages is coherent with the following evolutionary scenario. Firstly, tRNAs are recruited from the host chromosomes or from recombination with other phages co-infecting a bacterial cell. Secondly, these tRNAs are subject to frequent deletion following the deletional bias that is thought to predominate in the genomes of bacteria and phages (Lawrence et al., 2001, Mira et al., 2001). Yet, some tRNAs can provide for such an advantage as to counteract the effect of the deletion bias. As long as the advantage of carrying the tRNA overcomes the negative effect of increasing genome size and the deletion bias, the tRNA will be kept in the genome.

Our data indicate that tRNAs that are kept in phage genomes are those corresponding to codons abundant in the phage and rare in the host. This allows the phage to gain a clear-cut advantage over its competitors by translating its proteins more efficiently, reducing its latency time and increasing the reproduction rate. This may be balanced by the time necessary to replicate the tRNA sequence on the phage genome, but the latter effect must be weak since tRNA genes are very small.

For a phage to carry and express a tRNA that is already abundant in the host would give little benefit since it would have a small relative effect on the phage environment during infection. Instead, expressing a tRNA which is rare in the host may provide a decisive benefit to the phage if it corresponds to a highly frequent codon in its own genes. So the optimal configuration for a phage would be to carry tRNAs matching the codons it uses much more than its host, which is indeed what we observe.

It might be argued at this stage that the best strategy for the phage would be to perfectly mimic the host codon and tRNA usage. Although we did observe significant correlations between the codon usage bias of the phage and its host, this strategy may not be perfectly attainable for two reasons. Firstly, tRNA concentration and codon usage bias in bacteria vary with the physiological state of growth and in fast-growing bacteria they are mostly determined by the physiological requirements of the exponential phase (Dong et al., 1996, Kurland, 1991). These are not the conditions prevailing during the lytic cycle. Secondly, the genomes of phages tend to be AT richer than the genomes of their hosts, which necessarily affects codon usage. The reasons of this bias may be mutational, error prone polymerases or inefficient repair, or selective, adaptation to the AT richness of the bacterial cytoplasm. In any case, they are more important for virulent than for temperate phages (Rocha and Danchin, 2002). As a result, phage codon usage cannot perfectly fit the host translation machinery and the recruitment of the necessary tRNAs becomes adaptive. Such effect will be more important if the compositional gap is important, if the latency times are shorter, if the phage codon usage bias is higher and if the phage depends exclusively on horizontal transmission to reproduce. These conditions are met for virulent phages which, accordingly, contain more tRNAs than temperate ones.

We observed higher codon usage bias in virulent phages. Why should that be? We speculate that it is because virulence phages replicate faster and need to translate very efficiently their mRNAs. Although we could not find data on latency times (the average time it takes a phage to lyze the host after infection) for most our phages, we did find a recent work describing these values for some *E. coli* phages (De Paepe and Taddei, 2006). When comparing these latency times for dsDNA phages we found a statistically significant difference, with lower values for virulent phages (28 versus 54 minutes on average, $p < 0.02$, Wilcoxon test). Thus, increased codon usage bias in virulent phages might result from selection for lower latency times. Virulent phages would then tend to select more strongly for the presence of tRNA genes, both because they are more at odds with the host codon usage and because they are under stronger selection for codon usage bias.

Other models have been put forward to explain the presence of tRNAs among phage genomes, e.g. models where the presence of tRNAs allows the phage to resist to anticodon nucleases in the host (Blanga-Kanfi et al., 2006, Kaufmann, 2000); the employment of alternative genetic codes (see by example Bacher et al. (2003)); or a better integration of lysogenic phages in-

side the host chromosome (Canchaya et al., 2004). The first two hypotheses are based on very few observations and it is still unclear if they are indeed strategies to evade host response and if they are frequently found in nature. The last model is contradicted by our observation that lytic phages contain more tRNAs than temperate ones and that the populations of tRNA genes in phages are not random samples of the host repertoire. Moreover, most known temperate phages inserting in a tRNA gene of the host genome (e.g. *E. coli* phage P22, P4 or Lambda) have no tRNA genes. This shows that these genes are not necessary for phage integration in the bacterial chromosome. In contrast, our model is grounded on the well known advantage of carrying tRNA genes for translation optimization of the cognate codons and was confirmed by several different tests and controls.

Phages are a major vehicle of lateral gene transfer in bacteria. However, tRNA genes are essential, housekeeping and information related genes, which are expected to be the least prone to horizontal gene transfer (Jain et al., 1999). Their occurrence in phages may lead to the lateral transfer of tRNAs from one cellular genome to another, but our data indicate that their presence is much more likely to be caused by the advantage they confer to phages in hosts depleted in these tRNAs. Recently, a wealth of viruses and phages with large genomes has been discovered, highlighting the potential diversity in terms of their genome structure and different functionalities of viruses (Ghedin and Claverie, 2005). Here, we showed that these may include the recuperation of essential cellular genes from the host to optimize the expression of their own genes in view of infecting those same hosts. Thus, the very large viruses may contain a significant number of translation associated genes for selective purposes. Surely, the growing number of couples of host/virus complete genomes will reveal the extent and other variants of this evolutionary strategy.

Methods

Data

The genomes of phages and hosts were downloaded from Genbank. The assignment of a phage to a host and to a class of virulence was done using data collected from the literature. Sometimes this classification was impossible because the genes in the GenBank file were not functionally annotated and

no information was available in the literature. These phages (4 out of 37) were used in the generic analysis but were discarded in the comparison between virulent and temperate phages.

The tRNAs sequences in both the phage and the host genomes were detected using tRNAscan_SE (Lowe and Eddy, 1997) with default parameters for prokaryotic genomes. We started with a data set of 193 phage genomes, of which 48 contain a total of 214 tRNAs. We excluded from further analysis those tRNAs that could mislead the results of our statistical analysis, i.e. pseudogenes, tRNAs for SeC, undefined tRNAs and the rare tRNAs absent from the chosen host. We also removed all phages for which no complete sequence of a host was publicly available and those for which there was no annotation. We thus used a final dataset containing 37 phages corresponding to 15 hosts and including a total of 169 tRNAs. To build the dataset of host genomes we randomly selected for each phage one of its bacterial hosts, when several were fully sequenced, e.g. for *E. coli*. An alternative analysis using mixtures of several host genomes showed similar results. To test the robustness of the model, statistical tests and the master equation model analysis were also performed on average host genomes. For each host, we chose randomly one genome in each genus of the same family, and built an average genome by averaging between all these genomes the frequencies of codon usage and the tRNA gene content. Tests were then performed considering these genomes as the host, representing the putative or unknown wide host range of all phages. We detected tRNAs in the host genomes in the same way. The tRNAs located in the host genomes inside a prophagic region were removed from the pool of data to be analyzed, using published information on prophage locations (Canchaya et al., 2004). This avoids including a circularity in the analysis, i.e. comparing tRNAs of phages with those of their prophages. The table with the names and accession numbers of phage genomes and their classification into temperate and virulent are published as supplementary material.

tRNA alignment

Phage tRNAs were aligned against those of all hosts of the same genus as the chosen host. The alignments were done using the "needle" program (Rice et al., 2000) with a constant gap penalty of 10, which allowed a better alignment of the sequences. Similarities were measured using the same software.

Statistical tests

Comparison of the virulent and temperate phages total tRNA gene content

We used a Monte-Carlo method to test for a statistically significant difference in the number of tRNAs in virulent and temperate phages containing tRNAs. To this aim, we estimated the probability of finding as many tRNAs or more in a group, as observed in the real case. We considered the real distribution of tRNAs counts inside the phage genomes. Here, α denotes a phage, N_α is the number of tRNAs within phage α, and $N_{\alpha,\bar{x}}$ is the number of tRNAs having anticodon $\bar{x}$ in phage α. We drew at random in the $\{N_\alpha\}$ set 12 values, corresponding to our 12 virulent phages, and sum them. We then estimated the probability distribution of this sum, and computed the probability of having a sum superior or equal to 95, the observed number of tRNAs in the virulent phages.

Statistical test of the random uptake hypothesis

We designed two indicator variables, A_α and $B_{\alpha\beta}$, that allow testing respectively i) if tRNAs tend to correspond to over-represented codons or ii) if tRNAs correspond to codons used more in the phage than in its host. As a first order approximation, we only considered the correspondence between tRNAs of an anticodon and frequencies of the perfectly matching codons, as these are usually regarded as the optimal codons. We then computed A_α, the average frequency of codon usage restricted to codons for which a matching tRNA is present in the genome, and $B_{\alpha\beta}$, the average difference in codon usage between the phage and its host, for the same codons. $f_{\alpha,x}$ is the frequency of codon x in phage α, computed on all its genes and relative to all other codons. $\Delta f_{\alpha\beta,x}$ is the difference of the frequencies of codon usage, for codon x, between the phage α and its host β. $\bar{x}$ is the perfectly matching anticodon for codon x, using Watson-Crick pairing rules. We computed the indicators as:

$$A_\alpha = \frac{1}{N_\alpha} \sum\nolimits_x N_{\alpha,\bar{x}} f_{\alpha,x}, \tag{1a}$$

$$B_{\alpha\beta} = \frac{1}{N_\alpha} \sum\nolimits_x N_{\alpha,\bar{x}} \Delta f_{\alpha\beta,x}. \tag{1b}$$

To assess the statistical significance of these indicators, we drew at random from the host as many tRNAs as contained in the phage, N_α. By repeating

this procedure 100,000 times we obtained the expected distribution of tRNAs in the phage under a model where tRNAs are randomly sampled from the host genome. This allows obtaining the probability P_α^A and $P_{\alpha\beta}^B$ to have A_α or $B_{\alpha\beta}$ randomly greater or equal to the observed value of the indicators, in each phage. The significant departure of both sets of 37 probabilities $\{P_\alpha^A\}$ and $\{P_{\alpha\beta}^B\}$ from an uniform distribution was assessed by a Kolmogorov test.

Master equation model

The model

Suppose that the tRNAs of the phage α are taken at random among the tRNAs of its host β, with a rate r supposed to be unique for all phages and anticodons. All tRNAs for different anticodons $\bar{x}$ are considered independent. Without any lack of generality, we set the rate of loss of the tRNAs to 1 (changing this value would result in the same equation with a rescaling of time t). We denote by $H_{\beta,\bar{x}}$ the number of tRNAs of host β for anticodon $\bar{x}$. The probability $\mathcal{P}_{\alpha\beta,\bar{x}}(n)$ of phage α having n tRNAs of anticodon $\bar{x}$, is governed by the master equation:

$$\frac{\partial \mathcal{P}_{\alpha\beta,\bar{x}}(n)}{\partial t} = rH_{\beta,\bar{x}}\mathcal{P}_{\alpha\beta,\bar{x}}(n-1) + (n+1)\,\mathcal{P}_{\alpha\beta,\bar{x}}(n+1) - (rH_{\beta,\bar{x}} + n)\,\mathcal{P}_{\alpha\beta,\bar{x}}(n), \tag{2}$$

where the dependence of $\mathcal{P}_{\alpha\beta,\bar{x}}$ on time does not appear for easier reading. This initial equation is the mathematical formulation of the hypothesis that all tRNAs present in the genomes of phages are drawn at random from host genomes with a constant rate r, and lost at a rate normalized to unity.

We then added a selection parameter, s, to model how selection changes the probabilities of tRNAs being fixed in the populations of phages. This is achieved by allowing for three different processes which selection acts upon: i) the frequency $f_{\alpha,x}$ of codon x in the phage α; ii) the difference of codon frequencies between the host and the phage, $\Delta f_{\alpha\beta,x} = f_{\alpha,x} - f_{\beta,x}$; iii) the opposite of the frequency $f_{\beta,x}$ of codon x in the host genome. The quantity under selection is denoted hereafter by the symbol F (indicating $f_{\alpha,x}$, $\Delta f_{\alpha\beta,x}$ or $-f_{\beta,x}$ for the three cases (i), (ii) and (iii), respectively). All these models are described by the equation:

$$\begin{aligned}\frac{\partial \mathcal{P}_{\alpha\beta,\bar{x}}(n)}{\partial t} &= rH_{\beta,\bar{x}}\mathcal{P}_{\alpha\beta,\bar{x}}(n-1) + (n+1)e^{-sF}\mathcal{P}_{\alpha\beta,\bar{x}}(n+1) \\ &\quad - [rH_{\beta,\bar{x}} + ne^{-sF}]\mathcal{P}_{\alpha\beta,\bar{x}}(n).\end{aligned} \tag{3}$$

A positive value of s stands for a selective process tending to keep the tRNAs having a high value of the selected trait F. The exponential form of this selection rate is chosen for simplicity, and since the values of s are small it is equivalent to using a selection linear in s.

We also considered the hypothesis that a phage tRNA can multiply in the genome with rate c. The master equation corresponding to this case reads:

$$\begin{aligned}\frac{\partial \mathcal{P}_{\alpha\beta,\bar{x}}(n)}{\partial t} &= [rH_{\beta,\bar{x}} + c(n-1)]\mathcal{P}_{\alpha\beta,\bar{x}}(n-1) + (n+1)e^{-sF}\mathcal{P}_{\alpha\beta,\bar{x}}(n+1) \\ &\quad - [rH_{\beta,\bar{x}} + n(c+e^{-sF})]\mathcal{P}_{\alpha\beta,\bar{x}}(n).\end{aligned} \tag{4}$$

The stationary solution to equation (4) can be derived analytically. Here, we show the general solution, with $c = 0$ or $s = 0$ being special cases. To find the solution to (4), we first recast the master equation (4) in terms of the generating function $\phi(\lambda, t) = \sum_n e^{\lambda n}\mathcal{P}_{\alpha\beta,\bar{x}}(n,t)$. This gives a new differential equation, which turns out to be of the hypergeometric type and can thus be solved analytically. By applying the opposite transformation from the generating function to $\mathcal{P}_{\alpha\beta,\bar{x}}$, we finally obtain the following expression:

$$\mathcal{P}_{\alpha\beta,\bar{x}}(n) = \frac{1}{n!}\left(1 - ce^{sF}\right)^{\frac{rH_{\beta,\bar{x}}}{c}} (ce^{sF})^n \prod_{i=0}^{n-1}\left(\frac{rH_{\beta,\bar{x}}}{c} + i\right). \tag{5}$$

One can easily check by direct substitution of (5) into (4) that the right-hand term of (4) indeed vanishes. In the limit $c \to 0$, the previous expression reduces to the expected Poisson law for the stationary probability :

$$\lim_{c\to 0} \mathcal{P}_{\alpha\beta,\bar{x}}(n) = \frac{1}{n!}\varphi^n\, e^{-\varphi}. \tag{6}$$

with $\varphi = rH_{\beta,\bar{x}}e^{sF}$. This solution has the trivial limit $\varphi \to rH_{\beta,\bar{x}}$ as $s \to 0$. In this case r is the only parameter and we simply model random acquisition and loss of tRNAs.

Parameter fit

We use a maximum likelihood method to find the most probable values of the parameters r, s and c. Note that each of the 3 parameters is supposed to be identical for each phage and anticodon $\bar{x}$. Firstly the log–likelihood of the set of observed counts is computed :

$$\ln(\mathcal{L}(r,s,c)) = \sum_{\beta,\alpha,x} \ln(\mathcal{P}_{\alpha\beta,\bar{x}}(N_{\alpha,\bar{x}})), \tag{7}$$

where the dependence on r, s and c on the left hand side is brought by the expression (5). We verified that the log–likelihood landscape is relatively smooth, allowing us to maximize it by simply computing its value at every point of a 3 dimensional grid of constant step for each parameter, and verifying the maximum thus found by using a steepest gradient method. For very low values of c, which can be computationally tricky, we analytically compute $\frac{\ln(\mathcal{L})}{\partial c}\big|_{c=0}$ and find it always negative in a close neighborhood of the parameters r_{max} and s_{max} which maximize $\ln(\mathcal{L}(r,s,0))$. This analysis, combined to the absence of solutions found by the other methods for $c > 10^{-10}$, confirm that the most probable value of c is 0. Then, the two parameters r and s correspond to the zeros of the derivatives of $\ln(\mathcal{L}(r,s,0))$ relative to them. Computing the derivative of (7) and equating it to zero, gives the relation :

$$r = \frac{\sum_{\alpha,x} N_{\alpha,\bar{x}}}{\sum_{\beta,\alpha,x} H_{\beta,\bar{x}} e^{s\Delta f_{\alpha\beta,x}}}. \tag{8}$$

The most probable value of r is the one satisfying equation (8) and maximizing the log–likelihood (7). This result gives also directly the value of r when s is taken equal to 0, as in the first model.

The significance of including an additional parameter in the model is computed by the standard likelihood ratio method (Saporta, 1990).

Binary model

In the majority of cases, there is only one tRNA for a given anticodon, per phage genome. To confirm the results of the previous model we designed a simpler two-state model accounting only for presence (+) or absence (-) of tRNAs for a given anticodon. Using the same hypothesis of random uptake and selection as before (c is set to zero in this model), we have :

$$\frac{\partial \mathcal{P}^{(+)}_{\alpha\beta,\bar{x}}}{\partial t} = r H_{\beta,\bar{x}} \mathcal{P}^{(-)}_{\alpha\beta,\bar{x}} - e^{-s\Delta f_{\alpha\beta,x}} \mathcal{P}^{(+)}_{\alpha\beta,\bar{x}}, \tag{9a}$$

$$\frac{\partial \mathcal{P}^{(-)}_{\alpha\beta,\bar{x}}}{\partial t} = -r H_{\beta,\bar{x}} \mathcal{P}^{(-)}_{\alpha\beta,\bar{x}} + e^{-s\Delta f_{\alpha\beta,x}} \mathcal{P}^{(+)}_{\alpha\beta,\bar{x}}. \tag{9b}$$

Equations (9) are derived and analyzed as previously. The solution of this system is :

$$\mathcal{P}^{(+)}_{\alpha\beta,\bar{x}} = \frac{rH_{\beta,\bar{x}}}{rH_{\beta,\bar{x}} + e^{-s\Delta f_{\alpha\beta,x}}}, \tag{10a}$$

$$\mathcal{P}^{(-)}_{\alpha\beta,\bar{x}} = \frac{e^{-s\Delta f_{\alpha\beta,x}}}{rH_{\beta,\bar{x}} + e^{-s\Delta f_{\alpha,\beta,x}}}. \tag{10b}$$

In this case, the maximum of the log-likelihood was computed on a grid of precision 10^{-2} for s, sufficient to demonstrate that s is significantly different from 0.

Figure Legends

Figure 1: Distribution of the number of tRNAs inside phage genomes. Empty bars stand for temperate phages, grey ones for virulent phages and patterned ones for phage of unknown type. Note the heterogeneity of the counts, and the tendency for virulent phages to have more tRNAs than temperate ones. Names are indicated for phages with 19 or more tRNAs in the form "(host species) phage".

Figure 2: Distribution of the frequencies of codon usage in phage genomes. The solid line is the distribution of codon frequencies; the dot line, the distribution of codon frequencies, restricted to codons matching a tRNA on the considered phage genome. Note the peak around f=0.03. The dash-dot line is the frequency distribution for all codons of all hosts. In the inset, the cumulated probability distribution (probability that a random tRNA will have a frequency superior or equal to the one given in abscissa) of the tail of the frequency distributions of the phages for all codons (solid line), and only the ones matching a tRNA (dot line). Note the difference, indicating an excess of tRNAs matching high-frequency codons in phages.

Figure 3: Distribution of the frequencies of codon usage in virulent phage genomes (up) and temperate phage genomes (down). Light grey filled bars, the distribution of codon frequencies, for all codons; black empty histogram bars, the distribution of codon frequencies, restricted to codons matching a tRNA on the considered phage genome. Note the difference between the distributions in both the virulent and the temperate case.

Figures

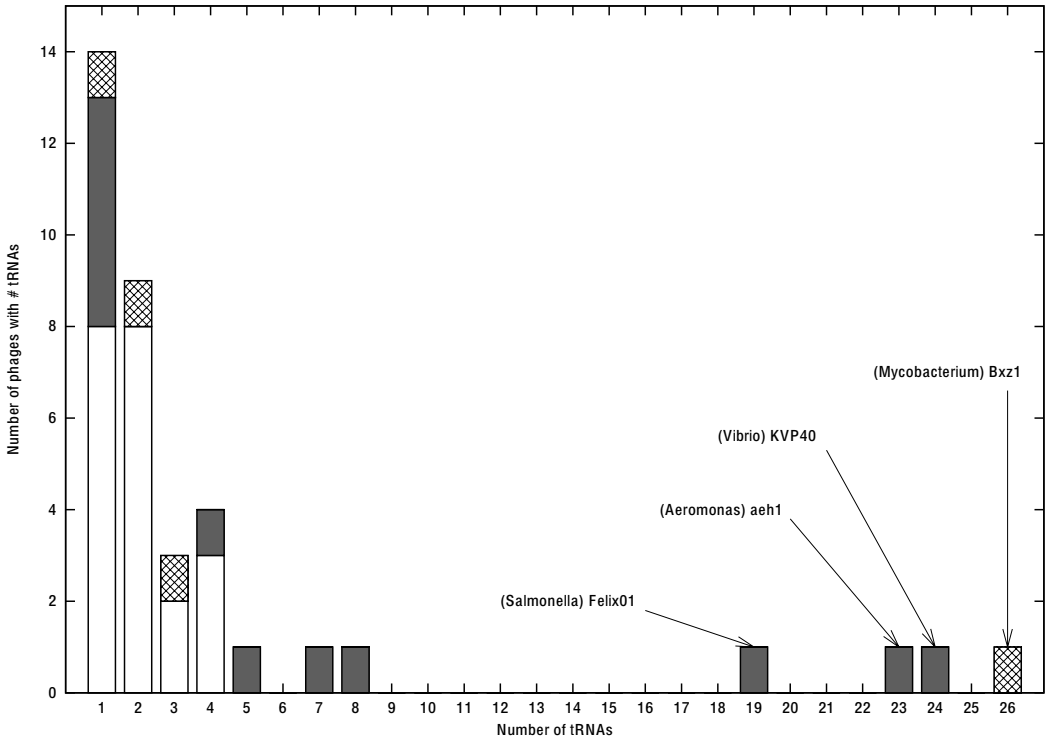


Figure 1.

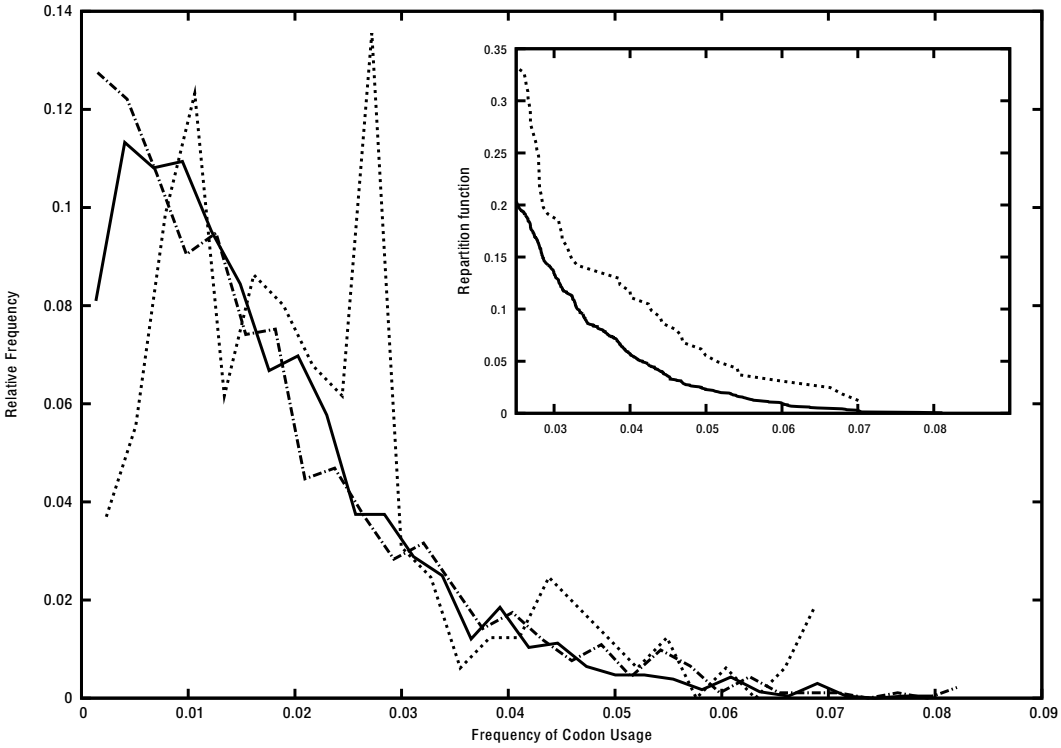

Relative Frequency
Frequency of Codon Usage
Repartition function
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.35
0.3
0.25
0.2
0.15
0.1
0.05

Figure 2.

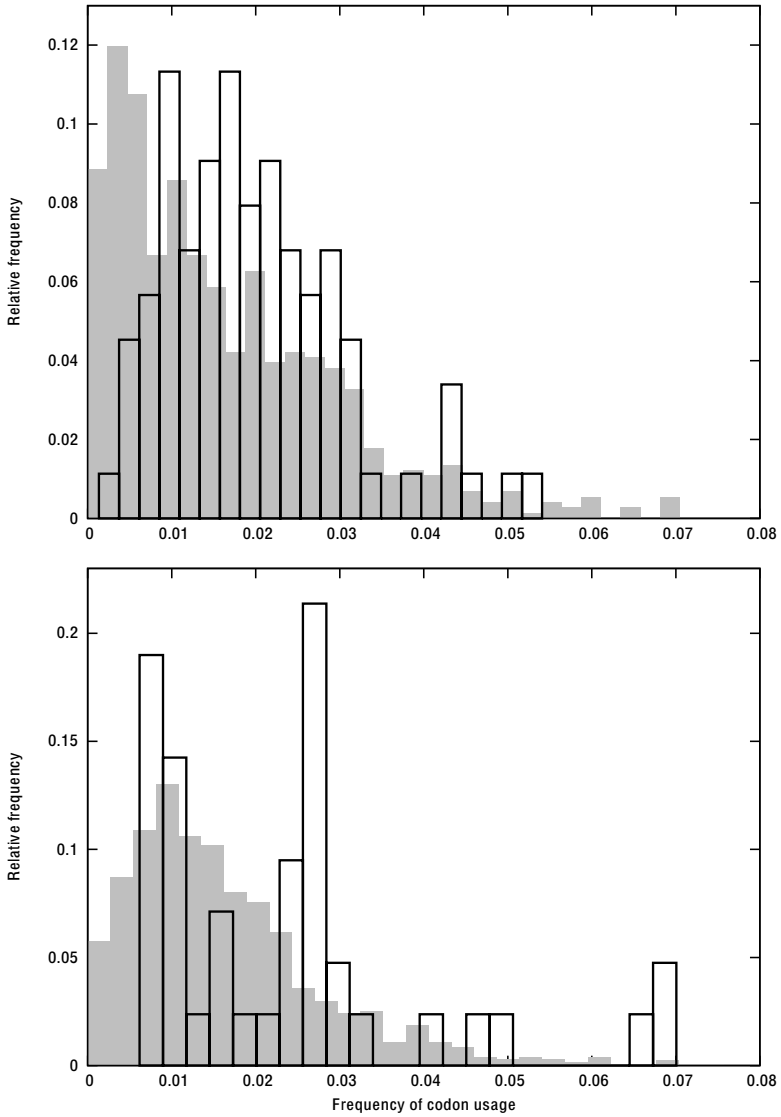

Relative frequency
0.12
0.1
0.08
0.06
0.04
0.02
0
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Relative frequency
0.2
0.15
0.1
0.05
0
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Frequency of codon usage

Figure 3.

References

Andersson, S. and Kurland, C.G. 1990. Codon preferences in free-living microorganisms. *Microbiol. Mol. Biol. Rev.* **54**: 198–210.

Bacher, J., Bull, J. and Ellington, A. 2003. Evolution of phage with chemically ambiguous proteomes. *BMC Evol. Biol.* **3**: 24–35.

Bailly-Bechet, M., Danchin, A., Iqbal, M., Marsili, M. and Vergassola, M. 2006. Codon usage domains over bacterial chromosomes. *PLoS Comp. Biol.* **2**: e37.

Blanga-Kanfi, S., Amitsur, M., Azem, A. and Kaufmann, G. 2006. PrrC-anticodon nuclease: functional organization of a prototypical bacterial restriction rnase. *Nucleic Acids Res.* **34**: 3209–3219.

Bulmer, M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**: 897–907.

Campbell, A.M. 1992. Chromosomal insertion sites for phages and plasmids. *J. Bacteriol.* **174**: 7495–7499.

Canchaya, C., Fournous, G. and Brussow, H. 2004. The impact of prophages on bacterial chromosomes. *Mol. Microbiol.* **53**: 9–18.

Casjens, S. 2003. Prophages and bacterial genomics: what have we learned so far? *Mol. Microbiol.* **49**: 277–300.

Cowe, E. and Sharp, P.M. 1991. Molecular evolution of bacteriophages: Discrete patterns of codon usage in T4 genes are related to the time of gene expression. *J. Mol. Evol.* **V33**: 13–22.

Daubin, V. and Ochman, H. 2004. Start-up entities in the origin of new genes. *Curr. Opin. Genet. Dev.* **14**: 616–619.

De Paepe, M. and Taddei, F. 2006. Viruses' life history: Towards a mechanistic basis of a trade-off between survival and reproduction among phages. *PLoS Biol.* **4**: e193.

Dong, H., Nilsson, L. and Kurland, C.G. 1996. Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J. Mol. Biol.* **260**: 649–663.

Drake, J. 1991. A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Nat. Acad. Sci. USA* **88**: 7160–7164.

Elf, J., Nilsson, D., Tenson, T. and Ehrenberg, M. 2003. Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science* **300**: 1718–1722.

Ghedin, E. and Claverie, J.M. 2005. Mimivirus relatives in the Sargasso sea. *Virol. J.* **2**: 62–67.

Gouy, M. and Gautier, C. 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* **10**: 7055–7074.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.* **9**: r43–74.

Ikemura, T. 1985. Codon usage and trna content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2**: 13–34.

Ikemura, T. 1981. Correlation between the abundance of *Escherichia coli* tRNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.* **146**: 1–21.

Jain, R., Rivera, M.C. and Lake, J.A. 1999. Horizontal gene transfer among genomes: The complexity hypothesis. *Proc. Natl. Acad. Sci. USA* **96**: 3801–3806.

Kaufmann, G. 2000. Anticodon nucleases. *Trends Biochem. Sci.* **25**: 70–74.

Knopf, C.W. 1998. Evolution of viral DNA-dependent DNA polymerases. *Virus Genes* **16**: 47–58.

Krakauer, D.C. and Jansen, V.A.A. 2002. Red Queen dynamics of protein translation. *J. Theor. Biol.* **218**: 97–109.

Kropinski, A.M. and Sibbald, M.J. 1999. Transfer RNA genes and their significance to codon usage in the *Pseudomonas aeurginosa* lambdoid bacteriophage D3. *Can. J. Microbiol.* **45**: 791–796.

Kunisawa, T. 1992. Synonymous codon preferences in bacteriophage T4: A distinctive use of transfer RNAs from T4 and from its host *Escherichia coli*. *J. Theor. Biol.* **159**: 287–298.

Kunisawa, T., Kanaya, S. and Kutter, E. 1998. Comparison of synonymous codon distribution patterns of bacteriophage and host genomes. *DNA Res.* **5**: 319–326.

Kurland, C.G. 1991. Codon bias and gene expression. *FEBS Lett.* **285**: 165–169.

Lawrence, J.G., Hendrix, R.W. and Casjens, S. 2001. Where are the pseudogenes in bacterial genomes? *Trends Microbiol.* **9**: 535–540.

Lowe, T. and Eddy, S. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**: 955–964.

Miller, E.S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T. and Ruger, W. 2003. Bacteriophage T4 Genome. *Microbiol. Mol. Biol. Rev.* **67**: 86–156.

Mira, A., Ochman, H. and Moran, N.A. 2001. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17**: 589–596.

Muto, A. and Osawa, S. 1987. The Guanine and Cytosine Content of Genomic DNA and Bacterial Evolution. *Proc. Natl. Acad. Sci. USA* **84**: 166–169.

Novembre, J.A. 2002. Accounting for Background Nucleotide Composition When Measuring Codon Usage Bias. *Mol. Biol. Evol.* **19**: 1390–1394.

Ohnishi, M., Tanaka, C., Kuhara, S., Ishii, K., Hattori, M., Kurokawa, K., Yasunaga, T., Makino, K., Shinagawa, H., Murata, T. et al. 1999. Chromosome of the enterohemorrhagic *Escherichia coli* O157:H7; comparative analysis with K-12 MG1655 revealed the acquisition of a large amount of foreign DNAs. *DNA Res.* **6**: 361–368.

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M. and Claverie, J.M. 2004. The 1.2-megabase genome sequence of Mimivirus. *Science* **306**: 1344–1350.

Rice, P., Longden, J. and Bleasby, A. 2000. EMBOSS: The european molecular biology open software suite. *Trends Genet.* **16**: 276–277.

Rocha, E.P.C. and Danchin, A. 2002. Base composition bias might result from competition for metabolic resources. *Trends Genet.* **18**: 291–294.

Rocha, E.P. 2004. Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.* **14**: 2279–2286.

Saporta, G. 1990. *Probabilités, Analyse des données et Statistique*. Editions Technip.

Sau, K., Gupta, S.K., Sau, S. and Ghosh, T.C. 2005. Synonymous codon usage bias in 16 *Staphylococcus aureus* phages: Implication in phage therapy. *Virus Res.* **113**: 123–131.

Sharp, P.M., Bailes, E., Grocock, R.J., Peden, J.F. and Sockett, R.E. 2005. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.* **33**: 1141–1153.

Sharp, P.M., Rogers, M.S. and McConnell, D.J. 1985. Selection pressures on codon usage in the complete genome of bacteriophage T7. *J. Mol. Evol.* **21**: 150–160.

Suttle, C.A. 2005. Viruses in the sea. *Nature* **437**: 356–361.

Turner, P.E. and Chao, L. 1999. Prisoner's dilemma in an RNA virus. *Nature* **398**: 441–443.

Waldor, M.K. and Mekalanos, J.J. 1996. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* **272**: 1910–1914.

Weinbauer, M.G. 2004. Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* **28**: 127–181.

Weiss, S.B., Hsu, W.T., Foft, J.W. and Scherberg, N.H. 1968. Transfer RNA coded by the T4 bacteriophage genome. *Proc. Natl. Acad. Sci. USA* **61**: 114–121.

Wilson, J.H. 1973. Function of the bacteriophage T4 transfer RNA's. *J. Mol. Biol.* **74**: 753–754.

aeh1	NC_005260	Lytic	Aeromonas_hydrophila_ATCC_7966	NC_008570
Bcep22	NC_005262	Lytic	Burkholderia_cenocepacia_AU_1054	CP000378
Bcep781	NC_004333	Lytic	Burkholderia_cenocepacia_AU_1054	CP000378
933W	NC_000924	Temperate	Escherichia_coli_K12	U00096
P27	NC_003356	Temperate	Escherichia_coli_K12	U00096
phiP27	NC_003356	Temperate	Escherichia_coli_K12	U00096
RB69	NC_004928	Temperate	Escherichia_coli_K12	U00096
Sf6	NC_005344	Temperate	Escherichia_coli_K12	U00096
Stx2	NC_003525	Temperate	Escherichia_coli_K12	U00096
T4	AF158101	Lytic	Escherichia_coli_K12	U00096
VT2-Sa	NC_000902	Temperate	Escherichia_coli_K12	U00096
HP1	U24159	Temperate	Haemophilus_influenzae	L42023
HP2	NC_003315	Temperate	Haemophilus_influenzae	L42023
phigle	NC_004305	Temperate	Lactobacillus_plantarum	AL935263
Lj928	NC_005354	Temperate	Lactobacillus_johnsonii_NCC_533	AE017198
Lj965	NC_005355	Temperate	Lactobacillus_johnsonii_NCC_533	AE017198
bil170.10	NC_001909	Lytic	Lactococcus_lactis	AE005176
bil311.10	NC_002670	Temperate	Lactococcus_lactis	AE005176
bil312.10	NC_002671	Temperate	Lactococcus_lactis	AE005176
sk1	AF011378	Lytic	Lactococcus_lactis	AE005176
TP901-1	NC_002747	Temperate	Lactococcus_lactis	AE005176
Bxz1	NC_004687	Unknown	Mycobacterium_smegmatis_MC2_155	NC_008596
Bxz2	NC_004682	Unknown	Mycobacterium_smegmatis_MC2_155	NC_008596
CJW1	NC_004681	Unknown	Mycobacterium_smegmatis_MC2_155	NC_008596
omega	NC_004688	Unknown	Mycobacterium_smegmatis_MC2_155	NC_008596
d29	AF022214	Lytic	Mycobacterium_tuberculosis_CDC1551	AE000516
I5	Z18946	Temperate	Mycobacterium_tuberculosis_CDC1551	AE000516
D3	NC_002484	Temperate	Pseudomonas_aeruginosa	AE004091
PaP3	NC_004466	Temperate	Pseudomonas_aeruginosa	AE004091
phiKZ	NC_004629	Lytic	Pseudomonas_aeruginosa	AE004091
FelixO1	NC_005282	Lytic	Salmonella_typhi	AL513382
ST64T	NC_004348	Temperate	Salmonella_typhi	AL513382
phiC31	NC_001978	Temperate	Streptomyces_avermitilis	BA000030
K	NC_005880	Lytic	Staphylococcus_aureus_Mu50	BA000017
G1	NC_007066	Temperate	Staphylococcus_aureus_Mu50	BA000017
KVP40	NC_005083	Lytic	Vibrio_fischeri_ES114	CP000020
VpV262	NC_003907	Lytic	Vibrio_parahaemolyticus	BA000031

Table S1: Accession numbers and classification of the virulence of the phages used, with the corresponding host accession number.

7.3 Perspectives

Notre travail a permis de clarifier les hypothèses concernant la distribution des ARNt dans les génomes de phages, et ses liens avec leur biais d'usage de codons. Notamment, il a permis de proposer une hypothèse de sélection des ARNt pour la compensation des différences d'usage du code entre le phage et son hôte, et a montré que cette sélection était plus importante pour les phages virulents, dont l'usage du code est plus éloigné de celui de leurs hôtes que pour les phages tempérés (Rocha and Danchin, 2002). Ces travaux sont d'une importance particulière car ils montrent que l'acquisition de composants du système de traduction peut apporter un bénéfice sélectif à des phages, ce qui permettrait d'expliquer partiellement le grand nombre de gènes codant pour des molécules impliquées dans la traduction trouvés récemment chez des virus géants de la famille de Mimivirus (Ghedin and Claverie, 2005; Raoult et al., 2004). De plus, on peut élaborer à partir de ces résultats d'autres hypothèses sur le transfert horizontal de gènes essentiels par des phages, dont on supposait qu'il devait être très restreint, les protéines essentielles d'un organisme étant impliquées dans trop d'interactions pour pouvoir être utilisées par d'autres espèces (Jain et al., 1999). Si les phages peuvent acquérir ce type de gènes, et qu'en contrepartie ces gènes leur apportent un avantage suffisant pour être conservés à long terme dans leur génome, on peut imaginer que les mutations sur ces gènes puissent s'accumuler dans le phage au fil du temps – alors qu'elles seraient contre-sélectionnées chez l'hôte originel –, finissant par autoriser leur transfert à d'autres espèces. De nombreux modèles peuvent donc être développés à partir de nos résultats, concernant à la fois l'évolution des phages et celle de leurs hôtes.

Épilogue

Cette thèse a, entre autres, permis de suggérer le rôle que pouvait avoir le biais d'usage du code dans la structuration des chromosomes bactériens, et dans la dynamique d'acquisition de matériel génétique par les phages. Ces deux aspects ouvrent de nombreuses possibilités de continuation. Pour le premier, l'analyse précise des bénéfices dû à l'organisation chromosomique, ou les implications de ces résultats dans un domaine comme la biologie synthétique, pourraient être abordées. Quant au second, il amène des questions qui s'intègrent parfaitement dans le cadre des recherches actuelles sur le transfert horizontal, et pourrait servir de base pour analyser plus en détails l'impact de ce phénomène sur les populations bactériennes, avec les questions sur la santé et l'environnement que cela implique ; ou encore être utilisé comme point de départ pour étudier l'évolution des virus et, peut-être, affiner les hypothèse actuelles à ce sujet. De nombreuses ouvertures de ces travaux existent donc, et feront l'objet de projets ultérieurs.

Les perspectives scientifiques de chacun de mes travaux étant déjà détaillées dans les sections correspondantes, je ne m'étendrai pas plus, et je profiterai de cette courte conclusion pour envisager le futur. D'un point de vue personnel, je me permettrai de dire que cette thèse a été une grande réussite. Ces trois années m'ont énormément appris, en terme à la fois de méthodes, de connaissances scientifiques, mais également – du moins en ai-je l'impression – en terme de qualités professionnelles et humaines au sens général. Elles m'ont également montré l'étendue de ce qu'il me restait à apprendre, et peut-être à faire. J'ai pu prendre part à l'explosion scientifique actuelle, dans tous les domaines que j'ai abordé, et dans lesquels je compte continuer à m'impliquer. Avec la reconnaissance du transfert horizontal comme moteur évolutif, l'étude de l'évolution bactérienne est complètement modifiée, et doit englober l'étude des virus et des bactériophages, ces organismes qui peuvent avoir un impact énorme sur la santé, aussi bien en termes négatifs que positifs. De plus, j'ai pu constater à la fois l'importance de la collaboration entre physiciens, informaticiens et biologistes – qui permet d'apporter les différentes expertises nécessaires à la compréhension des problèmes de génomique – et les difficultés qu'elle entraîne : les concepts, et le vocabulaire même, de domaines scientifiques *a priori* étrangers sont ardus à appréhender, et encore plus à intégrer à part entière parmi les connaissances que l'on peut avoir. J'espère donc avoir l'occasion, dans la suite de mon parcours, de continuer à participer aux interactions grandissantes entre ces disciplines, et éventuellement de tenter de faciliter les liens entre elles, et d'établir des ponts entre les champs scientifiques qu'elles recouvrent.

Annexe A

Modélisation d'un écosystème hydrothermal

Avant ma thèse, mon travail de DEA avait consisté à développer un simulateur d'écosystème hydrothermal profond, au laboratoire "Modélisation des systèmes Biologiques Intégrés" de l'Université Pierre et Marie Curie – Paris 6. Ces environnements particulièrement rudes sont situés à 2000 mètres de profondeur le long des dorsales Atlantique et Pacifique. Le simulateur devait permettre d'explorer la dynamique d'un tel écosystème, aussi bien au niveau biologique que hydrodynamique, et en particulier de modéliser l'influence de l'hydrodynamique générée par les fumeurs noirs – geysers marins desquels l'eau sort à une température de 300°C et une vitesse de 2 $m.s^{-1}$ (voir Fig. A.1) – sur les organismes biologiques. Durant ma thèse, j'ai continué ce projet en étudiant grâce à ce simulateur la dispersion larvaire à une échelle que nous avons nommée "bio-hydrodynamique", à savoir quelques mètres autour du fumeur noir. Un travail de modélisation à cette échelle n'a, à notre connaissance, jamais été effectué, et permettrait de mieux comprendre le fonctionnement de ces écosystèmes très particuliers et très difficiles à étudier.

A.1 Modèle hydrodynamique et développement du simulateur

Je vais résumer ici les principales caractéristiques du simulateur et du modèle hydrodynamique sous-jacent. Tous les détails sont donnés dans l'article, et notamment en ce qui concerne le modèle hydrodynamique dans la section "Matériels supplémentaires" ; je ne les réécrirai donc pas ici.

L'environnement d'un écosystème hydrothermal est complexe, formé de multiples cheminées desquelles partent des jets d'eau chauffés venus de la croûte terrestre. Nous prenons un modèle très simplifié : celui d'une seule cheminée, conique, seule au centre d'un carré de 30 mètres de côté (Fig. A.2). Le nombre de jets modélisés est également réduit, et nous ne considérons que 2 jets, le jet principal émergeant du sommet de la cheminée, rapide et chaud, et un diffuseur secondaire, placé sur le côté de la cheminée,au niveau de la colonie parentale. En effet, les organismes que nous étudions, *Alvinella pompejana*, vivent dans des creux du rocher placés sur les flancs de la cheminée principale ; l'eau qui circule dans ces anfractuosités est chauffée jusqu'à plusieurs dizaines de degrés, donnant un environnement plus favorable que la température ambiante de l'eau à cette profondeur, 2°C, et

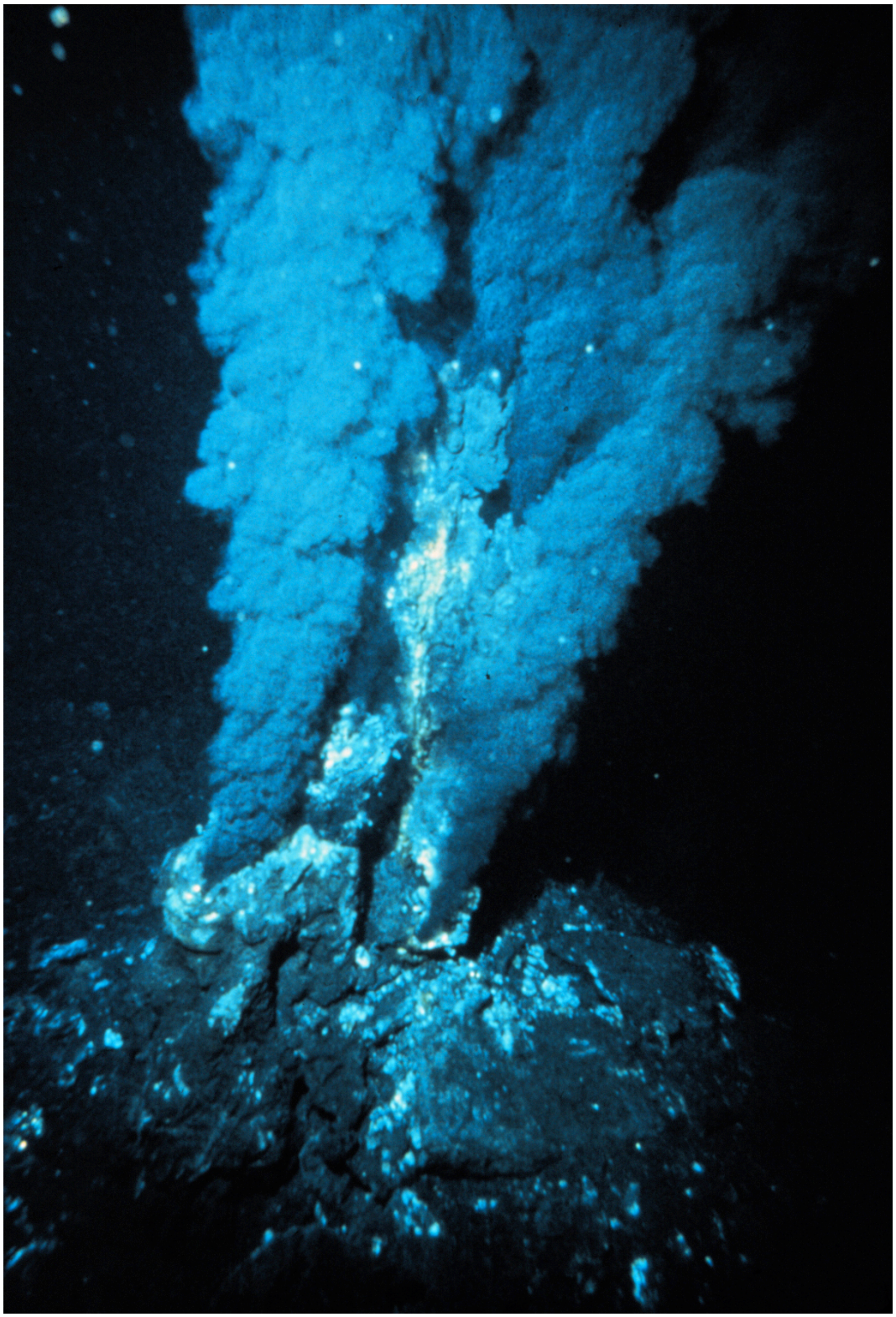


FIG. A.1 – Fumeur noir à deux jets. L'eau s'échappant de la cheminée est à une température de 300°C et à une pression de 200 atm.

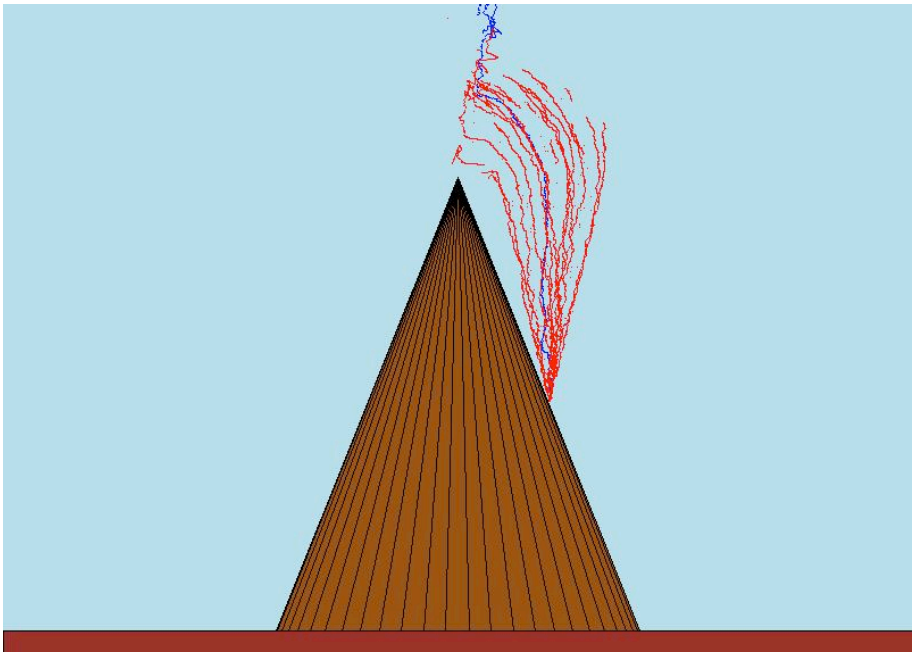


FIG. A.2 – Exemple de la sortie graphique après une simulation. Au centre, la cheminée hydrothermale. En rouge, les trajectoires des larves encore situées dans le cadre de la simulation. En bleu, les trajectoires des larves expulsées loin de la cheminée par le jet principal.

créant des mouvements de fluide. C'est également cet endroit qui sert de point de départ aux larves.

Nous ne considérons que trois sources d'impulsion possibles pour le fluide environnant :

- Le jet primaire, vertical, localisé au sommet de la cheminée centrale.
- Le diffuseur secondaire, placé au niveau de la colonie parentale sur le côté de la cheminée.
- Les courants de fond.

Les courants de fond sont très simplement modélisés comme unidirectionnels et stationnaires à notre échelle de temps. La modélisation du jet principal et celle du diffuseur secondaire sont basées sur la théorie des jets (Batchelor, 1970; Landau and Lifshitz, 1959; Tritton, 1977). Le modèle est le même dans les deux cas, mais la dynamique créée est différente, à cause des valeurs relatives de la température et de la vitesse à l'origine de ces jets. L'idée clef de notre modèle est, pour chaque jet, de déterminer son régime stationnaire grâce aux valeurs de paramètres comme sa température et sa vitesse au niveau de l'orifice. Pour cela, on utilise l'analyse dimensionnelle, et on cherche la force qui domine la mécanique du jet : il peut s'agir de la force inertielle, auquel cas le jet se comporte comme un jet d'eau dans une fontaine, ou de la force convective. Dans ce cas, le mouvement du jet est dominé par la différence de température entre le fluide dans le jet et en dehors : le "jet" est alors une bulle d'eau chaude qui s'élève à cause de sa différence de densité avec le milieu extérieur.

Une fois le régime stationnaire de chaque jet reconstitué, on peut, en utilisant les lois de conservation de la masse, calculer la vitesse moyenne du fluide en tout point. Une composante stochastique est ensuite incorporée dans ce modèle stationnaire sous la

forme d'une vitesse aléatoire, proportionnelle à la vitesse locale en chaque point, sensée représenter de manière simplifiée la turbulence. À partir de cela, le simulateur fonctionne simplement en modélisant le mouvement de chaque larve, en fonction de sa position et des courants qu'elle rencontre. Les larves sont indépendantes et passives, sans motilité propre, et leur mouvement est uniquement dû à la vitesse du courant qu'elles ressentent. Le point de départ des larves est fixé au niveau du diffuseur secondaire, et leurs trajectoires sont enregistrées jusqu'à ce qu'elles sortent du cadre de la simulation ou rencontrent une paroi rocheuse, où elles vont pouvoir coloniser.

A.2 L'article

Voici l'article final exposant les résultats de nos analyses, actuellement soumis à la revue *Journal of Theoretical Biology*. Nous avons testé l'importance de nombreux paramètres, à la fois topographiques, hydrodynamiques et biologiques, sur la dispersion larvaire et en particulier sur le taux de colonisation des larves larguées par la colonie parentale sur le mur de la cheminée hydrothermale. Nos principales conclusions sont :

- Les courants de fond latéraux peuvent dominer la dynamique des larves, et selon leur direction soit les entraîner loin du fumeur originel, soit les rabattre sur sa paroi et favoriser la colonisation.
- Si les courants de fond sont faibles, les larves ont tendance à être entraînées verticalement par absorption dans le jet principal. Le taux de colonisation des larves augmente alors en général avec la pente des parois du fumeur, et avec la vitesse de sortie du fluide chauffé au niveau de la cheminée.
- Contrairement à ce qui avait été supposé, le taux de flottaison des larves – leur vitesse de sédimentation – n'affecte pas le taux de colonisation, à l'exception de cas particuliers aux conditions hydrodynamiques très modérées.

A modeling approach of the influence of local hydrodynamic conditions on larval dispersal at hydrothermal vents

Marc Bailly-Bechet[1], Michel Kerszberg,
Françoise Gaill, Florence Pradillon[2]

UMR CNRS 7138, Systématique, Adaptation et Evolution

Univ. Pierre et Marie Curie, 7 quai Saint Bernard, 75252 Paris Cedex 05, France

(mkersz@ccr.jussieu.fr, francoise.gaill@snv.jussieu.fr, florence.pradillon@snv.jussieu.fr)

[1] Present address: Institut Pasteur, Unité Génétique in silico, 25 rue du Docteur Roux 75724 Paris Cedex 15, France (mbailly@pasteur.fr)

[2] Corresponding author. Tel: +33-1-44-27-35-02. Fax: +33-1-44-27-58-01.

Running title: Larval dispersal at bio-hydrodynamical scale

Key words: Larval dispersal, colonization, hydrothermal vents, agent-based 3D model, bio-hydrodynamical scale

Abstract

Larval dispersal is key in understanding how deep-sea hydrothermal vent communities function and are maintained. To date, numerical approaches developed to simulate larval dispersal have been conducted at ocean ridge scales. However, hydrothermal vents have complex and dynamic local physico-chemical environments. These smaller-scale, but significant variations may influence larval fate in its early stages after release, and hence have a knock-on effect on both dispersal and colonization processes. Here we present a new numerical approach to the study of larval dispersal, considering a "bio-hydrodynamical" scale ranging from a few centimeters to a few meters around hydrothermal sources. We use a physical model for the vent based on jet theory and compute the turbulent velocity field around the smoker. Larvae are considered as passive particles whose trajectories are affected by hydrodynamics, topography of the vent chimney and their biological properties. Our model predicts that bottom currents often dominate all other factors by entraining all larvae away from the vent. When bottom currents are very slow (< 1 mm.s^{-1}), general larvae motion is upwards due to entrainment by the main smoker jet. In this context, smokers with vertical slopes favor retention of larvae. Additionally, larval retention rates increase with velocity of the main smoker jet. This is because entrainment in this high-speed plume is preceded by a phase when larvae are attracted towards the smoker wall, where they can eventually settle. Finally, the buoyancy rate of the larvae, measured to be in the range of 0.01 mm.s^{-1}, is generally irrelevant unless hydrodynamic conditions are in equilibrium, i.e if the buoyancy rate is comparable to both the bottom current speed and the local water velocity due to entrainment by close smokers. Overall, our model evidences the strong effect of the release point of larvae on their future entrainment within local fluxes. Larvae released from smoker walls might have an entirely different fate than those released further away in the water column, which are not, or less, affected by near-chimney hydrodynamics.

1 Introduction

Species inhabiting isolated or unstable environments rely on their dispersal capabilities to colonize new habitats and maintain their populations. Deep-sea hydrothermal vent ecosystems are islands spread along oceanic ridges, with short, decade-long lifespans (Haymon et al., 1993; MacDonald, 1982). Therefore, dispersal and colonization are a critical phase of the life cycle of species endemic to vents (Tyler and Young, 1999, 2003). Since most vent species are sessile as adults, they must disperse predominantly in their larval stages through the water column (Lutz et al., 1984). In order to understand how some of the vent species have persisted through geological time and over such a wide geographical range, dispersal processes were examined through a number of different approaches. *In situ* collections using net tows (Kim et al., 1994; Mullineaux and France, 1995), water pumps (Mullineaux et al., 2005), traps (Khripounoff et al., 2000; Metaxas, 2004) or colonization devices (Berg and Van Dover, 1987; Mullineaux et al., 2003) allowed the collection of larval and post-larval stages of vent species and the estimation of their distribution both through the water column and on the bottom. Previous genetic studies have confirmed migrant fluxes between populations inhabiting distant vent sites (Jollivet et al., 1995; Vrijenhoek, 1997; Young et al., 2003). Development studies have provided information on larval life spans and potential dispersal phase duration (Marsh et al., 2001; Pradillon et al., 2001). Finally, recently published measurements were used to estimate the potential distance over which larvae might travel (Chevaldonné et al., 1997; Kim and Mullineaux, 1998; Mullineaux et al., 2002; Thomson et al., 2003). Larvae might be transported between vents either in bottom currents channeled within the axial valley of the ridge (Kim and Mullineaux, 1998; Thomson et al., 2003), or in currents present at 200 to 300 meters above the ridge crest after they have been entrained to this level by rising buoyant hydrothermal plumes (Kim et al., 1994; Mullineaux and France, 1995). Data gathered from these experiments provided input parameters for computational approaches developed to predict the dispersal potential of vent larvae. Taking into account measured bottom current, observed spatial vent distribution along ridges and known reproductive characteristics, Chevaldonné et al. (1997) and Jollivet et al. (1999) modeled propagule fluxes between vent sites for polychaete species of the Alvinellid family. Dispersal models based on current data and using Lagrangian approaches were used to predict the sorts of distances larvae would be able to travel along ridges (Marsh et al., 2001; Mullineaux et al., 2002). To date all approaches have been conducted at the ridge segments scale, i.e. over tens to hundreds of kilometers, with the exception of the model developed by Kim and Mullineaux (1998), where vertical entrainment of larvae present in the water

column was considered at the vent chimney scale.

Organisms living at hydrothermal vents are exposed to a complex physico-chemical environment due to the mixing between sea water and hydrothermal fluid (Le Bris et al., 2003, 2005; Sarradin et al., 1998). For species living directly on the chimney wall, local fluxes may have a strong influence on larval fate early after their release, while they are still in the vicinity of the smoker, hence affecting dispersal and colonization processes. So far, all modeling approaches assumed that larvae released from vent species were floatting in the water column around smokers, from which point they could be entrained by currents or by the rising smoker plume. However, when released from a smoker wall, a larvae might be trapped by topographic features of the chimney, and would therefore not be able to disperse. Here, we develop a model to qualitatively study how local physical constraints may influence the structure of the vent smoker community through their effects on larval dispersal. We work at and around the smoker chimney scale, i.e metre scale. Larvae in this "bio-hydrodynamic" range can still be considered as being in the vicinity of the smoker. To our knowledge, this is the first attempt at modeling phenomena on this intermediate scale. Our physical model is based on jet theory, and we use numerical methods to describe the hydrodynamic velocity fields around smokers. The passive larvae are entrained by the turbulent fluid, and may be deposited on the mineral surfaces. Using this modeling approach, we aim to identify which factors significantly affect larval fate in the early stages after release. It is what happens during this phase that can influence larger scale dispersal processes or colonization patterns. Factors tested here include hydrodynamic features such as smoker jet velocities and temperatures, smoker topography, and larval characteristics (e. g. buoyancy rates).

2 Computational framework and biological hypotheses

The program is a 3 dimensional agent-based simulator of a hydrothermal vent ecosystem. The numerical simulation models the hydrothermal ecosystem as a 30 m sided cube, centered around a vent chimney (Fig. 1). Two sources of hydrothermal fluid are modeled as jets on this vent chimney. The main one is emitted vertically from the apex of the vent edifice. The secondary source, also called secondary diffuser, is located on the chimney side, where the adult population grows. The flow is there emitted perpendicularly to the wall.

The aim of the simulations is to model the possible trajectories of the released larvae entrained by the heated water emitted by the secondary diffuser.

Each simulation represents one day, divided in 172800 steps of 0.5 s, a delay long enough for our model with simplified turbulence to be reasonable, while suficiently short to allow the detailed simulation of larval trajectories. Each simulation starts by the creation of 1000 larvae, with no initial speed, on the central axis of the secondary diffuser, located on the chimney side. This corresponds to the release of offspring by the adults living there. Then their trajectories at each time step are computed and recorded, according to the hydrodynamic properties of the water around them. At every time step, the movement of each larva is computed independently. Larvae are considered as inert particles entrained by the surrounding currents. Therefore, the speed of each larva at any time step is equal to the velocity of the fluid at the same place and time, plus the buoyancy of the larva. Experimental studies conducted on the development of two vent species, *Riftia pachyptila* and *Alvinella pompejana*, showed that early embryos do not have any locomotion structures (Marsh et al., 2001; Pradillon et al., 2001). Therefore, since we focus on the early events, in the hours following embryos release, it is realistic to model them as inert particles without proper motility. All larvae being independent, the choice of 1000 larvae was done according to available comupter time. To get better statistics, some simulations were repeated 3 to 5 times.

2.1 Physical environment

2.1.1 Mineral environment

The mineral environment is composed of two parts: the oceanic floor, which is defined as a flat surface, and the chimney, which is modeled as a cone defined by the coordinates of its summit $\vec{x}_s$, its base center $\vec{x}_b$ and its base ray R_c. In all simulations there is one chimney located at the center of the oceanic floor. The slope α of the chimney is the angle between the oceanic floor and a generatrix; we have $\alpha = \arctan\left(\frac{\|\vec{x}_s - \vec{x}_b\|}{R_c}\right)$.

Chimney surface irregularities might significantly enhance larval settlement when the larvae are driven close to the chimney surface by turbulent fluids . For this reason, irregularities, like small rock outcrops, are modeled on the chimney. To keep the number of parameters small, only one parameter is used to simulate surface irregularities: the maximal texture depth R, representing the maximal height of the irregularities on the surfaces, also called surface roughness. The distribution of the size of irregularities is represented in a probabilistic way to simplify the simulations.

More details about surface roughness are given in the section 2.2.1.

2.1.2 Hydrodynamic environment

In our simulations, fluid flows originate from 3 different sources:

1. The primary jet, corresponding to the main hydrothermal fluid emission, located on top of the chimney.

2. The secondary jet, with its outlet on the side of the chimney, located amongst adult populations. This secondary jet represents the diffusion of diluted hot hydrothermal fluids observed among animal communities. The velocity and temperature at the outlet of this diffuser are much lower than those observed for the primary jet.

3. The bottom currents, with constant direction at our timescale (see below).

At any given point $\vec{r}$, the fluid velocity is the sum of two parts: a deterministic component, the *entrainment* velocity v_e, and a random one, the *turbulence* velocity v_t.

The entrainment velocity $\vec{v_e}(\vec{r})$ is constant in time during one simulation. It depends on the position $\vec{r}$ of the point relative to the jets, and on the parameters of the simulation. The entrainment velocity is defined to be :

$$\vec{v_e}(\vec{r}) = \vec{v_1}(\vec{r}) + \vec{v_2}(\vec{r}) + \vec{v_b} \tag{1}$$

where $\vec{v_j}(\vec{r})$ is the velocity caused by jet j at position $\vec{r}$ ($j = 1$ for the primary jet, $j = 2$ for the secondary diffuser), and $\vec{v_b}$ is the constant velocity of the bottom currents, considered as unidirectional and parallel to the sea floor. This hypothesis is based on the observation that the highest frequency recorded for reversal of bottom currents directions is semidiurnal (i.e. about 12h, Kim and Mullineaux (1998); Thomson et al. (2003)).

In situ observations indicate that for of active smokers, constant velocity of fluid flows at the smoker output are found over periods of a few hours, with relatively low variations. Here we simulate larval trajectories over 1 day. However, in most cases, the larval trajectories within the simulation field are determined in a few hours only. Therefore, it is here reasonable here to assume that both jets and bottom currents velocities are constant during each simulation.

Summing the velocities generated by the two jets at point $\vec{r}$ may seem simplistic as one would expect interactions between the two flows. However, as these interactions effects depends on both speeds, we can suppose that they are relevant (when

compared to the higher speed flow) only when $\vec{v}_1(\vec{r})$ and $\vec{v}_2(\vec{r})$ are both of a suficient magnitude. This never happens in our simulations, knowing that the velocity of the secondary jet is usually much lower than that of the main jet. At any given point, either one jet is much stronger than the other, or both have a very low velocity, when compared for example to the bottom current speed. Therefore, the two jets dynamics are considered independently, and velocities from both jets are added up.

A jet is parametrized by: the ray of its outlet (supposed circular) R_0, the typical velocity of the fluid $\vec{U}_0$, and the difference in temperature with the external media ΔT_0, both taken at the outlet. The temperature of the surrounding media is constant and equal to 2°C, which is the temperature of abyssal sea-water at the depth at which vents occur. The jets are supposed turbulent and expanding in a conical shape from the outlet, with an increasing ray and a constant slope θ_j (half angle of aperture). Details about the jet geometry and the conical shape assumption are given in the Supplementary materials. The values of θ_j for both jets are the same in all simulations. We use the computation of the typical scale of variation of the gaussian profile of speed in infinite jets by Morton et al. (1956) to infer the value of θ_j, which is set to $\theta_j = 10°$.

The way the velocity caused by jet j at point $\vec{r}$ is computed is detailed in the Supplementary Materials. Here is a summary of the hypothesis made and the formulae found for $\vec{v}_j(\vec{r})$, valid for both the main and the secondary jets. For simplicity we adopt the notation $[\vec{k}] = \frac{\vec{k}}{\|\vec{k}\|}$ for unitary vectors. We note $\vec{x}_j$ the origin of the jet. The jet is modeled as a cone of moving fluid expanding from $\vec{x}_j$. This origin is located below the jet outlet and inside the chimney, at a distance such that the ray of the cone is R_0 at the level of the outlet. As the jets are supposed conical, the ray of the jet, at a vertical distance l from $\vec{x}_j$, is simply $R(l) = l\tan(\theta_j)$.

Jet dynamics can be dominated either by inertial forces (inertial jets), which behave like an ooze, or convective forces (convective jets), which behave like a bubble of heated water. The dominant force is estimated at one given point $\vec{r}$ inside the jet by comparing the overall Richardson number $Ri = \frac{g\alpha\Delta T(\vec{r})l}{U^2(\vec{r})}$ to 1, with α being the thermal expansivity of water, g the intensity of gravity, l the distance from the outlet along the jet axis, $\Delta T(\vec{r})$ the difference in temperature between inside and outside the jet at point $\vec{r}$, and $\vec{U}(\vec{r})$ the water speed at point $\vec{r}$. $Ri < 1$ means inertial forces are dominant, $Ri > 1$ that convective forces are dominant.

The speed generated by the jet j at point $\vec{r}$ is computed as follows : if $\vec{r}$ is inside the jet itself (that is, $[\vec{r} - \vec{x}_j] \cdot [\vec{U}_0] < \cos(\theta_j)$), for an inertial flow, we have :

$$\vec{v}_j(\vec{r}) = \frac{U_0}{R_0} \frac{1}{\|\vec{r} - \vec{x}_j\|} [\vec{r} - \vec{x}_j] \tag{2}$$

and, for a convective flow, we have :

$$\vec{v}_j(\vec{r}) = \frac{U_0}{R_0} \left(\frac{1}{\|\vec{r} - \vec{x}_j\|} \right)^{1/3} [\vec{r} - \vec{x}_j] \tag{3}$$

The same type of relationship is derived for temperature variation inside each type of jet (see Supplementary materials).

This formulae imply that Ri increases with l for inertial jets: the dominant force changes and inertial jets tend to become more and more convective with increasing l. For a convective jet $Ri(l)$ is constant and the jet remains convective. If a jet has inertial dynamics at its outlet, the critical distance d_c at which it qualitatively changes to become convective is :

$$d_c = \sqrt{\frac{R_0 U_0^2}{g\alpha\Delta T_0}} \tag{4}$$

Jets with two different dynamics consecutive in space are numerically modeled as two different but continuous jets, the convective one above the inertial. For $l \leq d_c$, the equation (2) for inertial jets are used directly. For $l > d_c$, the formula (3) for a convective jet is used, with R_0 replaced by the ray of the jet at $l = d_c$ (which gives $R_0 = d_c \tan(\theta_j)$). U_0 is replaced by the velocity at the same point, which is where the inertial and convective part of the jet merge. Note that the axis of the convective part of the jet is vertical, even if the inertial part had an inclined axis (see Fig. 1a).

Such a definition of the speed field implies a discontinuity in velocity between the inside and the outside of the jet. At the bundary, the flow is turbulent and creates eddies entraining the surrounding water inside the jet. This increases its cross-section flow with distance l to the outlet. We model this absorption by a water speed field, orthogonal to the jet cone sides, all around the jet. We compute the speed field far from the jet using water incompressibility. The speed at a lateral distance r_{lat} from the jet is calculated such that the total flow directed inwards, in the sense of $(\vec{r}_{p,j} - \vec{r})$ on Fig. 1b, is the same as the increase in flow inside the jet cone at the same height, computed using the jet properties. This spatial part of the vent surrounding, where the flow is directed towards the main jet, is later refered to as the zone of absorption.

One simplifying assumption made is that the water speed field around the jet cone is orthogonal to its boundary surface, and has the same circular symmetry as the speed field inside the cone (Fig. 1b). Because of this assumption, there is still a discontinuity in the speed field at the cone boundary. However, this occurs due to the absence of a detailed modeling of the turbulence. It could be a limitation of the model for studying the particular dynamics of entrainment of water inside the thermal plumes. Usually larvae arriving at this point are most likely to be entrained by the main jet thermal plume far from any mineral surface. Hence, the precise modelling of the turbulence at the jet boundaries being of no consequence on the larval settlement, it was not aborded here.

We calculate the speed of any point $\vec{r}$ inside the zone of absorption, relative to the speed of its orthogonal projection on the side of the conical jet. We call this point $\vec{r}_{p,j}$ (Fig. 1b). We note $l = \|\vec{r}_{p,j} - \vec{x}_j\|$, the distance to the jet start (which is almost equal to the distance to the jet outlet) and r_{lat} the distance between $\vec{r}$ and $\vec{r}_{p,j}$. We have:

$$\vec{v}_j(\vec{r}) \simeq \frac{U_0 R_0}{2\left(r_{lat} + R(l)\right)} \tan^2(\theta_j)[\vec{r} - \vec{r}_{p,j}] \tag{5}$$

at point $\vec{r}$, if $\vec{r}_{p,j}$ is on the border of the inertial part of a jet. The dependence of $\vec{v}_j(\vec{r})$ with l is there hidden in the $R(l)$ term.

If $\vec{r}_{p,j}$ is on the border of a convective jet, we have:

$$\vec{v}_j(\vec{r}) \simeq \frac{5}{6} \tan^2(\theta) U_0 \frac{R_0^{1/3} l^{2/3}}{\left(r_{lat} + R(l)\right)} [\vec{r} - \vec{r}_{p,j}] \tag{6}$$

Both these speeds are computed using standard jet properties. Detailed calculations can be found in the Supplementary materials.

Finally, outside of the absorption zone, we consider that the water movement generated by the jet j is only due to viscous entrainment of the water layer in contact with the absorption zone. The location where this phenomenon applies is called the *entrainment zone* on Fig. 1b. Taking now $\vec{r}$ to be in this zone, we compute the speed at point $\vec{r}$ relative to the speed at $\vec{r}_{p,a}$, defined to be the orthogonal projection of $\vec{r}$ on the absorption and entrainment zone limit (see Fig. 1b). The entrainment has to decrease with distance to the absorption zone $\|\vec{r} - \vec{r}_{p,a}\|$. In order to minimize the number of parameters, and to prevent estimation of the decrease in water speed due to viscosity, we arbitrarily chose:

$$\vec{v}_j(\vec{r}) = \frac{1}{1 + 5\|\vec{r} - \vec{r}_{p,a}\|} [\vec{v}(\vec{r}_{p,a})] \tag{7}$$

Tests have been run replacing the factor 5 by 1 with no difference on the results. $\vec{v}(\vec{r}_{p,a})$ is computed using (5) or (6), depending respectively if the corresponding jet is inertial or convective at its outlet.

Velocity fields are obviously not generated across a mineral structure. This is particularly important for the secondary jet, located on one side of the chimney, which only has one-sided influence.

The speed generated by jet j at point $\vec{r}$ is estimated in a deterministic way. As stated before, turbulence is added to prevent the model being entirely deterministic, which would give rise to unrealistic dynamics, with all larvae having the same trajectories due to the same entrainment speed. Our way of modelling turbulence, while simple, gives us the essential features needed for the model. The turbulence speed is proportional to the entrainment speed. It models the random movement of a particle in turbulent eddies. The turbulent speed is defined as :

$$\vec{v_t}(\vec{r}) = k\|\vec{v_e}(\vec{r})\|\,[\vec{u}] \tag{8}$$

where k is the turbulence coefficient, and $[\vec{u}]$ a random unitary vector. k is higher inside than outside the jets. Values increasing from 0.5 to 2 outside and 3 to 12 inside were tested to assess the importance of turbulence. As expected, colonization rates are increasing with greater turbulence (data not shown). In order to analyse the effect of other parameters that could remain unseen with high turbulence, the turbulence coefficient was set to a low constant value: 0.5 outside the jets and 3 inside, in the presented simulations.

A new random unitary vector $\vec{u}$ is computed every time a new speed is considered, so the turbulence is a white noise proportional to the entrainment speed at each point $\vec{r}$. As a consequence, there are no correlations between two successive computations of $\vec{u}$: at our scale the details of turbulent structures are completely neglected, even inside the jets.

2.2 Larval characteristics

Observation of eggs and early embryos showed that their buoyancy rate –defined as "the sinking speed of the larvae due to their density"– differs according to species (Cary et al., 1989; Marsh et al., 2001; Pradillon et al., 2004), and such characteristics have been put forward to explain different dispersal capabilities. We take into account the buoyancy rate $\vec{b}$ of larvae by adding it to the final speed $\vec{v_l}$ of the larvae:

$$\vec{v_l} = \max(\vec{v}(\vec{r}), \epsilon) + \vec{b} \tag{9}$$

Here the final speed of a larva at point $\vec{r}$ is noted $\vec{v_l}(\vec{r})$, and $\vec{v}(\vec{r})$ is the final hydrodynamic speed, taking into account the entrainment speed, the turbulence and the bottom currents speed. Buoyancy $\vec{b}$ can be directed downwards (sinking larvae) or upwards (buoyant larvae).

As the fluid speed $\vec{v}(\vec{r})$ can approach 0 far from both jets, and in the absence of bottom currents, a minimal random speed of larvae was set up to prevent the unrealistic case of static larvae. If $\|\vec{v}(\vec{r})\| < \epsilon$, the jet-induced part of larval speed is replaced by a minimum cut-off speed of random orientation and norm $\epsilon = 0.1$ mm.s^{-1}. The choice of the value of ϵ is such that its value is low when compared to any typical speed of a jet.

2.2.1 Larval settlement

If the displacement of a larva (computed as $\vec{v_l}.dt = 0.5\vec{v_l}$) ends at a distance d smaller than the maximum texture depth R of a mineral surface (defined in section 2.1.1), the larva is in a position to settle. Here "settling larvae" is not used in its proper sense since at our time scales, released larvae are at most early embryos, not yet competent, and thus not ready to settle. Settling is here related to a trapping effect. As a first order approximation, we suppose that the distribution of the irregularities sizes is random, and the probability of settling P_s is:

$$P_s = p_{col}(1 - \frac{d}{R}) \tag{10}$$

where p_{col} is the maximum probability of settling, the probability that it will settle if it makes contact with the chimney itself. Due to the random part of their movement induced by turbulence, larvae close to mineral surfaces tend to hit them multiple times (Berg, 1993) and colonize them even if p_{col} is low. The value of p_{col} is set in all tests at 0.5.

It should be noted that, this part of the model being probabilistic, a larva can settle at a given distance d from the mineral surface, while larvae subsequently passing by the same point or even closer to the mineral surface will not settle. This is contrary to what would be expected with a simple model of automatic settling upon contact with the surface. In this way we model an average roughness of the mineral surface with only one parameter, R, with no need to calculate multiple configurations to take irregularities into account.

2.2.2 Larval mortality

Finally, experimental studies also demonstrated sensitivity of embryos to temperature, with temperatures over 20°C being lethal in *A. pompejana* (Pradillon, 2002; Pradillon et al., 2005). In our simulation, fluid temperature outside the jets is set to 2°C, neglecting the heat transmitted by conduction from the rocks or from the jet to the ocean water. Inside the jet, temperature is modeled like speed as a decreasing function of the distance to the outlet (see Supplementary Materials).

A probabilistic model of larval death due to temperature is included in the simulation. At each time step, at the end of the larval movement, the probability for a larva located at point $\vec{r}$ to die is defined as :

$$P_{death} = 0.0125T(\vec{r}) - 0.25 \tag{11}$$

with $T(\vec{r})$ being the water temperature at point $\vec{r}$. This formula makes P_{death} increase linearly from 0 at $T = 20$°C to 1 at $T = 100$°C, allowing the biologically safe assumption that the shortest possible exposition to temperatures greater or equal than 100° C is lethal. At our timescale of a few hours, temperatures below 20° C are considered harmless. No mortality due to predation is assumed in this version of the model.

2.3 Simulations

2.3.1 Simulation parameters

There are four possible outcomes for each larva at the end of a simulation:

1. Settled on a mineral surface.
2. Dead (due to high temperature exposure).
3. Entrained by the fluid outside of the simulation field.
4. Still in the simulation field, but not in any of the first three situations.

The forth case is due to the time limit put on the simulations. However, in all simulation, more than 95% of the larvae are in one of the 3 first cases after 24h.

In the simulations, some of the parameters are set to fixed values, chosen from average values reported in the literature. We checked that variations in the values of these parameters have no significant influence on the outcomes of simulations. The half opening angle θ_j of each of the two jets is thus set to 10°. The radius R_0

of the opening of the main smoker and secondary fluid emission are set to 0.05 m and 0.01 m respectively. All plotted results (in percentages) represent one of the categories over all 4. The ranges of values tested in the simulations for the different parameters cover the real values reported in the litterature, and are shown in Table 1.

2.3.2 Real cases

Simulations were done to predict the larval trajectories of two species living on a smoker at the Genesis site, 13°N East Pacific Rise, described by Sarradin et al. (1998). This is a 9 metres high smoker, with a near vertical slope, which expulses hydrothermal fluid at 270°C. The top 2 metres of the chimney are covered with alvinellid worms including *A. pompejana*. *R. pachyptila* tubeworms occupy the base of the chimney wall.

In the case of the *A. pompejana* larvae simulations, we set larval release point 2 metres below the main jet. Temperature in the secondary jet is set at 40°C, which represents an average value from measurements taken in the alvinellid colony (7-91°C (Le Bris et al., 2003; Sarradin et al., 1998)). Velocity of the secondary jet is set to 1 cm.s^{-1} according to average values recorded in diffusion areas. Negative larval buoyancy rates of -0,03 mm.s^{-1} are applied (Pradillon et al., 2004).

In the case of *R. pachyptila*, we set the larval release point at the base of the smoker. Temperature in the secondary jet is set to 10°C, which represents an average value from temperatures recorded in the *Riftia* tubeworms (5-15°C, Le Bris et al. (2003); Sarradin et al. (1998)). Velocity of the secondary jet is set at 0.5 cm.s^{-1}. Positive larval buoyancy rates of +0,03 mm.s^{-1} are applied (Marsh et al., 2001). No experimental data being available, we set the chimney wall texture depth to 1 cm.

3 Results

The main objective of our model is to identify which parameters are significant with respect to larval trajectories in the vicinity of the smoker from which they were released, within a few hours after spawning. Our simulation field consists of a 30 metre sided cube, in the center of which stands an active vent chimney. The larvae are emerging from a secondary diffuser located on the chimney side. We tested different parameters including: velocity and temperature of the ejected fluids from the main jet or secondary diffuser; velocity of bottom currents; slope of the chimney; effect of relief features on chimney walls; position of the release point of larvae and

larval buoyancy. The outcome of each simulation gives the proportion of larvae which: settled within the simulation field, settled out of the simulation field (i.e. considered as dispersing); unsettled but still were in the simulation field; and died. Trajectories and larvae settling points are also given by the simulation, and allow us to identify which forces have the strongest influence.

Large scale models of larval dispersal usually assume that larvae are carried up in the water column, entrained by rising warm fluids and by absorption of the surrounding seawater into the vent plume. However, depending on local conditions, a significant proportion of the released larvae might not be entrained and rather settle on the site of origin. Here we show in details which conditions can give rise to an increased or decreased larval settlement.

3.1 Topography

At the scale of centimetres to metres around a vent structure, topographic features might stand within the larvae trajectories. These features are expected to trap them, preventing their attraction within the rising plume and their large scale dispersal. The maximal height of the rock outcrops at the surface of the chimney wall is defined as the maximal texture depth R. Larvae are released in the secondary diffuser, at a distance λR from the wall. We can distinguish two cases in the simulations:

- For $\lambda \leq 1$, i.e. when the larvae are released closer to the wall surface than the maximal height of the rock relief, the percentage of settling larvae increases with R and diminishes with λ (Fig. 2). This diminution is due to the lower number of larvae caught just after being released, while they are still close to the smoker wall, with higher values of λ.

- For values of $\lambda > 1$, when larvae are released beyond all relief features, the settling rate diminishes with R and λ (Fig. 2). In this case, the diminution relative to R can be explained if we hypothesize that most of the colonization takes place close to the release point. With higher values of R (and $\lambda > 1$), larvae are released further away from the chimney wall, resulting in fewer opportunities to settle.

Since we consider embryos of vent organisms inhabiting chimney walls, the larvae are most probably released within a distance where they might encounter relief features. However, a too low value of λ would increase the settling close to the release point so much so that no other phenomena could be studied. So we decided to set λ to 1 in further simulations. For the same reason, the maximal texture

depth R is set at a low value ($R = 0.001$ m), in order to allow more larvae to leave the chimney wall after being released, and evaluate more precisely the effect of hydrodynamics. In real cases, the values of λ and R are not known exactly, but our results show that the exact position of the larval release point relative to the chimney wall surface irregularities can be of great importance to local larval colonization around the adult patch.

We next test the importance of the chimney shape on colonization. To assess the importance of the slope of the smoker wall, two types of chimneys representing different topographies are modelled: a small flat edifice 5 metres high with a wall slope of 45°, and a tall vertical edifice 15 metres high with a slope of 70° from the horizontal. All other parameters being equal, the simulations show a higher proportion of larvae settling on the wall of the tall vertical smoker than on the flat one (compare Fig. 2a and 2b, and Fig. 3a and 3b). This trend does not depend on the release point of the larvae (Fig. 3). One tentative explanation is that, the more vertical a smoker, the longer the larvae remain close to the chimney wall during their vertical ascent in the heated water coming from the secondary diffuser, and hence the more possibilities they have of colonization. On a flat edifice, distance between larval trajectories and smoker wall increases immediately, thus limiting the number of settling events along the chimney wall.

3.2 Hydrodynamics

As larvae are considered as passive particles, they are entrained by local currents. Their entrainment within the smoker plume depends on the main jet regime, which drives all the ecosytem hydrodynamics. Interestingly, the proportion of settling larvae increases with the velocity of the main jet (Fig. 3). Temperature of the main jet fluids does not affect this pattern (daya not shown). Indeed, if the water in this jet increases in velocity, more water will be entrained laterally inside it due to local turbulence. This increase in the entrainment of the surrounding waters propagates to strengthen the global water flux directed towards the jet and the chimney. Therefore, the faster the main jet velocity is, the earlier the larvae are attracted towards the smoker during their ascent. With a high speed main jet, they are attracted while their vertical level is still below the top of the edifice. This allows them to make contact with the smoker wall and settle before being entrained inside the rising plume of the main smoker.

The relative importance of this phenomenon is modulated by the distance between the main jet and the larval release point. When this distance increases, the proportion of settling larvae on smokers with a high speed main jet falls to the values

obtained for the low speed ones (Fig. 3). In those cases, the larvae are too far to be entrained early towards the chimney wall, and there is no colonization close to the main jet outlet on the wall. Supporting this hypothesis, we observe two main regimes instead of a continuous increase of the colonization rates with the main jet velocity : the settlement rates remain low for main jet velocities up to 10 cm.s^{-1}, and abruptly increase for higher velocities (Fig. 3). Such dual regime also occurs when the larvae are released further down on the smoker wall, but the transition between the two regimes occurs for higher main jet velocities. When the smoker is steeper sided, the difference between the two regimes diminishes (Fig 3a).

The bottom currents strongly affect the settling rates. With a lateral bottom current entraining the larvae towards the wall (i.e. for colonies located on the side of the chimney facing to the bottom current), the colonization rate is always 100% (data not shown). If orientated in the opposite direction, the bottom current can be a dominant force, in competition even with the strongest main jets (Fig. 4). In all tested configurations with typical bottom current velocities between 2 and 20 cm.s^{-1} (Cannon et al., 1991; Kim and Mullineaux, 1998), the bottom currents dominate over the attraction by the plume, and export the larvae laterally out of the smoker surroundings.

3.3 Biology

High temperatures can be deleterious to larvae (Marsh et al., 2001; Pradillon et al., 2005, 2001). In our simulations, we tested the effect of exposure to high temperature. Mortality rate is proportional to both the temperature and the duration of exposure. First, we remarked that mortality due to high temperature in the main jet is almost never observed (data not shown). Considering larval trajectories, we assumed that this was due to the fact that most larvae enter the plume at a distance from the jet outlet where temperature has already dropped. However, we observed up to 40% mortality in released larvae due to the temperature in hot secondary jets (Fig. 5). The velocity of the secondary jet plays a key role this case. A faster jet expulses the larvae quickly, and the exposure time to high temperature is reduced, leading to lower mortality. On the contrary, low velocity jets leave the larvae exposed to lethal conditions for longer, thus increasing mortality rates (Fig. 5).

The buoyancy rates observed for vent embryos are in the range of tens of micrometres per second (Marsh et al., 2001; Pradillon et al., 2005, 2001), i.e. very low compared to currents and jet velocities which are typically in the range of centimetres to metres per second. In our simulations, larval buoyancy does not influence larval dispersal or colonization rates when these typical values are considered. When bot-

tom currents oriented so as to export larvae are taken into account, settling rates become low whatever the hydrodynamic properties of the larvae. All are equally entrained outside of the smoker surroundings. In another type of dynamic environment, with a high velocity main jet, all larvae are attracted and pushed towards the smoker walls, thus leading to very high settling rates, whatever larval buoyancy (Fig. 6b). However, in areas with low jet velocities, buoyancy rates might significantly influence colonization (Fig. 6). Without bottom currents and main jet velocity as low as 0.1 $m.s^{-1}$, negative buoyancy rates – within the range of those observed for vent species – significantly increase the proportion of colonists, whereas positive buoyancy decreases the colonization rates (Fig. 6a).

3.4 Simulations of real cases

In addition to the general prediction of the parameters dominantly affecting larval trajectories in the vicinity of smokers, our model can be used to simulate larval fate in real environmental cases. Such a tool might be helpful for conducting *in situ* experiments, such as choosing the deployment site of larval colonization experiments according to predicted larval trajectories. For example, in Fig. 7, we present the trajectories of larvae released from a smoker from the Genesis site at 13°N on the East Pacific Rise. The smoker (PP HOT2) was described by Sarradin et al. (1998). The top two metres of the chimney are covered with alvinellid worms including *A. pompejana*. *R. pachyptila* tubeworms occupy the base of the chimney wall.

A. pompejana and *R. pachyptila* are two emblematic annelid species of East Pacific Rise vent sites. They usually colonize different parts of the vent ecosystem and are believed to have different larval strategies. By simulating the trajectories of larvae released from the populations of these two species, we show different fate, which might be partly responsible for the different dispersal capabilities of these species. During our simulations almost all *A. pompejana* larvae are retained on the smoker. This pattern is due to a combination of characteristics which favour attraction towards the smoker wall including: proximity of the adult colony to a high speed main jet, vertical slope of the smoker, and negative buoyancy of the larvae. All conditions push the larvae towards the chimney and prevent them from being expelled from the system. However, the situation might be completely different when the smoker is exposed to bottom currents (Fig. 7c), in which case the larvae can be exported out of the ecosystem.

In the case of *R. pachyptila*, the proportion of retained larvae is much lower (Fig. 7d, e, f). Almost half of the larvae are entrained in the plume of the main jet (Fig. 7d), and this proportion increases to almost 100% with high main jet velocities of 1

$m.s^{-1}$ (Fig. 7e). When the bottom currents are simulated, all larvae are entrained laterally rather than vertically if they are released on the hidden face of the smoker (Fig. 7f), or settle if they are on the wall facing the current, as before.

4 Discussion

The model presented here is the first attempt to analyse larval dispersal at a local scale, i.e. within centimetres to metres from a vent chimney. Simulations allow us to identify the main parameters that influence dispersal and settling rates of vent larvae just after their release. We show that bottom currents are the dominant phenomenon, driving on site colonization or dispersal depending on their direction relative to the colony's position on the smoker. When bottom currents are slow, the geometry of the chimney wall, particularly its slope, influences mostly colonization; vertical smokers have higher colonization rates. As expected, irregularities on the chimney wall may catch larvae before they are entrained by local hydrodynamics and thus increase colonization. High velocity of the main jet also gives rise to increased colonization rates due to the wide zone of absorption it generates. The influence of the main smoker on the larval fate is smoothed by increasing distance of the larval release point from the main jet. Finally, only when all hydrodynamic conditions are in a lower range of values, can the buoyancy rates affect the fate of larvae.

Previous approaches by Kim et al. (1994) have shown that the hydrothermal plume would concentrate larvae and drive their dispersal. In their model, larvae are considered to be initially in the water surrounding the chimney. We suggest that vent larvae might be affected differently by vent hydrodynamics. Due to their initial location close to vent chimney walls, their trajectories might be stopped by relief and topography features. These features might catch released larvae and prevent their entrainment within the rising plume. Indeed, in simulations where larval starting point was set far from the chimney wall, larvae were mainly entrained within the plume. This agrees with the results of Kim et al. (1994) (Fig. 2, $\lambda = 20$). However, when larvae are released from the chimney wall, larval fate differs. By simulating vent larvae behaviour at a bio-hydrodynamic scale, our model indicates that due to their specific release point vent larvae should not be considered equivalently with other abyssal larvae, as initial trajectories within the vent surroundings are specifically affected by topography and local hydrodynamics.

Bottoms currents can transport larvae between vents (Kim and Mullineaux, 1998; Mullineaux and France, 1995). In our simulations, bottom currents appear as the major force driving larvae outside of the vent vicinity. From the larval point of

view, bottom currents, whose velocities are typically in the range of some cm.s^{-1} (Cannon et al., 1991; Kim and Mullineaux, 1998), are more powerful than the local convection fluxes and plume absorption. In the presence of bottom currents, vent larvae should then be expected either to disperse within a few metres above the sea floor – rather than at the level of the spreading buoyant plume – or to colonize the chimney wall facing the currents. Indeed, field observations are in agreement with this pattern. Higher abundances of vent larvae have been detected a few metres above the bottom along the ridge axial valley than at the level of the spreading of vent plumes, 200-300 meters above the bottom (Mullineaux et al., 2005).

When bottom currents are negligible compared to the smoker main jet, the larvae are always entrained vertically higher in the system. This is a logical consequence of the hypothesis that all water fluxes are created by jets having an upwards component. Due to absorption of the surrounding fluid by the main jet, the larvae tend to be entrained vertically inside the rising plume. Such trajectories make the slope of the chimney walls important in two ways. First, at the expulsion of larvae, a chimney with a vertical slope will offer settling possibilities for an extended period of time, since larval ascent is parallel to the wall. This will not happen with a flat edifice. For the larvae that do not settle early but rise in the water column, the other possibility of settling is just before their absorption into the main jet, if they are close to the mineral structure again. In this case, vertical surfaces at the top of the chimney will offer more colonization possibilities, as the larvae entrained towards the plume from beneath will make contact there before getting inside the plume. These two phases are occurring independently, which might locally modulate the colonization rates. In natural smokers, the slope may vary at different heights on the same edifice, and increased colonization can occur separately at the expulsion point or near to the top. Such colonization configurations were observed with *Alvinellid* colonies (Ex Elsa PPH1 EPR 13°N, personal observation by FP).

The absorption effect increases with the velocity of the main jet, but decrease with the distance from the plume. Therefore, depending on their release position, and on the main jet speed, the larvae will be entrained more or less early towards the plume, leading respectively to settling on the top of the smoker or entrainment inside the plume.

The buoyancy rates of embryos and larvae have been put forward to explain different dispersal strategies in different vent species. For example, eggs and early embryos of *Alvinella pompejana*, are negatively buoyant (Pradillon et al., 2004), leading to the hypothesis that larvae of this species would sink to the base of chimney and develop there (Chevaldonné and Jollivet, 1993; Pradillon et al., 2005) and have limited dispersal (Chevaldonné et al., 1997). Conversely, *Riftia pachyptila* eggs and

embryos are positively buoyant (Marsh et al., 2001), leading to the hypothesis of large dispersal. However, our model shows that strong hydrodynamic motion close to the hydrothermal edifices make these tendencies negligible.

By simulating larval trajectories of two vent species, *A. pompejana* and *R. pachyptila*, in the vicinity of a real smoker, we showed that the position of adult populations influences the type of dispersal larvae might have. With a vertical configuration of the smoker wall, *A. pompejana* embryos would be quite easily trapped and would not disperse. Indeed, young active smokers typically have a vertical elongated shape due to the rapid mineral accretion (FP personal observation). In such cases, dispersal of embryos emitted by populations growing on the smoker walls would be limited. On the contrary, less active edifices grow much more slowly and tend to have a more flat or round shape, sometines refered to as "snowball". In that case, more dispersal would occur. This would therefore favor migration towards new sites when activity is declining. *R. pachyptila* occurs much more rarely on smoker walls, and would therefore be almost always in a configuration where larvae would be exported outside of the vent.

The results obtained here allow us to evaluate classical dispersal scenarios and to formulate new hypotheses based on the parameters significantly influencing colonization. In the future, the model will offer possibilities to test other processes at the bio-hydrodynamic scale, including: reproduction, predation, species interactions, and temporal dynamics of colonization.

References

Berg, C.J., Van Dover, C.L., 1987. Benthopelagic macrozooplankton communities at and near deep-sea hydrothermal vents in the eastern Pacific Ocean and the gulf of California. Deep-Sea Res. 34, 379–401.

Berg, H.C., 1993. Random Walks in biology. Princeton University Press.

Cannon, G.A., Pashinski, D.J., Lemon, M., 1991. Middepth flow near hydrothermal venting sites on the southern Juan de Fuca Ridge. J. Geophys. Res. 96, 12815–12831.

Cary, C.S., Felbeck, H., Holland, N.D., 1989. Observations on the reproductive biology of the hydrothermal vent tubeworm *riftia pachyptila*. Mar. Ecol. Prog. Ser. 52, 89–94.

Chevaldonné, P., Jollivet, D., 1993. Videoscopic study of deep-sea hydrothermal vent alvinellid polychaete populations : biomass estimation and behaviour. Mar. Ecol. Prog. Ser. 95, 251–262.

Chevaldonné, P., Jollivet, D., Vangriesheim, A., Desbruyères, D., 1997. Hydrothermal-vent alvinellid polychaete dispersal in the eastern Pacific. influence of vent site distribution, bottom currents, and biological patterns. Limnol. Oceanogr. 42, 67–80.

Converse, D., Holland, H., Edmond, J., 1984. Flow rates in the axial hot springs of the East Pacific Rise (21°N): implications for the heat budget and the formation of massive sulfide deposits. Earth Planet. Sc. Lett. 69, 187–191.

Fouquet, Y., Auclair, G., Cambon, P., Etoubleau, J., 1988. Geological settings and mineralogical and geochemical investigations on sulfide deposits near 13°N on the East Pacific Rise. Mar. Geol. 84, 145–178.

Haymon, R.M., Fornari, D.J., Von Damm, K.L., Lilley, M.D., Perfit, M.R., Edmond, J.M., Shanks III, W.C., Lutz, R.A., Grebmeier, J.M., Carbotte, S., Wright, D., McLaughlin, E., Smith, M., Beedle, N., Olson, E., 1993. Volcanic eruption of the mid-ocean ridge along East Pacific Rise crest at 9° 45-52' N: direct submersible observations of seafloor phenomena associated with an eruption event in April, 1991. Earth Planet. Sc. Lett. 119, 85–101.

Jollivet, D., Chevaldonné, P., Planque, B., 1999. Hydrothermal-vent Alvinellid polychaete dispersal in the Eastern Pacific. 2. A metapopulation model based on habitat shifts. Evolution 53, 1128–1142.

Jollivet, D., Desbruyères, D., Bonhomme, F., Moraga, D., 1995. Genetic differenciation of deep-sea hydrothermal vent alvinellid populations (Annelida : Polychaeta) along the East Pacific Rise. Heredity 74, 376–391.

Juniper, S., Tebo, B., Karl, D., 1995. The Microbiology of Deep-Sea Hydrothermal Vents. CRC Press.

Khripounoff, A., Comtet, T., Vangriesheim, A., Crassous, P., 2000. Near-bottom biological and mineral particule flux in the Lucky Strike hydrothermal vent area (Mid-Atlantic Ridge). J. Mar. Syst. 25, 101–118.

Kim, S.L., Mullineaux, L.S., 1998. Distribution and near-bottom transport of larvae and other plankton at hydrothermal vents. Deep-Sea Res. II 45, 423–440.

Kim, S.L., Mullineaux, L.S., Helfrich, K.R., 1994. Larval dispersal via entrainment into hydrothermal vent plumes. J. Geophys. Res. 99(C6), 655–665.

Le Bris, N., Sarradin, P.M., Caprais, J.C., 2003. Contrasted sulphide chemistries in the environment of the 13°N EPR vent fauna. Deep-Sea Res. I 50, 737–747.

Le Bris, N., Zbinden, M., Gaill, F., 2005. Processes controlling the physico-chemical micro-environments associated with pompeii worms. Deep-Sea Res. I 52, 1071–1083.

Lutz, R.A., Jablonski, D., Turner, R.D., 1984. Larval development and dispersal at deep-sea hydrothermal vents. Science 226, 1451–1453.

MacDonald, K., 1982. Mid-ocean ridges: fine scale tectonic, volcanic and hydrothermal processes within the plate boundary zone. Ann. Rev. Earth Pl. Sc. 10, 155–190.

Marsh, A.G., Mullineaux, L.S., Young, C.M., Manahan, D.T., 2001. Larval dispersal potential of the tubeworm *riftia pachyptila* at deep-sea hydrothermal vents. Nature 411, 77–80.

Metaxas, A., 2004. Spatial and temporal patterns in larval supply at hydrothermal vents in the northeast Pacific Ocean. Limnol. Oceanogr. 49, 1949–1956.

Morton, B., Taylor, G., Turner, FRS amd Turner, J., 1956. Turbulent gravitational convection from maintained and instantaneous sources. Proc. R. Soc. Lond. Ser. A 234, 1–23.

Mullineaux, L.S., France, S.C., 1995. Dispersal mechanisms of deep-sea hydrothermal vent fauna. In: S.E. Humphris, R.A. Zierenberg, L.S. Mullineaux, R.E. Thomson (Eds.), Seafloor hydrothermal systems: physical, chemical, biological and geological interactions. Geological Monograph 91, Geophysical Monograph 91. American Geophysical Union, Washington, DC, 408–424.

Mullineaux, L.S., Mills, S.W., Sweetman, A.K., Beaudreau, A.H., Metaxas, A., Hunt, H.L., 2005. Vertical, lateral and temporal structure in larval distribution at hydrothermal vents. Mar. Ecol. Prog. Ser. 293, 1–16.

Mullineaux, L.S., Peterson, C.H., Micheli, F., Mills, S.W., 2003. Successional mechanism varies along a gradient in hydrothermal fluid flux at deep-sea vents. Ecol. Monogr. 73, 523–542.

Mullineaux, L.S., Speer, K.G., Thurnherr, A.M., Maltrud, M.E., Vangriesheim, A., 2002. Implications of cross-axis flow for larval dispersal along mid-ocean ridges. Cah. Biol. Mar. 43, 281–283.

Pradillon, F., 2002. Données sur les processus de reproduction et de développement précoce d'un eucaryote thermophile *Alvinella pompejana*. Doctoral, University Pierre et Marie Curie.

Pradillon, F., Le Bris, N., Shillito, B., Young, C.M., Gaill, F., 2005. Influence of environmental conditions on early development of the hydrothermal vent polychaete *alvinella pompejana*. J. Exp. Biol. 208, 1551–1561.

Pradillon, F., Shillito, B., Chervin, J.C., Hamel, G., Gaill, F., 2004. Pressure vessels for in vitro studies of deep-sea fauna. High Pressure Res. 24, 237–246.

Pradillon, F., Shillito, B., Young, C.M., Gaill, F., 2001. Developmental arrest in vent worm embryos. Nature 413, 698–699.

Sarradin, P.M., Caprais, J.C., Briand, P., Gaill, F., Shillito, B., Desbruyères, D., 1998. Chemical and thermal description of the environment of the Genesis hydrothermal vent community (13°N, EPR). Cah. Biol. Mar. 39, 159–167.

Thomson, R., Mihaly, S., Rabinovich, A., McDuff, R., Veirs, S., Stahr, F., 2003. Constrained circulation at endeavour ridge facilitates colonization by vent larvae. Nature 424, 545–549.

Tivey, M., 1995. Modeling chimney growth and associated fluid flow at seafloor hydrothermal vent sites. In: S.E. Humphris, R.A. Zierenberg, L.S. Mullineaux, R.E. Thomson (Eds.), Seafloor hydrothermal systems: physical, chemical, biological and geological interactions. Geological Monograph 91. American Geophysical Union, Washington, DC, 158–177.

Tyler, P.A., Young, C.M., 1999. Reproduction and dispersal at vents and cold seeps. J. Mar. Biol. Ass. U. K. 79, 193–208.

Tyler, P.A., Young, C.M., 2003. Dispersal at hydrothermal vents: a summary of recent progress. Hydrobiologia 503, 9–19.

Vrijenhoek, R.C., 1997. Gene flow and genetic diversity in naturally fragmented metapopulations of deep-sea hydrothermal vents animals. J. Hered. 88, 285–293.

Young, C.M., Lutz, R.A., Vrijenhoek, R.C., Won, Y., 2003. Dispersal barriers and isolation among deep-sea mussel populations (Mytilidae: Bathymodiolus) from Eastern Pacific hydrothermal vents. Mol. Ecol. 12, 169–184.

Tables and Figures

Figure 1: Hydrodynamics created by the jets. a) Representation of the inertial and convective parts of the jets; $d_{c,1}$ and $d_{c,2}$ are the critical distances at which inertial jets become convective, respectively for the main and secondary jets. b) Left, a qualitative view of the speed field created by the main jet in the three differents zones: the fluid velocity decreases far from the outlet both laterally and vertically. Right, a summary of the notations used in the paper. In inset, a zoom on the apex of the chimney, with the outlet of the main jet.

Figure 2: Effect of the maximum texture depth R and the larval release point distance from the chimney wall, on the colonization rate. Larvae are released from a vertical (a) or a flat (b) smoker, at different distances from the chimney wall, which is represented by the value of λ (see the text): black circles, squares, triangles and stars are respectively for λ values equal to 0.5, 1, 1.5, and 20.

Figure 3: Effect of the main smoker jet velocity on the colonization rate. Larvae are released from a diffuser located on the side of either a vertical (a) or a flat (b) smoker, at different distances from the top of the chimney (main jet output): 0.5 metres below the main jet (black circles), at mid-height of the chimney (black squares) and 0.5 metres above the base of the chimney (black triangles).

Figure 4: Effect of bottom current velocity on the colonization rate. Bottom current orientation is set such that larval release point is on the face of the smoker hidden from the current. Larvae are released from the wall of a chimney, at different distances from the main smoker output: 0.5 metres below the main smoker (black circles and stars), at mid-height of the chimney (black squares) and 0.5 metres above the base of the chimney (black triangles). The main jet speed is 2 $m.s^{-1}$. Stars stand for the same simulation, but with a main jet speed of 0.1 $m.s^{-1}$ and a larval release point 0.5 metres below the main jet output. If the colony is located on the wall facing the bottom current, colonization rate is invariably 100% (not shown).

Figure 5: Effect of a secondary jet's temperature and velocity at the output on larvae mortality rate. The larvae are released from the chimney wall, 0.5 meters below the main smoker. Secondary jet temperatures are 10°C (black circles), 20°C

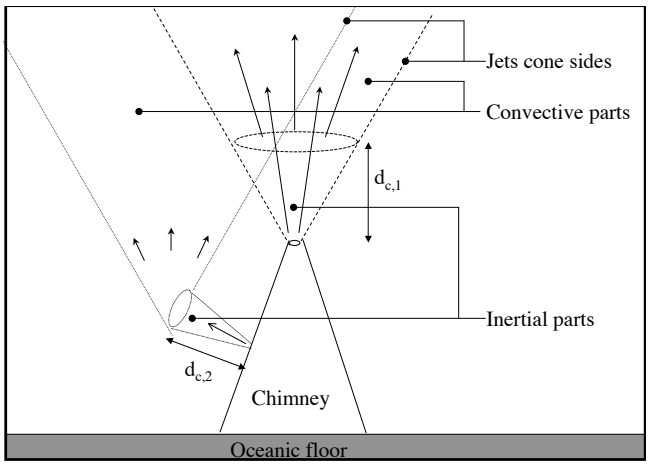
(open squares), 40°C (black triangles), 60°C (stars), 80°C (open losanges), 100°C (crosses). These results are independent of the position of the larvae release point.

Figure 6: Effect of the larvae buoyancy rates in different hydrodynamic contexts. Larvae are released from the wall of a chimney whose main smoker jet velocity is 0.1 $m.s^{-1}$ (a) or 2 $m.s^{-1}$ (b), 0.5 metres below the main smoker. Bottom currents with different velocities, directed so that the larvae are on the hidden face of the smoker, are compared: no current (black circles), 0.001 $m.s^{-1}$ (open squares), 0.01 $m.s^{-1}$ (black triangles) and 0.1 $m.s^{-1}$ (stars).

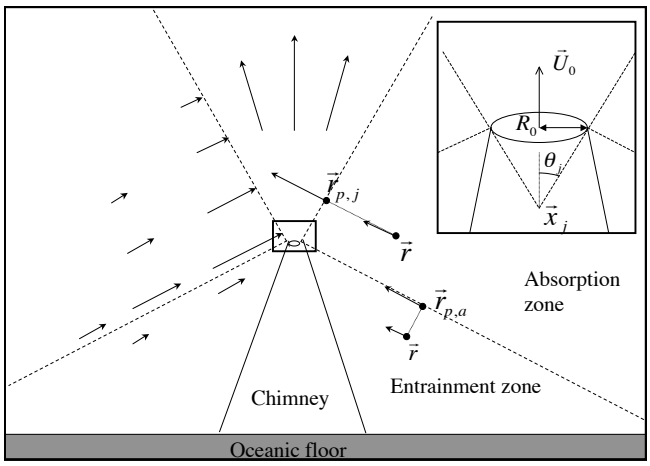
Figure 7: Prediction of larval trajectories of two vent species, *Alvinella pompejana* and *Riftia pachyptila*, at a 13°N EPR vent site. The mineral environment is modeled so as to represent the chimney PP HOT2 at the Genesis site described by Sarradin et al. (1998) (see text). *A. pompejana* larvae predicted trajectories are given on the left. *R. pachyptila* trajectories are given on the right. In (a) and (d), the main jet velocity is set to 3 $m.s^{-1}$. In (b) and (e), the main jet velocity is lower, 1 $m.s^{-1}$. In (c) and (f) we add bottom currents directed towards the right side of the figure, with speed 0.01 $m.s^{-1}$. In this case the main jet speed is also 3 $m.s^{-1}$. Trajectories of the larvae which settled or were ejected at the end of the simulation are respectively in green and blue. Percentages of settled larvae are indicated.

	Minimal values	Maximal values	Literature reported values & refs
Topography			
Distance between jets outlets $\|\vec{x}_1 - \vec{x}_2\|$	0.5 m	15 m	0.1 to several metres
Chimney wall texture depth R	0.001 m	0.05 m	no data
Hydrodynamics			
Primary jet speed	0.01 m.s^{-1}	4 m.s^{-1}	0.5 - 3.5 m.s^{-1} (Converse et al., 1984; Fouquet et al., 1988)
Primary jet temperature	50°C	400°C	150 - 400°C (Tivey, 1995)
Secondary jet speed	0.001 m.s^{-1}	0.05 m.s^{-1}	0.005 - 0.01 m.s^{-1} (Juniper et al., 1995)
Secondary jet temperature	10°C	100°C	5 - 90°C (Le Bris et al., 2003, 2005; Sarradin et al., 1998)
Bottom current speed v_b	0.0001 m.s^{-1}	1 m.s^{-1}	0.005 - 0.5 m.s^{-1} (Cannon et al., 1991; Kim and Mullineaux, 1998; Thomson et al., 2003)
Biological characteristics			
Larval buoyancy b	$\pm$0.00005 m.s^{-1}	$\pm$0.005 m.s^{-1}	$\pm$0.00003 m.s^{-1} (Marsh et al., 2001; Pradillon et al., 2004)

Table 1: Minimum and maximum values used in simulations for each parameter tested. The values for fixed parameters are given in Materials and Methods.

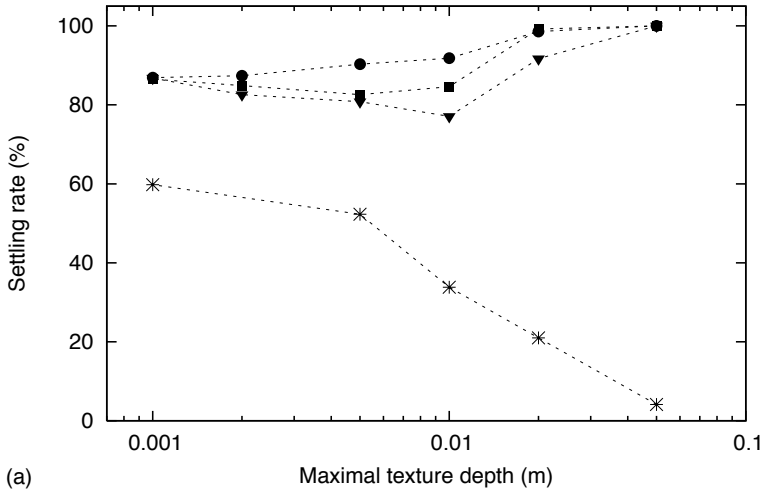

Jets cone sides
Convective parts
$d_{c,1}$
Inertial parts
$d_{c,2}$
Chimney
Oceanic floor

(a)

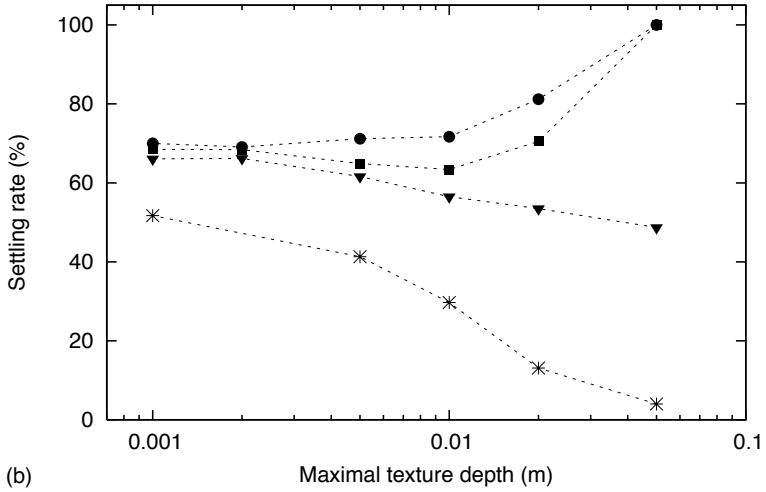

$\vec{U}_0$
R_0
θ_j
$\vec{x}_j$
$\vec{r}_{p,j}$
$\vec{r}$
$\vec{r}_{p,a}$
$\vec{r}$
Absorption zone
Entrainment zone
Chimney
Oceanic floor

(b)

Figure 1.

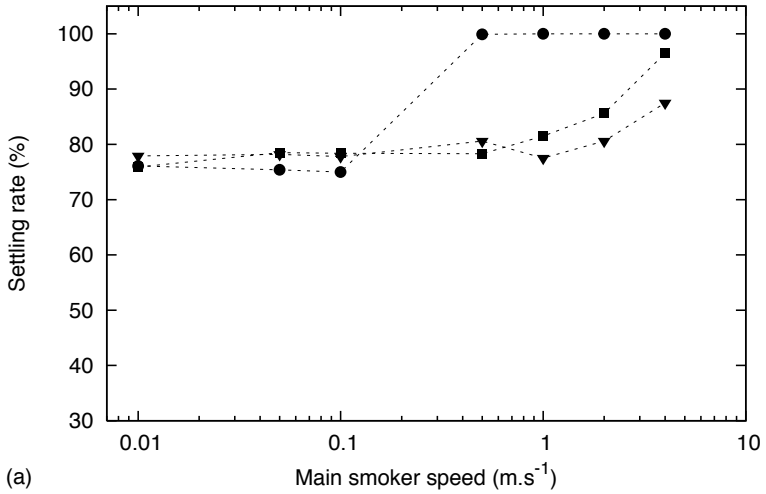

Settling rate (%)
100
80
60
40
20
0
0.001
0.01
0.1
Maximal texture depth (m)

(a)

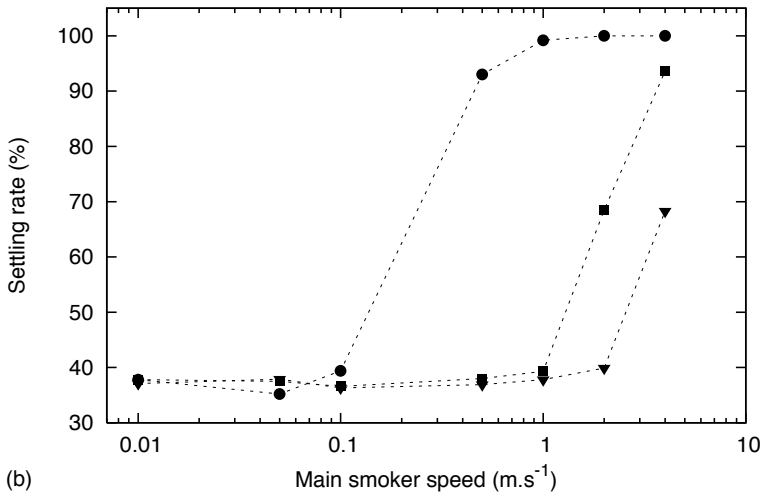

Settling rate (%)
100
80
60
40
20
0
0.001
0.01
0.1
Maximal texture depth (m)

(b)

Figure 2.


Settling rate (%)
100
90
80
70
60
50
40
30
0.01
0.1
1
10
Main smoker speed (m.s-1)

(a)


Settling rate (%)
100
90
80
70
60
50
40
30
0.01
0.1
1
10
Main smoker speed (m.s-1)

(b)

Figure 3.

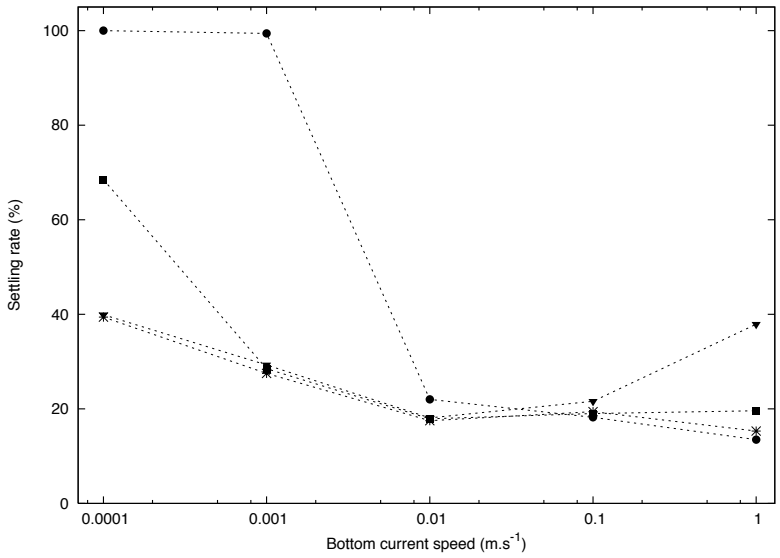

100
80
60
40
20
0
Settling rate (%)
0.0001
0.001
0.01
0.1
1
Bottom current speed (m.s-1)

Figure 4.

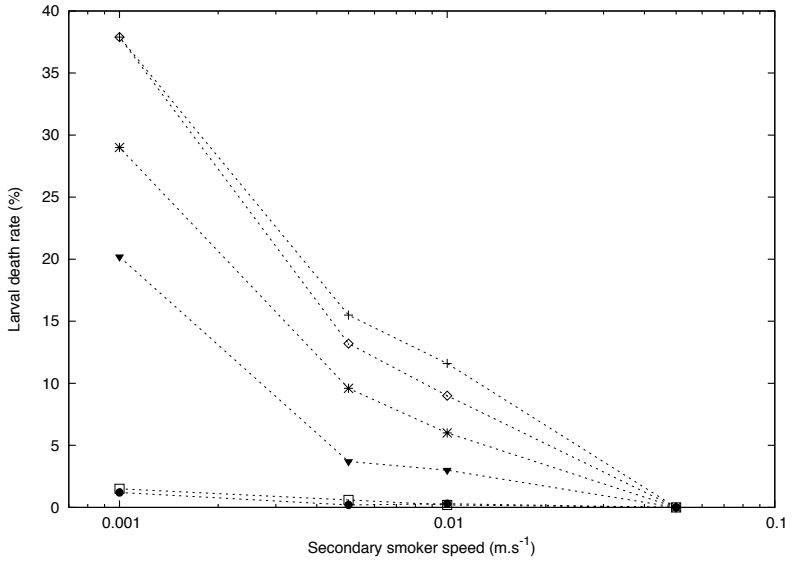

Larval death rate (%)
40
35
30
25
20
15
10
5
0
0.001
0.01
0.1
Secondary smoker speed (m.s-1)

Figure 5.

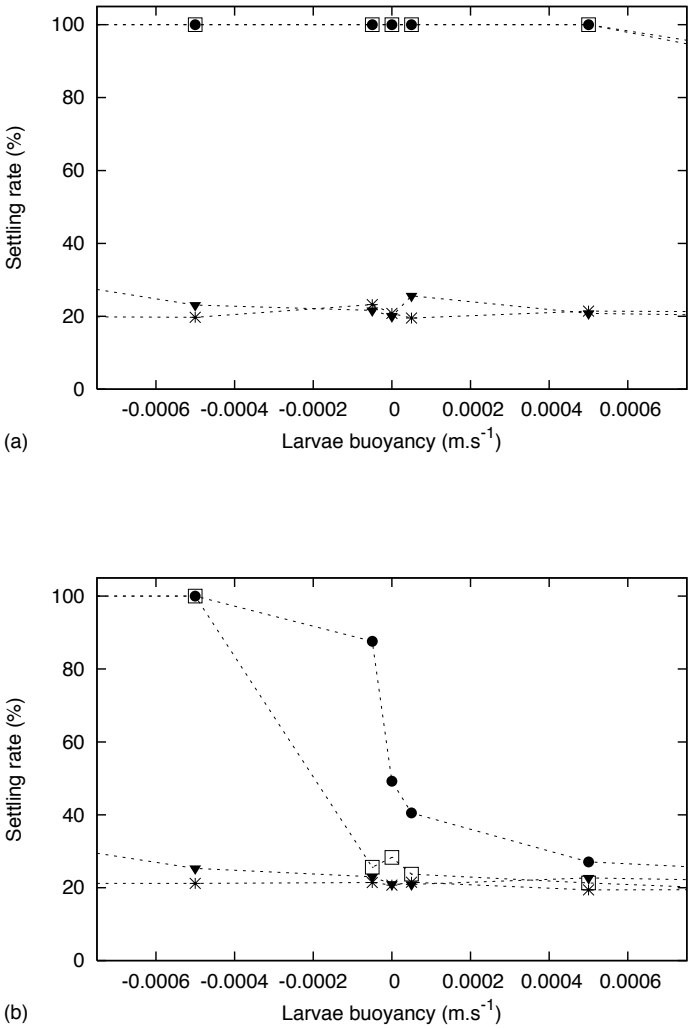

100
80
60
40
20
0
Settling rate (%)
-0.0006 -0.0004 -0.0002 0 0.0002 0.0004 0.0006
Larvae buoyancy (m.s-1)
(a)
100
80
60
40
20
0
Settling rate (%)
-0.0006 -0.0004 -0.0002 0 0.0002 0.0004 0.0006
Larvae buoyancy (m.s-1)
(b)

Figure 6.

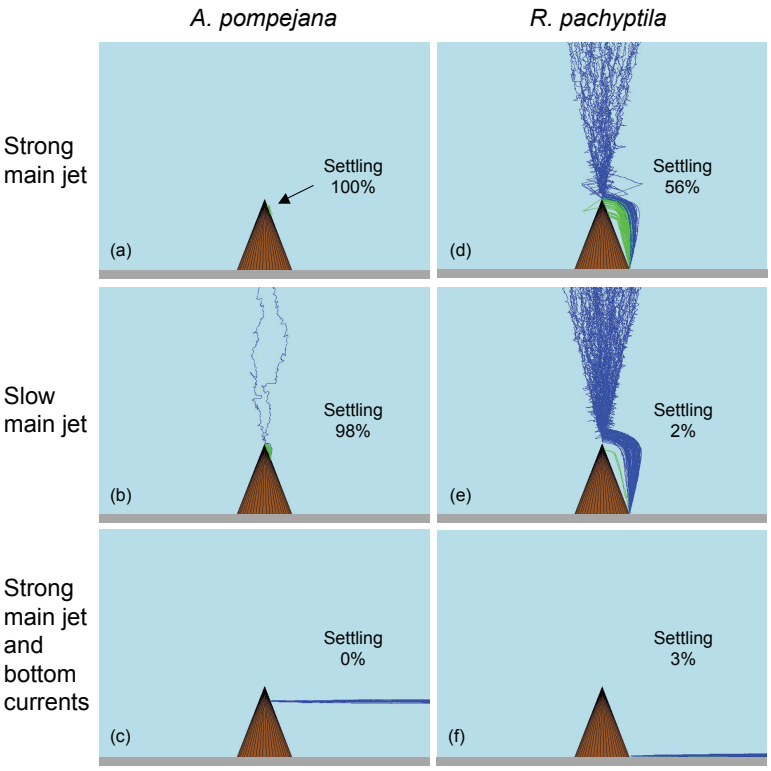

A. pompejana
R. pachyptila
Strong main jet
Slow main jet
Strong main jet and bottom currents
Settling 100%
Settling 56%
Settling 98%
Settling 2%
Settling 0%
Settling 3%
(a)
(b)
(c)
(d)
(e)
(f)

Figure 7.

Supplementary Materials for "A modeling approach of the influence of local hydrodynamic conditions on larval dispersal at hydrothermal vents"

Marc Bailly-Bechet, Michel Kerszberg,
Françoise Gaill, Florence Pradillon

1 Relevant dimensionless numbers

Deep hydrothermal sources are modeled as heated water jets stationary in time. As hydrodynamic equations are very time and computer intensive to solve, an empirical approach was employed to describe the fluid dynamics. The speed and temperature field, as the geometry of the jets, are determined using dimensional analysis and conservation laws. These features only depend on few parameters, such as the jets speed and temperature at the outlet. An approach based on dimensionless number analysis can seem simplistic, but is justified because i) little is known about the precise hydrodynamics of hydrothermal vents, and ii) this type of analysis allows focus only on the more important phenomena.

To evaluate the speed and temperature field in and out of the heated water jets, we use the Boussinesq approximation of the convection phenomena:

$$\vec{\nabla} \cdot \vec{u} = 0 \tag{1a}$$

$$\frac{\partial \vec{u}}{\partial t} + (\vec{u} \cdot \vec{\nabla})\vec{u} = -\frac{1}{\rho}\vec{\nabla}P + \nu\nabla^2\vec{u} - \vec{g}\alpha\Delta T \tag{1b}$$

$$\frac{\partial T}{\partial t} + (\vec{u} \cdot \vec{\nabla})T = \kappa\nabla^2 T \tag{1c}$$

We use the classical notations: $\vec{u}$ is the water speed, ρ the water density, P the pressure, ν the viscosity, $\vec{g}$ is gravity, α is the coefficient of thermal expansion of

the water, ΔT the difference in temperature with the surrounding media, and κ the thermal diffusivity of water. The smokers we are studying have been reported to have semidiurnal activity periods, so we assume them stationary in time at the hour scale, and discard the time dependent part of these equations. We also discard the pressure term because the solution of these equations does not depend on this hydrostatic term. We identify the typical scale of each term of these equations by rewriting them in function of the typical parameters of the jet. We have a jet size of order of magnitude L, its speed $\vec{u} = U\vec{v}$, and gravity $\vec{g} = g\vec{k}$ with $\vec{v}$ and $\vec{k}$ unitary. By considering gradients are significant on the typical size L, and by noting the dimensionless gradient $\vec{\nabla} = \frac{\vec{\nabla_L}}{L}$, we have:

$$\frac{U^2}{L}(\vec{v}\cdot\vec{\nabla_L})\vec{v} = \frac{U\nu}{L^2}\nabla_L^2\vec{v} - g\alpha\Delta T\vec{k} \tag{2}$$

$$\frac{U}{L}(\vec{v}\cdot\vec{\nabla_L})T = \frac{\kappa}{L^2}\nabla_L^2 T \tag{3}$$

Equation (1a), representing water incompressibility, is not used now, but will be later. We then rewrite these equations respectively in function of the Reynolds number $Re = \frac{UL}{\nu}$, the Richardson number $Ri = \frac{g\alpha\Delta TL}{U^2}$, and the Peclet number $Pe = \frac{UL}{\kappa}$. Note that the Richardson number employed here is the *overall* Richardson number, as defined by Tritton (1977), p 168, and not the *gradient* Richardson number commonly used to consider entrainment against stratification in stratified fluid problems. We have:

$$(\vec{v}\cdot\vec{\nabla_L})\vec{v} = \frac{1}{Re}\nabla_L^2\vec{v} - Ri\,\vec{k} \tag{4}$$

$$(\vec{v}\cdot\vec{\nabla_L})T = \frac{1}{Pe}\nabla_L^2 T \tag{5}$$

Now, we consider the range of variation of these dimensionless numbers, to understand which phenomenon is dominant at the bio-hydrodynamical scale. We take typical values at the outlet of the smoker for this analysis: 10^{-2} m.s^{-1} $< U <$ 1 m.s^{-1} for the speed of these jets at the outlet and 10^{-2} m $< L <$ 1 m for the ray of the outlet. We know $\nu \simeq 10^{-6}$ m^2.s^{-1} from Cho et al. (1999), $\kappa \simeq 10^{-7}$ m^2.s^{-1} and α ranging from 10^{-4} K^{-1} at 20°C to 10^{-3} K^{-1} at 300°C (Irvine and Duigan, 1985). It gives, at the outlet of the smoker:

$$10^2 < Re < 10^6 \tag{6a}$$
$$10^3 < Pe < 10^7 \tag{6b}$$
$$10^{-4} < Ri < 10^4 \tag{6c}$$

The values found for the Reynolds number show that, at these scales, the flow is turbulent as viscosity is dominated by inertia forces. The values of the Peclet number mean that temperature is transported by convection, and not conduction. On the contrary, values for the Richardson number vary over a wide range around 1. This implies that the dynamics of the jets, at this scale, can be either dominated by inertial or convective effects, and depends on the value of the Richardson number Ri for each jet. We now use these results to decide which phenomena are modeled in our jet simulation.

2 Turbulent jets geometry

To explain how jet dynamics are simulated, we take the example of the primary jet, corresponding to the main smoker. This jet is located at the top of the chimney and has a vertical axis. All the results are applied similarly to the secondary jet, in the corresponding geometry, as the secondary jet is modeled with a non-vertical axis.

Turbulent jets are modeled with a conical divergent shape, with the point source located inside the chimney (justification for the conical shape is given below). At the apex of the chimney the jet has a ray R_0 and a speed $\vec{U_0}$, directed upwards. The properties of the water being significanlty different inside and outside the jet, the hydrodynamics inside and outside of the jet are modeled separately. The first quantity to be estimated is the ray of the jet $R(l)$ at a distance l of the point source, i.e the shape of the jet. Note here that we model the jets as having a shape expanding from the source point, and not as infinite in width with, by example, a gaussian profile. As the jet is turbulent, R cannot be dependent on the viscosity ν. The only relevant parameters are water speed $\vec{U}$, density ρ and the distance l from the outlet. Dimensional analysis (see Tritton (1977), p89) tells us that $R(l, \rho, \vec{U}) \sim l$ is the only possible dependence for R . So, by defining θ_j as the half-angle of opening of the jet at the point source, we have:

$$R(l) = l \tan(\theta_j) \tag{7}$$

which justifies our choice of a conical shape for the turbulent jets. This choice is also verified by experiments which have shown that the shape of a turbulent jet is conical (Tritton, 1977).

3 Jet dynamics

To estimate the speed and temperature field inside a heated water jet, we find the dominant momentum source inside the jet. We consider it as the only momentum source, which allows us to quantify both fields.

3.1 Inertial jets

If $Ri < 1$, the source of movement is inertia, and heat is transported by forced convection. No external forces are applied on the water and the momentum is conserved along the jet, giving across any perpendicular cross-section S of the jet at distance l from the outlet (Tritton, 1977):

$$\iint_S \rho U^2(l) dS = \pi R_0^2 \rho U_0^2 \tag{8}$$

On the other side, as no heat is generated along the jet, we also have:

$$\iint_S \Delta T(l) U(l) dS = \pi R_0^2 \Delta T_0 U_0 \tag{9}$$

where $\Delta T(l)$ is the difference in temperature between the jet and the surrounding media. U_0 is the speed at the outlet and ΔT_0 the temperature difference with the ocean water. Based on these equations, and using $R(l) = l \tan(\theta_j)$, one can find with dimensional analysis the dependence of speed $U(l)$ and temperature $\Delta T(l)$ with l. We have:

$$U(l) \sim l^{-1} \Rightarrow U(l) \simeq U_0 \frac{R_0}{l} \tag{10a}$$

$$\Delta T(l) \sim l^{-1} \Rightarrow \Delta T(l) \simeq \Delta T_0 \frac{R_0}{l} \tag{10b}$$

The multiplicative constant have been omitted here to emphasize that these relations give us the typical behavior of the jet in function of the parameters, not an exact solution. We call this case an *inertial* jet. The previous equations are theoretically

valid only far from the outlet, for $l \gg R_0$. They will be used for all l in our case. This simplification limits the validity of our conclusions far enough of the jet outlet. However, the qualitative behavior of the jet close to its outlet is the same as predicted by these equations, allowing us to tentatively generalize our results. One direct consequence of these equations is an increase of the jet flow $F(l)$ with l :

$$F(l) \sim U(l)R^2(l) \sim U_0 R_0 l \tag{11}$$

This increase means that the cold water surrounding the jet is entrained inside the jet. The qualitative explanation for this phenomenon is that local turbulence along the sides of the jet cone creates a local speed fields directed towards the jet axis (see Landau and Lifshitz (1959); Tritton (1977)).

The speed field around the conical jet can be estimated by using indirectly the formula (11). $\vec{r}_{p,j}(l)$ is a point on the side of the jet at a distance l from its outlet. The speed $\vec{U}(\vec{r}_{p,j})$ of the water outside of the jet at this point, is defined as locally orthogonal to the surface of the conical jet and directed towards the jet axis (see Fig. 1). Conservation of momentum gives:

$$\frac{\partial F(l)}{\partial l} \simeq 2\pi R(l) U(\vec{r}_{p,j}) \tag{12}$$

which results after replacement in:

$$U(\vec{r}_{p,j}) \simeq \frac{U_0 R_0}{2l} \tan(\theta_j) \tag{13}$$

This movement orthogonal to the boundary of the jet cone propagates far from the jet by continuity. Using mass conservation, i.e water incompressibility, one finds, at a lateral distance r_{lat} from the jet side and vertical distance l from the outlet:

$$U(l, r_{lat}) = U(\vec{r}_{p,j}) \frac{R(l)}{r_{lat} + R(l)} \simeq \frac{U_0 R_0}{2(r_{lat} + R(l))} \tan^2(\theta_j) \tag{14}$$

This speed field is valid for the points who have a projection $\vec{r}_{p,j}$ on the side of the jet cone, i.e for points inside the *zone of absorption* (Fig. 1). It models effectively both a vertical and an horizontal entrainment, as the streamlines in this zone are orthogonal to the jet cone surface. The vertical entrainment denotated by this equation is also found in more classical jet models, with a $\cosh^{-2}$ or gaussian profile.

In the zone below this absorption zone (that is, the entrainment zone in Fig. 1), the speed field is estimated by continuity with the absorption zone field. Water here is put to movement by viscous contact with the layer of water above it in the

zone of absorption. So the water speed there depends directly on the water speed in the absorption zone. In this zone $\vec{U}(\vec{r})$ is computed as follows. We note $\vec{r}_{p,a}$ the projection of $\vec{r}$ on the line starting at the basis of the jet, and perpendicular to it (see Fig. 1). This line is the separation line between the absorption and entrainment zones. $\vec{U}(\vec{r})$ is supposed to decrease with $\|\vec{r} - \vec{r}_{p,a}\|$. To keep the number of parameters as low as possible, a simple formula with a non-parametric asymptotic decrease is proposed:

$$\vec{U}(\vec{r}) = \frac{1}{1 + 5\|\vec{r} - \vec{r}_{p,a}\|}[\vec{v}(\vec{r}_{p,a})] \tag{15}$$

As before, the notation $[\vec{k}]$ represents the unitary vector $\frac{\vec{k}}{\|\vec{k}\|}$. The choice of the value 5 is arbitrary. Simulations have been done with a value of one with no significant differences, letting us believe that the particular formula employed here is not of great importance for the larval dispersal problem.

3.2 Convective jets

We now expose the study of another type of jet, the convective jets. We follow the same methodology as in the previous chapter, and only document the main results here. If $Ri > 1$, heat generates momentum inside the jet, and the jet is modeled using free convection laws. The quantity of main interest is the heat transported per unit time over each cross-section S of the jet. The speed and temperature fields, as the flow, inside the jet cone, are found using conservation integrals. The speed field outside of the jet in the different regions of space is found using the same technique as for the inertial jets.

First the dependence of $U(\vec{r})$ and $\Delta T(\vec{r})$ with l are found using conservation integrals. Conduction being neglected, heat is conserved along the jet, and momentum increases due to convection. Across each cross-section S of the jet, we have (see Landau and Lifshitz (1959) p272):

$$\iint_S U(l)\Delta T(l)dS = \pi U_0 \Delta T_0 R_0^2 \tag{16a}$$

$$\frac{\partial}{\partial l}\iint_S \rho U^2(l)dS = \rho g \alpha \Delta T(l) \tag{16b}$$

Dimensional analysis gives:

$$\Delta T(l) \sim l^{-\frac{5}{3}} \Rightarrow \Delta T(l) \simeq \Delta T_0 \left(\frac{R_0}{l}\right)^{-\frac{5}{3}} \tag{17}$$

$$U(l) \sim l^{-\frac{1}{3}} \Rightarrow U(l) \simeq U_0 \left(\frac{R_0}{l}\right)^{\frac{1}{3}} \tag{18}$$

Following the same lines as previously, we deduce the speed $\vec{U}(l, r)$ for every point with a projection on the cone of the jet, at vertical distance l and lateral distance r_{lat} from the jet cone side. We have:

$$U(l, r) \simeq \tan^2(\theta_j) U_0 \frac{R_0^{1/3} l^{2/3}}{r_{lat} + R(l)} \tag{19}$$

Finally the speed in the entrainment zone is computed as before, based on equation (19) instead of (14).

4 Jet evolution

Plugging the dependencies for speed and ray with the distance l to the outlet into the formula for Ri, one finds that Ri is an increasing function of the distance l to the outlet, with $Ri \propto l^2$ for an inertial jet and Ri constant in function of l for a convective one. This leads to say that, far enough from the outlet, all jets are convectively driven, whether they were already convective or inertial at their outlet. The critical distance d_c at which a jet changes of dominant force, from inertia to convection, is defined such that $Ri(d_c) = 1$.

At the beginning of the simulations, we first define if, at the outlet, a jet is inertial or convective. We compute its Richardson number at the outlet, with the values for speed and temperature chosen for the simulation. If it is convective, it will remain so, and is modeled on all its length by the equations (17) and (18). If it is inertial, d_c is computed, and two cases can happen:

1. $d_c < R_0$. We consider the jet to be convective from its outlet to its end. As our equations are true only far from the outlet, we cannot consider as significant a critical distance d_c smaller than the minimal ray of the jet R_0, which represents the scale at which our simplifications fall. So we consider the inertial part of this jet negligibly small and compute its speed and temperature fields with equations (17) and (18) from the beginning.

2. $d_c > R_0$. We model the part of the jet for $0 \leq l \leq d_c$ with the equations (10a) and (10b) for the inertial jets, and the part for $l > d_c$ with the ones for a convective jet, i.e (17) and (18). The values R_0 and U_0 for the convective part are found by continuity with the inertial part at $l = d_c$, which gives $R_{0,conv} = R(l = d_c)$ and $U_{0,conv} = U_{inertial}(l = d_c)$. Note that the convective part of a jet is always directed upwards, whether or not the first part is.

This final point allows us to model jets with variable dynamics relative to the distance to the outlet. In general, the main smoker located on the top of the chimney is a jet with a change in dominant force, while the secondary jet is already convective at its outlet. These are not rules, but observations from the typical values of the speeds and temperature at the outlet of these jets. The precise jet behavior is computed at the beginning of each simulation.

References

Cho, C., Urquidi, J., Robinson, G., 1999. Molecular-level description of temperature and pressure effects on the viscosity of water. J. Chem. Phys. 111.

Irvine, T., Duigan, M., 1985. Isobaric thermal expansion coefficients for water over large temperature and pressure ranges. Int. Comm. Heat Mass Transfer 12.

Landau, L., Lifshitz, E., 1959. Mécanique des fluides. Pergamon.

Tritton, D., 1977. Physical fluid dynamics. ELBS.

A.3 Perspectives

Le simulateur que nous avons développé permet d'explorer de nombreux autres problèmes que celui de la dispersion larvaire. Étant très modulable, il peut prendre en compte des topographies différentes, être utilisé pour l'étude de problèmes de relations entre espèces, et peut être élargi pour quantifier les interactions entre différents processus écologiques comme la reproduction, la prédation, la compétition pour les ressources. Plusieurs voies on été évoquées pour des projets de recherche ultérieurs basés sur le même type de simulations. Il s'agit par exemple de :

- L'analyse de la dynamique spatio-temporelle de la colonisation sur un fumeur. Des simulations plus précises et plus poussées sont possibles maintenant que nous avons identifié les paramètres hydrodynamique et topographiques d'importance dans les écosystèmes hydrothermaux. Ceci nous permettrait d'affiner nos premiers résultats, et de déterminer plus précisément les facteurs influençant la survie des larves dans les premières heures après la colonisation, ce qui peut être d'une importance considérable sur la dynamique des populations de ces espèces.
- La modélisation, avec le même simulateur, des diffuseurs de faible puissance que l'on observe parfois autour des fumeurs principaux, sur le fond océanique. Ces diffuseurs jouent peut-être un rôle dans la dispersion des espèces entre écosystèmes hydrothermaux voisins.
- La reconstitution de sites réels, qui permettrait de réaliser des prédictions sur la distribution spatiale des organismes, et de comparer à court terme (quelques mois) l'évolution de ses sites. Une telle connaissance serait très utile aux océanographes de terrain, mais le peu de données topographiques auxquelles nous avons accès pour le moment rendent difficile ce travail.

Annexe B

Méthode de classification des gènes

Notre méthode de classification vise à partitionner en S groupes un ensemble de N gènes, de façon à maximiser la cohérence de l'usage des codons par les gènes d'un même groupe, et à rendre ces groupes les plus différents possibles les uns des autres.

On note $n_{g,a}(\ell)$ le nombre de fois où le codon ℓ est employé dans le gène g pour coder pour l'acide aminé a. La dégénérescence de l'acide aminé a est notée q_a : par exemple, on a $q_{Leu} = 6$. Supposons maintenant qu'il existe une structure sous-jacente de groupes dans l'emploi des codons à l'intérieur des gènes ; cette structure est caractérisée par les probabilités $p_{s,a}(\ell)$, qui donnent au sein de chaque groupe C_s la probabilité que le codon ℓ soit employé pour coder l'acide aminé a. Cette distribution donne les fréquences relatives d'usage de chaque codon qui caractérisent le groupe. On a bien sûr :

$$\sum_{\ell} p_{s,a}(\ell) = 1. \tag{B.1}$$

Calculons maintenant la vraisemblance qu'un gène g donné ait été généré par la distribution des $p_{s,a}(\ell)$ du groupe s. On a une distribution multinomiale des comptes, ce qui nous donne :

$$\mathcal{L}(g|p_{s,a}(\ell)) = \prod_{a=1}^{18} \frac{\Gamma\left(\sum_{\ell} n_{g,a}(\ell) + 1\right)}{\prod_{\ell} \Gamma\left(n_{g,a}(\ell) + 1\right)} \prod_{l=1}^{q_a} p_{s,a}(\ell)^{n_{g,a}(\ell)}. \tag{B.2}$$

La vraisemblance d'une classification particulière de tous les gènes $\{C_s\}$, sachant que cette structure sous-jacente en groupes existe, est donc donnée par le produit de cette vraisemblance sur tous les éléments des groupes :

$$\mathcal{L}(\{C_s\}|p_{s,a}(\ell)) = \prod_{s} \prod_{g \in C_s} \mathcal{L}(g|p_{s,a}(\ell)). \tag{B.3}$$

Nous allons employer la formule de Bayes pour inverser la vraisemblance conditionnelle B.2, et calculer $\mathcal{L}(p_{s,a}(\ell)|C_s)$, la distribution des valeurs des probabilités $p_{s,a}(\ell)$ de chaque groupe C_s sachant les gènes qu'il contient. On peut voir cette vraisemblance comme la distribution des $p_{s,a}(\ell)$ qui représente le mieux les gènes présents dans le groupe, la structure qui correspond le mieux à une classification donnée.

L'application de la formule de Bayes nécessite la définition d'une distribution *a priori*, à savoir la distribution des $p_{s,a}(\ell)$ quand aucune information sur le contenu des groupes n'est connue. Dans notre cas, le choix *a priori* doit seulement contenir l'information que

tous les codons existent et peuvent être utilisés. Pour cela, on choisit une distribution uniforme :

$$\mathcal{P}_0\left(p_{s,a}(\ell)\right) = \Gamma(q_a)\,\delta\left(\sum_\ell p_{s,a}(\ell) - 1\right), \tag{B.4}$$

qui n'est rien d'autre que la formulation mathématique de l'uniformité des $p_{s,a}(\ell)$ et du fait qu'elles soient normalisées. L'emploi d'une distribution *a priori* uniforme dans l'espace des logarithmes n'est pas nécessaire ici, car on sait que tous les codons peuvent être utilisés (Jaynes, 1967). On applique ensuite la formule de Bayes pour calculer la distribution *a posteriori* pour chaque groupe, que l'on notera $\mathcal{P}_{post}$:

$$\mathcal{P}_{post} = \mathcal{L}(p_{s,a}(\ell)|C_s) = \frac{\prod_{g\in C_s}\mathcal{L}(g|p_{s,a}(\ell))\,\mathcal{P}_0(p_{s,a}(\ell))}{\int\prod_{g\in C_s}\mathcal{L}(g|p_{s,a}(\ell))\,\mathcal{P}_0(p_{s,a}(\ell))\,dp_{s,a}(\ell)}. \tag{B.5}$$

L'intégration au dénominateur peut être calculée analytiquement grâce à la formule de Dirichlet :

$$\int\prod_{k=1}^{N} p_k^{n_k}\,\delta\left(\sum_{p_k} p_k - 1\right)\,dp_k = \frac{\Gamma(\sum_{k=1}^{N} n_k + N)}{\prod_{k=1}^{N}\Gamma(n_k+1)} \tag{B.6}$$

Pour faciliter la lecture, nous allons employer les notations $N_a^s(\ell) \equiv \sum_{g\in\mathcal{C}_s} n_{g,a}(\ell)$ pour le nombre de codons ℓ employés dans chaque groupe et $N_a^s \equiv \sum_\ell N_a^s(\ell)$ pour le nombre d'acides aminés employés dans chaque groupe. En réalisant l'intégration dans (B.5) et en remplaçant, on obtient la formule suivante pour la distribution *a posteriori* :

$$\mathcal{P}_{post}(p_{s,a}(\ell)) = \frac{\Gamma\left(N_a^s + q_a\right)}{\prod_{\ell=1}^{q_a}\Gamma(N_a^s(\ell)+1)}\delta\left(\sum_l p_{s,a}(\ell) - 1\right)\prod_{\ell=1}^{q_a} p_{s,a}(\ell)^{N_a^s(\ell)}. \tag{B.7}$$

À partir de ce point deux formalismes différents conduisent au même critère de partition, l'un basé sur la maximisation de l'énergie libre de la classification, l'autre sur la théorie de l'information. Je vais les détailler successivement.

Maximisation de l'énergie libre Le système que l'on étudie est un système désordonné : les probabilités associées aux groupes $p_{s,a}(\ell)$ sont des paramètres dont on ne peut estimer que la distribution. Or, la vraisemblance d'une classification est fonction de ces probabilités, car elle reflète l'adéquation entre la structure sous-jacente en groupes, représentée par les $p_{s,a}(\ell)$, et la classification, donnée par les attributions des gènes dans les groupes. On cherche la classification la plus vraisemblable des gènes : il nous faut donc maximiser une fonction de la vraisemblance conditionnée aux valeurs des $p_{s,a}(\ell)$, moyennée relativement aux valeurs possibles des $p_{s,a}(\ell)$. Or les comptes des gènes nous ont permis de déterminer la distribution *a posteriori* de ces probabilités ; le moyennage va donc avoir lieu sur cette distribution.

La dernière question à laquelle il nous faut répondre est celle du choix de la fonction de la vraisemblance à maximiser. On recherche la classification typique des gènes : il nous faut donc maximiser le logarithme de la vraisemblance, et non pas la vraisemblance elle-même, pour éviter que les valeurs très peu probables des $p_{s,a}(\ell)$ ne biaisent la valeur de la vraisemblance. De plus, ce choix nous permettra de faire le lien avec la seconde partie par la suite. La vraisemblance d'observer chaque classification est donc :

$$\log\left(\mathcal{L}(\{C_s\})\right) = \sum_s\int\log\left(\mathcal{L}(C_s|p_{s,a}(\ell))\right)\,\mathcal{P}_{post}(p_{s,a}(\ell))\,dp_{s,a}(\ell), \tag{B.8}$$

où la somme sur les groupes vient du passage au logarithme dans l'intégrale. Le calcul de cette intégrale peut être calculée en employant le "replica trick"[1] :

$$\langle \log(f) \rangle = lim_{n\to 0} \frac{\langle f^n \rangle - 1}{n}, \tag{B.9}$$

où la moyenne est l'intégration sur une distribution de probabilités, ici (B.7). Nous allons calculer le terme intégral au numérateur, sans tenir compte du -1. Il est égal à :

$$\sum_s \int \mathcal{L}^n(C_s|p_{s,a}(\ell))\, \mathcal{P}_{post}(p_{s,a}(\ell))\, dp_{s,a}(\ell). \tag{B.10}$$

En remplaçant les termes par leurs expressions (B.2) et (B.7), on obtient :

$$\sum_s \prod_a \frac{\Gamma(N_a^s + q_a)}{\prod_{\ell=1}^{q_a} \Gamma(N_a^s(\ell)+1)} \int \prod_\ell p_{s,a}(\ell)^{N_a^s(\ell)(n+1)}\, \delta\left(\sum_p p_{s,a}(\ell) - 1\right)\, dp_{s,a}(\ell), \tag{B.11}$$

ce qui, après application de la formule de Dirichlet, donne :

$$\sum_s \prod_a \frac{\Gamma(N_a^s + q_a)}{\prod_\ell \Gamma(N_a^s(\ell)+1)} \frac{\prod_\ell \Gamma((n+1)N_a^s(\ell)+1)}{\Gamma((n+1)N_a^s + q_a)}. \tag{B.12}$$

À ce stade, on prend la limite $n \to 0$; pour cela on peut employer la formule

$$lim_{n\to 0}\Gamma(x+ny) = \Gamma(x)(1 + ny\,\psi(x)), \tag{B.13}$$

où ψ est la dérivée logarithmique de la fonction Γ, telle que $\frac{d\ln(\Gamma(x))}{dx} = \frac{\Gamma'(x)}{\Gamma(x)} = \psi(x)$. On obtient alors, après simplification des termes produits sur ℓ, le produit :

$$\lim_{n\to 0}(B.12) = \sum_s \prod_a \prod_\ell \frac{1 + nN_a^s(\ell)\,\psi(N_a^s(\ell)+1)}{1 + nN_a^s\,\psi(N_a^s + q_a)}. \tag{B.14}$$

Le développement au premier ordre en n de l'intégrale B.10 est donc :

$$1 + n\sum_s \sum_a \left(\sum_\ell \left(N_a^s(\ell)\,\psi\left(N_a^s(\ell)+1\right)\right) - N_a^s\,\psi(N_a^s + q_a) \right). \tag{B.15}$$

Finalement, en appliquant B.9, on obtient pour le critère à maximiser :

$$\begin{aligned}\langle \log \mathcal{L} \rangle_{post}(\{\mathcal{C}_s\}) = &\sum_{s=1}^{S} \sum_{a=1}^{A} \sum_{\ell=1}^{q_a} N_a^s(\ell)\Psi\left(1 + N_a^s(\ell)\right) \\ &- \sum_{s=1}^{S} \sum_{a=1}^{A} N_a^s \Psi\left(q_a + N_a^s\right)\ .\end{aligned} \tag{B.16}$$

Ce qui achève notre calcul. On voit que ce critère ne dépend que des comptes à l'intérieur de chaque groupe.

[1] Cette formule est nommée ainsi car elle a grandement été employée dans l'étude des verres de spin avec une technique dite de "répliques", voir par exemple Binder and Young (1986).

Maximisation de l'information acquise L'autre façon de voir les choses, à partir du moment où l'on a les deux distributions (B.4) et (B.7) est le suivant. L'écart entre ces deux distributions nous donne une mesure de l'influence des données, les gènes, sur les groupes observés. En effet, au départ, la structure supposée est homogène, caractérisée par la même distribution de probabilités *a priori* dans chaque groupe. Par la suite, le fait de classer les données d'une certaine façon modifie les distributions de probabilités à l'intérieur des groupes, par l'intermédiaire du calcul de la distribution *a posteriori*. Une façon de trouver la meilleure classification est donc de maximiser l'écart entre les distributions *a priori* et *a posteriori*, donc de maximiser l'information acquise par l'observation des gènes dans les groupes dans lesquels ils apparaissent. Ce gain d'information, une différence entre deux distributions de probabilités, peut naturellement être mesuré en utilisant la distance de Kullback-Leibler symmétrisée. On cherche donc maintenant à maximiser l'expression :

$$\frac{1}{2}\sum_s\sum_a\left(\int \mathcal{P}_{post}\log\left(\frac{\mathcal{P}_{post}}{\mathcal{P}_0}\right)\,dp_{s,a}(\ell)+\int \mathcal{P}_0\log\left(\frac{\mathcal{P}_0}{\mathcal{P}_{post}}\right)\,dp_{s,a}(\ell)\right). \tag{B.17}$$

On va voir comment cette expression nous ramène au même critère que précédemment. Chacune des deux intégrales peut être calculée en employant le "replica trick", qui va simplifier les logarithmes comme ci-dessus. Pour la première intégrale, on calcule donc tout d'abord :

$$\frac{\mathcal{P}_{post}^{n+1}}{\mathcal{P}_0^n}(p_{s,a}(\ell))=\frac{1}{(\Gamma(q_a))^n}\left(\frac{\Gamma(N_a^s+q_a)}{\prod_\ell\Gamma(N_a^s(\ell)+1)}\right)^{(n+1)}\prod_\ell p_{s,a}(\ell)^{(n+1)N_a^s(\ell)}\,\delta\left(\sum_p p_{s,a}(\ell)-1\right). \tag{B.18}$$

L'intégration de ce terme sur les $p_{s,a}(\ell)$ par la formule de Dirichlet, et la double sommation sur les acides aminés et sur les groupes nous donnent :

$$\sum_s\sum_a\frac{1}{(\Gamma(q_a))^n}\left(\frac{\Gamma(N_a^s+q_a)}{\prod_\ell\Gamma(N_a^s(\ell)+1)}\right)^{(n+1)}\frac{\prod_l\Gamma((n+1)N_a^s(\ell)+1)}{\Gamma((n+1)N_a^s+q_a)}. \tag{B.19}$$

La limite pour $n\to 0$ est calculée pour le terme de droite comme en (B.12). La limite des deux termes de gauche est calculée en utilisant l'égalité $\lim_{n\to0} f^n = 1+n\log(f)$, ce qui donne :

$$\lim_{n\to0}(B.19)=\sum_{s,a}\frac{1}{1+n\log\left(\Gamma(q_a)\right)}\,\frac{1+n\log\left(\Gamma(N_a^s+q_a)\right)}{\prod_l\,\left(1+n\log\left(\Gamma(N_a^s(l)+1)\right)\right)}\,\frac{\prod_\ell\left(1+nN_a^s(\ell)\,\psi(N_a^s(\ell)+1)\right)}{1+nN_a^s\,\psi(N_a^s+q_a)}. \tag{B.20}$$

Le terme de droite est celui qui va nous redonner le critère de partition déjà trouvé en (B.16). Nous allons simplement montrer maintenant que le développement au premier ordre des termes de gauche va s'annuler avec la deuxième intégrale dans (B.17). Ce développement est :

$$1-n\sum_{s,a}\left(\log\left(\Gamma(q_a)\right)-\log\left(\Gamma(N_a^s+q_a)\right)+\sum_\ell\log\left(\Gamma(N_a^s(l)+1)\right)\right). \tag{B.21}$$

Calculons maintenant la deuxième intégrale, et montrons qu'elle va annuler ce terme. On commence par calculer :

$$\frac{\mathcal{P}_0^{n+1}}{\mathcal{P}_{post}^n}=\frac{\Gamma(q_a)^{n+1}\prod_\ell\Gamma^n(N_a^s(\ell)+1)}{\Gamma^n(N_a^s+q_a)}\prod_\ell p_{s,a}^{-nN_a^s(\ell)}(\ell). \tag{B.22}$$

De la même manière que précédemment, l'intégrale va se décomposer en un terme non integré avec des fonctions Γ à la puissance n, et un terme provenant de l'intégration. La limite du terme non integré est :

$$\begin{aligned}&\lim_{n\to 0}\sum_{s,a}\left(\frac{\Gamma^n(q_a)}{\Gamma^n(N_a^s+q_a)}\prod_l\Gamma^n\left(N_a^s(\ell)+1\right)\right)\\ =&1+n\sum_{s,a}\left(\log(\Gamma(q_a))-\log(\Gamma(N_a^s+q_a))+\sum_\ell\log(N_a^s(\ell)+1)\right).\end{aligned} \tag{B.23}$$

Le terme facteur de n est exactement l'opposé du terme que l'on cherche à annuler, à savoir B.21. Il reste à ajouter la limite du terme venant de l'intégrale, qui est :

$$\lim_{n\to 0}\sum_s\sum_a\frac{\prod_l\Gamma(-nN_a^s(\ell)+1)}{\Gamma(-nN_a^s+q_a)}=\sum_s\sum_a\frac{\prod_\ell(1-nN_a^s(\ell))}{\Gamma(q_a)(1-nN_a^s\psi(q_a))}. \tag{B.24}$$

Il est facile de voir que ce second terme est une constante, ne dépendant que du nombre total de gènes à classer. La seconde intégrale est donc égale à (B.23), qui est bien l'opposé de (B.21). Ceci achève de démontrer comment on peut effectivement retrouver le critère (B.16) en se basant sur la maximisation de la distance entre la distribution *a priori* et la distribution *a posteriori*.

Bibliographie

Agris, P.F. 2004. Decoding the genome : a modified view. *Nucleic Acids Research* **32** : 223–238.

Agris, P.F., Vendeix, F.A.P. and Graham, W.D. 2007. tRNA's wobble decoding of the genome : 40 years of modification. *Journal of Molecular Biology* **366** : 1–13.

Akashi, H. 2001. Gene expression and molecular evolution. *Current Opinion in Genetics & Development* **11** : 660–666.

Akashi, H. and Eyre-Walker, A. 1998. Translational selection and molecular evolution. *Current Opinion in Genetics & Development* **8** : 688–693.

Akashi, H. and Gojobori, T. 2002. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proceedings of the National Academy of Sciences, USA* **99** : 3695–3700.

Altuvia, S., Weinstein-Fischer, D., Zhang, A., Postow, L. and Storz, G. 1997. A small, stable RNA induced by oxidative stress : role as a pleiotropic regulator and antimutator. *Cell* **90** : 43–53.

Ambrogelly, A., Palioura, S. and Soll, D. 2007. Natural expansion of the genetic code. *Nature Chemical Biology* **3** : 29–35.

Andersson, D.I., Verseveld, H.W., Stouthamer, A.H. and Kurland, C.G. 1986. Suboptimal growth with hyper-accurate ribosomes. *Archives of Microbiology* **144** : 96–101.

Andersson, S. and Kurland, C.G. 1990. Codon preferences in free-living microorganisms. *Microbiology and Molecular Biology Reviews* **54** : 198–210.

Archetti, M. 2004a. Genetic robusteness and selection at the protein level for synonymous codons. *Journal of Evolutionary Biology* **9** : 353–365.

Archetti, M. 2004b. Selection on codon usage for error minimization at the protein level. *Journal of Molecular Evolution* **59** : 400–415.

Bacher, J.M. and Schimmel, P. 2007. An editing-defective aminoacyl-tRNA synthetase is mutagenic in aging bacteria via the sos response. *Proceedings of the National Academy of Sciences, USA* **104** : 1907–1912.

Batchelor, G. 1970. *An introduction to fluid dynamics*. Cambridge Press University.

Baudouin-Cornu, P., Surdin-Kerjan, Y., Marliere, P. and Thomas, D. 2001. Molecular Evolution of Protein Atomic Composition. *Science* **293** : 297–300.

Bennetzen, J. and Hall, B. 1982. Codon selection in yeast. *Journal of Biological Chemistry* **257** : 3026–3031.

Berg, J., Tymoczko, J. and Stryer, L. 2002. *Biochemistry.* W. H. Freeman and Company, New York.

Beyer, A. 1997. Sequence analysis of the AAA protein family. *Protein Science* **6** : 2043–2058.

Bilgin, N., Ehrenberg, M. and Kurland, C. 1988. Is translation inhibited by noncognate ternary complexes ? *FEBS Letters* **233** : 95–99.

Binder, K. and Young, A. 1986. Spin glasses : experimental facts, theoretical concepts and open questions. *Reviews of Modern Physics* **58** : 801–976.

Blahut, R. 1972. Computation of channel capacity and rate distorsion function. *IEEE Transactions on Information Theory* **IT18** : 460–473.

Blattner, F.R., Plunkett, Guy, I., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F. et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277** : 1453–1462.

Blochl, E., Rachel, R., Burggraf, S., Hafenbradl, D., Jannasch, H.W. and Stetter, K.O. 1997. *Pyrolobus fumarii*, gen. and sp. nov., represents a novel group of Archaea, extending the upper temperature limit for life to 113 °C. *Extremophiles* **1** : 14–21.

Bock, A., Faiman, L.E. and Neidhardt, F.C. 1966. Biochemical and genetic characterization of a mutant of *Escherichia coli* with a temperature-sensitive valyl ribonucleic acid synthetase. *Journal of Bacteriology* **92** : 1076–1082.

Boone, D., Liu, Y., Zhao, Z., Balkwill, D., Drake, G., Stevens, T. and Aldrich, H. 1995. *Bacillus infernus* sp. nov., an Fe(III)- and Mn(IV)-reducing anaerobe from the deep terrestrial subsurface. *International Journal of Systematic and Evolutionary Microbiology* **45** : 441–448.

Bouma, J.E. and Lenski, R.E. 1988. Evolution of a bacteria/plasmid association. *Nature* **335** : 351–352.

Brewer, B.J. 1988. When polymerases collide : Replication and the transcriptional organization of the *E. coli* chromosome. *Cell* **53** : 679–686.

Brown, T.A. 2007. *Genomes 3.* Garland Science Publishing, New York.

Brussow, H., Canchaya, C. and Hardt, W.D. 2004. Phages and the evolution of bacterial pathogens : from genomic rearrangements to lysogenic conversion. *Microbiology and Molecular Biology Reviews* **68** : 560–602.

Brussow, H. and Hendrix, R.W. 2002. Phage genomics : Small is beautiful. *Cell* **108** : 13–16.

Bulmer, M. 1987a. Codon usage and intragenic position. *Journal of Theoretical Biology* **133** : 67–71.

Bulmer, M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129** : 897–907.

Bulmer, M. 1987b. Coevolution of codon usage and transfer rna abundance. *Nature* **325** : 728–730.

Calendar, R.L. (ed.) 2005. *The Bacteriophages.* Oxford University Press.

Campbell, A.M. 1992. Chromosomal insertion sites for phages and plasmids. *Journal of Bacteriology* **174** : 7495–7499.

Campbell, A.M. 2002. Preferential orientation of natural lambdoid prophages and bacterial chromosome organization. *Theoretical Population Biology* **61** : 503–507.

Canchaya, C., Desiere, F., McShan, W.M., Ferretti, J.J., Parkhill, J. and Brussow, H. 2002. Genome analysis of an inducible prophage and prophage remnants integrated in the *Streptococcus pyogenes* strain sf370. *Virology* **302** : 245–258.

Carafa, Y.d., Brody, E. and Thermes, C. 1990. Prediction of rho-independent *Escherichia coli* transcription terminators : A statistical analysis of their rna stem-loop structures. *Journal of Molecular Biology* **216** : 835–858.

Carbone, A., Kepes, F. and Zinovyev, A. 2005. Codon bias signatures, organization of microorganisms in codon space, and lifestyle. *Molecular Biology and Evolution* **22** : 547–561.

Carbone, A., Zinovyev, A. and Kepes, F. 2003. Codon adaptation index as a measure of dominating codon bias. *Bioinformatics* **19** : 2005–2015.

Casjens, S. 2003. Prophages and bacterial genomics : what have we learned so far ? *Molecular Microbiology* **49** : 277–300.

Cavalcanti, A.R.O., Leite, E.S., Neto, B.B. and Ferreira, R. 2004. On the classes of aminoacyl-tRNA synthetases, amino acids and the genetic code. *Origins of Life and Evolution of Biospheres* **34** : 407–420.

Charles, H., Calevro, F., Vinuelas, J., Fayard, J.M. and Rahbe, Y. 2006. Codon usage bias and tRNA over-expression in *Buchnera aphidicola* after aromatic amino acid nutritional stress on its host *Acyrthosiphon pisum*. *Nucleic Acids Research* **34** : 4583–4592.

Chechetkin, V.R. 2006. Genetic code from tRNA point of view. *Journal of Theoretical Biology* **242** : 922–934.

Chen, S.L., Lee, W., Hottes, A.K., Shapiro, L. and McAdams, H.H. 2004. Codon usage between genomes is constrained by genome-wide mutational processes. *Proceedings of the National Academy of Sciences, USA* **101** : 3480–3485.

Chou, T. 2003. Ribosome recycling, diffusion, and mRNA loop formation in translational regulation. *Biophysical Journal* **85** : 755–773.

Cilibrasi, R. and Vitanyi, P.M.B. 2005. Clustering by compression. *IEEE Transactions on Information Theory* **51** : 1523–1545.

Claverie, J.M. 2006. Viruses take center stage in cellular evolution. *Genome Biology* **7**.

Cole, S.T., Eiglmeier, K., Parkhill, J., James, K.D., Thomson, N.R., Wheeler, P.R., Honore, N., Garnier, T., Churcher, C., Harris, D. et al. 2001. Massive gene decay in the leprosy bacillus. *Nature* **409** : 1007–1011.

Cooper, G.M. 1999. *La Cellule : Une approche moléculaire*. De Boeck Université.

Cover, T.M. and Thomas, J.A. 1991. *Elements of Information Theory*. J. Wiley & sons, Inc., City College of New York.

Cowe, E. and Sharp, P.M. 1991. Molecular evolution of bacteriophages : Discrete patterns of codon usage in T4 genes are related to the time of gene expression. *Journal of Molecular Evolution* **V33** : 13–22.

Craigen, W.J. and Caskey, C.T. 1986. Expression of peptide chain release factor 2 requires high-efficiency frameshift. *Nature* **322** : 273–275.

Crick, F. 1965. Codon-anticodon pairing : the wooble hypothesis. *Journal of Molecular Biology* **19** : 548–555.

Crick, F. 1966. The genetic code : Yesterday, today and tomorrow. In *Cold Spring Harbor Symp. Quant. Biol.*, vol. 31. 3–9.

Cristianini, N. and Shawe-Taylor, J. 2000. *An introduction to Support Vector Machines*. Cambridge Press University.

Dagan, T. and Graur, D. 2005. The comparative method rules ! Codon volatility cannot detect positive darwinian selection using a single genome sequence. *Molecular Biology and Evolution* **22** : 496–500.

Danchin, A. 1990. *Une aurore de pierres*. Editions du Seuil.

Danchin, A. 2007. Archives or palimpsests ? bacterial genomes unveil a scenario for the origin of life. *Biological Theory* **2** : 52–61.

Daubin, V. and Ochman, H. 2004. Start-up entities in the origin of new genes. *Current Opinion in Genetics & Development* **14** : 616–619.

De Paepe, M. and Taddei, F. 2006. Viruses' life history : Towards a mechanistic basis of a trade-off between survival and reproduction among phages. *PLoS Biology* **4**.

Denamur, E. and Matic, I. 2006. Evolution of mutation rates in bacteria. *Molecular Microbiology* **60** : 820–827.

Dethlefsen, L. and Schmidt, T. 2005. Differences in codon bias cannot explain differences in translational power among microbes. *BMC Bioinformatics* **6** : 3.

d'Herelle, F. 1922. *The Bacteriophage ; its role in immunity*. Williams and Wilkins, Baltimore.

Di Giulio, M. 2007. The universal ancestor and the ancestors of Archaea and Bacteria were anaerobes whereas the ancestor of the Eukarya domain was an aerobe. *Journal of Evolutionary Biology* **20** : 543–548.

Di Giulio, M. 2000. The late stage of genetic code structuring took place at a high temperature. *Gene* **261** : 189–195.

Di Giulio, M. 2005a. The origin of the genetic code : theories and their relationships, a review. *Biosystems* **80** : 175–184.

Di Giulio, M. 2005b. Structuring of the genetic code took place at acidic ph. *Journal of Theoretical Biology* **237** : 219–226.

Dittmar, K.A., Sørensen, M.A., Elf, J., Ehrenberg, M. and Pan, T. 2005. Selective charging of tRNA isoacceptors induced by amino-acid starvation. *EMBO Reports* **6** : 121–157.

Dobrindt, U., Blum-Oehler, G., Nagy, G., Schneider, G., Johann, A., Gottschalk, G. and Hacker, J. 2002. Genetic structure and distribution of four pathogenicity islands (PAI I536 to PAI IV536) of uropathogenic *Escherichia coli* strain 536. *Infection and Immunity* **70** : 6365–6372.

Dong, H., Nilsson, L. and Kurland, C.G. 1996. Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *Journal of Molecular Biology* **260** : 649–663.

Doolittle, W.F. 1999. Phylogenetic classification and the universal tree. *Science* **284** : 2124–2128.

Doring, V., Mootz, H.D., Nangle, L.A., Hendrickson, T.L., de Crecy-Lagard, V., Schimmel, P. and Marliere, P. 2001. Enlarging the Amino Acid Set of Escherichia coli by Infiltration of the Valine Coding Pathway. *Science* **292** : 501–504.

Dubnau, D. 1991. Genetic competence in *Bacillus subtilis*. *Microbiology and Molecular Biology Reviews* **55** : 395–424.

Eagon, R.G. 1962. *Pseudomonas natriegens*, a marine bacterium with a generation time of less than 10 minutes. *Journal of Bacteriology* **83** : 736–737.

Eddy, S.R. 1999. Noncoding RNA genes. *Current Opinion in Genetics & Development* **9** : 695–699.

Elf, J., Nilsson, D., Tenson, T. and Ehrenberg, M. 2003. Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science* **300** : 1718–1722.

Ellis, R.J. 2006. Protein folding : Inside the cage. *Nature* **442** : 360–362.

Filee, J., Siguier, P. and Chandler, M. 2007. I am what I eat and I eat what I am : acquisition of bacterial genes by giant viruses. *Trends in Genetics* **23** : 10–15.

Finlay, B. and Falkow, S. 1997. Common themes in microbial pathogenicity revisited. *Microbiology and Molecular Biology Reviews* **61** : 136–169.

Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B., Merrick, J. et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269** : 496–512.

Forterre, P. 2006. Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes : A hypothesis for the origin of cellular domain. *Proceedings of the National Academy of Sciences, USA* **103** : 3669–3674.

Fox, G.E., Magrum, L.J., Balch, W.E., Wolfe, R.S. and Woese, C.R. 1977. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proceedings of the National Academy of Sciences, USA* **74** : 4537–4541.

Fox, T.D. 1987. Natural variation in the genetic code. *Annual Review of Genetics* **21** : 67–91.

Francino, M.P. and Ochman, H. 1997. Strand asymmetries in DNA evolution. *Trends in Genetics* **13** : 240–245.

Frank, A.C. and Lobry, J.R. 1999. Asymmetric substitution patterns : a review of possible underlying mutational or selective mechanisms. *Gene* **238** : 65–77.

Frederico, L.A., Kunkel, T.A. and Shaw, B.R. 1990. A sensitive genetic assay for the detection of cytosine deamination : determination of rate constants and the activation energy. *Biochemistry* **29** : 2532–2537.

Freeland, S.J. and Hurst, L.D. 1998. The genetic code is one in a million. *Journal of Molecular Evolution* **47** : 238–248.

French, S. 1992. Consequences of replication fork movement through transcription units in vivo. *Science* **258** : 1362–5.

Friedman, R. and Hughes, A.L. 2005. Codon volatility as an indicator of positive selection : Data from eukaryotic genome comparisons. *Molecular Biology and Evolution* **22** : 542–546.

Fuller, D.N., Rickgauer, J.P., Jardine, P.J., Grimes, S., Anderson, D.L. and Smith, D.E. 2007. From the cover : Ionic effects on viral dna packaging and portal motor function in bacteriophage {varphi}29. *Proceedings of the National Academy of Sciences* **104** : 11245–11250.

Galtier, N. and Lobry, J.R. 1997. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *Journal of Molecular Evolution* **44** : 632–636.

Gamow, G. 1954. Possible relation between deoxyribonucleic acid and protein structures. *Nature* **173** : 318–318.

Gautier, C., Gouy, M. and Louail, S. 1985. Non-parametric statistics for nucleic acid sequence study. *Biochimie* **67** : 449–453.

Ghedin, E. and Claverie, J.M. 2005. Mimivirus relatives in the Sargasso sea. *Virology Journal* **2** : 62–67.

Giglione, C., Boularot, A. and Meinnel, T. 2004. Protein N-terminal methionine excision. *Cellular and Molecular Life Sciences* **61** : 1455–1474.

Gilchrist, M.A. and Wagner, A. 2006. A model of protein translation including codon bias, nonsense errors, and ribosome recycling. *Journal of Theoretical Biology* **239** : 417–434.

Gladitz, J., Shen, K., Antalis, P., Hu, F.Z., Post, J.C. and Ehrlich, G.D. 2005. Codon usage comparison of novel genes in clinical isolates of *Haemophilus influenzae*. *Nucleic Acids Research* **33** : 3644–3658.

Gollnick, P. and Babitzke, P. 2002. Transcription attenuation. *Biochimica et Biophysica Acta - Gene Structure and Expression* **1577** : 240–250.

Goodarzi, H., Katanforoush, A., Torabi, N. and Najafabadi, H.S. 2007. Solvent accessibility, residue charge and residue volume, the three ingredients of a robust amino acid substitution matrix. *Journal of Theoretical Biology* **245** : 715–725.

Goodarzi, H., Najafabadi, H.S., Hassani, K., Nejad, H.A. and Torabi, N. 2005. On the optimality of the genetic code, with the consideration of coevolution theory by comparison of prominent cost measure matrices. *Journal of Theoretical Biology* **235** : 318–325.

Gordon, A.D. 1999. *Classification*. Chapman and Hall/CRC Press.

Gottesman, S. 2004. The small RNA regulators of *ESCHERICHIA COLI* : Roles and mechanisms*. *Annual Review of Microbiology* **58** : 303–328.

Gouy, M. and Gautier, C. 1982. Codon usage in bacteria : correlation with gene expressivity. *Nucleic Acids Research* **10** : 7055–7074.

Gouy, M., Gautier, C. and Milleret, F. 1985. System analysis and nucleic acid sequence banks. *Biochimie* **67** : 433–436.

Gouy, M. and Grantham, R. 1980. Polypeptide elongation and tRNA cycling in *Escherichia coli* : A dynamic approach. *FEBS Letters* **115** : 151–155.

Granick, S. 1957. Speculations on the origin and evolution of photosynthesis. *Annals of the New York Academy of Sciences* **69** : 292–308.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Research* **9** : r43–74.

Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pavé, A. 1980. Codon catalog usage and the genome hypothesis. *Nucleic Acids Research* **8** : r49–r62.

Gribskov, M., Devereux, J. and Burgess, R.R. 1984. The codon preference plot : graphic analysis of protein coding sequences and prediction of gene expression. *Nucleic Acids Research* **12** : 539–549.

Groeneveld, H., Oudot, F. and van Duin, J. 1996. RNA phage KU1 has an insertion of 18 nucleotides in the start codon of its lysis gene. *Virology* **218** : 141–147.

Grosjean, H. and Fiers, W. 1982. Preferential codon usage in prokaryotic genes : the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene* **18** : 199–209.

Grunwald, P. and Vitanyi, P. 2004. Shannon information and kolmogorov complexity.

Gultepe, E. and Kurnaz, M. 2005. Monte-carlo simulation and statistical analysis of genetic information coding. *Physica A* **357** : 525–533.

Guy, L. and Roten, C.A.H. 2004. Genometric analyses of the organization of circular chromosomes : a universal pressure determines the direction of ribosomal rna genes transcription relative to chromosome replication. *Gene* **340** : 45–52.

Haebel, P.W., Gutmann, S. and Ban, N. 2004. Dial tm for rescue : tmRNA engages ribosomes stalled on defective mRNAs. *Current Opinion in Structural Biology* **14** : 58–65.

Haig, D. and Hurst, L.D. 1991. A quantitative measure of error minimization in the genetic code. *Journal of Molecular Evolution* **33** : 412–417.

Hallier, M., Desreac, J. and Felden, B. 2006. Small protein b interacts with the large and the small subunits of a stalled ribosome during trans-translation. *Nucleic Acids Research* **34** : 1935–1943.

Häusler, T. 2006. *Viruses vs Superbugs : A Solution to the antibiotics crisis ?* MacMillan, NY.

Hendrix, R.W. 1998. Bacteriophage DNA packaging : RNA gears in a DNA transport machine. *Cell* **94** : 147–150.

Hendrix, R.W. 2002. Bacteriophages : Evolution of the majority. *Theoretical Population Biology* **61** : 471–480.

Hendrix, R.W. 2003. Bacteriophage genomics. *Current Opinion in Microbiology* **6** : 506–511.

Hendrix, R.W., Lawrence, J.G., Hatfull, G.F. and Casjens, S. 2000. The origins and ongoing evolution of viruses. *Trends in Microbiology* **8** : 504–508.

Hohn, M.J., Park, H.S., O'Donoghue, P., Schnitzbauer, M. and Soll, D. 2006. Emergence of the universal genetic code imprinted in an RNA record. *Proceedings of the National Academy of Sciences of the USA* **103** : 18095–18100.

Humphrey, S., Stanton, T., Jensen, N. and Zuerner, R. 1997. Purification and characterization of VSH-1, a generalized transducing bacteriophage of *Serpulina hyodysenteriae*. *Journal of Bacteriology* **179** : 323–329.

Hurst, L.D. and Merchant, A.R. 2001. High guanine-cytosine content is not an adaptation to high temperature : a comparative analysis amongst prokaryotes. *Proceedings of the Royal Society of London, Series B : Biological Sciences* **268** : 493–497.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution* **2** : 13–34.

Ikemura, T. 1981a. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes : A proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *Journal of Molecular Biology* **151** : 389–409.

Ikemura, T. 1981b. Correlation between the abundance of *Escherichia coli* tRNAs and the occurrence of the respective codons in its protein genes. *Journal of Molecular Biology* **146** : 1–21.

Ikemura, T. 1982. Correlation between the abundance of yeast tRNAs and the occurrence of the respective codons in protein genes : Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting tRNAs. *Journal of Molecular Biology* **158** : 573–597.

Ingraham, J., Maaloe, O. and Neidhardt, F. 1983. *Growth of the Bacterial Cell*. Sinauer, Sunderland, Mass.

Itzkovitz, S. and Alon, U. 2007. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Research* **17** : 405–412.

Jacob, F. and Monod, J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology* **3** : 318.

Jain, R., Rivera, M.C. and Lake, J.A. 1999. Horizontal gene transfer among genomes : The complexity hypothesis. *Proceedings of the National Academy of Sciences, USA* **96** : 3801–3806.

Jaynes, E. 1967. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics* **SSC-4** : 227–241.

Jensen, R.B. and Gerdes, K. 1995. Programmed cell death in bacteria : proteic plasmid stabilization systems. *Molecular Microbiology* **17** : 205–210.

Jia, M. and Li, Y. 2005. The relationship among gene expression, folding free energy and codon usage bias in *Escherichia coli*. *FEBS Letters* **579** : 5333–5337.

Jones, L.M., McDuff, C.R. and Wilson, J.B. 1962. Phenotypic alterations in the colonial morphology of *Brucella abortus* due to a bacteriophage carrier state. *Journal of Bacteriology* **83** : 860–866.

Juhala, R.J., Ford, M.E., Duda, R.L., Youlton, A., Hatfull, G.F. and Hendrix, R.W. 2000. Genomic sequences of bacteriophages HK97 and HK022 : pervasive genetic mosaicism in the lambdoid bacteriophages. *Journal of Molecular Biology* **299** : 27–51.

Kaiser, C.M., Chang, H.C., Agashe, V.R., Lakshmipathy, S.K., Etchells, S.A., Hayer-Hartl, M., Hartl, F.U. and Barral, J.M. 2006. Real-time observation of trigger factor function on translating ribosomes. *Nature* **444** : 455–460.

Kan, J., Kano-Sueoka, T. and Sueoka, N. 1968. Characterization of leucine transfer ribonucleic acid in escherichia coli following infection with bacteriophage t2. *Journal of Biological Chemistry* **243** : 5584–5590.

Kanaya, S., Yamada, Y., Kudo, Y. and Ikemura, T. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs : gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* **238** : 143–155.

Kano-Sueoka, T., Lobry, J.R. and Sueoka, N. 1999. Intra-strand biases in bacteriophage T4 genome. *Gene* **238** : 59–64.

Kano-Sueoka, T. and Sueoka, N. 1969. Leucine trna and cessation of escherichia coli protein synthesis upon phage t2 infection. *Proceedings of the National Academy of Sciences, USA* **62** : 1229–1236.

Kapp, L.D. and Lorsch, J.R. 2004. The molecular mechanics of eukaryotic translation. *Annual Review of Biochemistry* **73** : 657–704.

Karlin, S. 2001. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends in Microbiology* **9** : 335–343.

Karlin, S. and Mrazek, J. 2000. Predicted highly expressed genes of diverse prokaryotic genomes. *Journal of Bacteriology* **182** : 5238–5250.

Karlin, S., Mrazek, J. and Campbell, A.M. 1998. Codon usages in different gene classes of the *Escherichia coli* genome. *Molecular Microbiology* **29** : 1341–1355.

Kimura, M. 1968. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetical Research* **11** : 247–269.

King, J. and Jukes, T. 1969. Non-darwinian evolution. *Science* **164** : 788–798.

Kirnos, M.D., Khudyakov, I.Y., Alexandrushkina, N.I. and Vanyushin, B.F. 1977. 2-aminoadenine is an adenine substituting for a base in s-2l cyanophage dna. *Nature* **270** : 369–370.

Knight, R., Freeland, S. and Landweber, L. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biology* **2** : research0010.1–research0010.13.

Knopf, C.W. 1998. Evolution of viral DNA-dependent DNA polymerases. *Virus Genes* **16** : 47–58.

Koonin, E., Senkevich, T. and Dolja, V. 2006. The ancient virus world and evolution of cells. *Biology Direct* **1**.

Koonin, E.V. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Reviews : Microbiology* **1** : 127–136.

Koonin, E.V. and Galperin, M.Y. 2002. *Sequence. Evolution. Function. Computational approachs in comparative genomics.* Kluwer Academic Publisher.

Koonin, E.V., Makarova, K.S. and Aravind, L. 2001. Horizontal gene transfer in prokaryotes : Quantification and classification. *Annual Review of Microbiology* **55** : 709–742.

Kozak, M. 2005. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* **361** : 13–37.

Kreitman, M. and Comeron, J.M. 1999. Coding sequence evolution. *Current Opinion in Genetics & Development* **9** : 637–641.

Kropinski, A.M. and Sibbald, M.J. 1999. Transfer RNA genes and their significance to codon usage in the *Pseudomonas aeurginosa* lambdoid bacteriophage D3. *Canadian Journal of Microbiology* **45** : 791–796.

Kunisawa, T. 1992. Synonymous codon preferences in bacteriophage T4 : A distinctive use of transfer RNAs from T4 and from its host *Escherichia coli*. *Journal of Theoretical Biology* **159** : 287–298.

Kunisawa, T., Kanaya, S. and Kutter, E. 1998. Comparison of synonymous codon distribution patterns of bacteriophage and host genomes. *DNA Research* **5** : 319–326.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessieres, P., Bolotin, A., Borchert, S. et al. 1997. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390** : 249–256.

Kurland, C.G. 1991. Codon bias and gene expression. *FEBS Letters* **285** : 165–169.

Landau, L. and Lifshitz, E. 1959. *Mecanique des fluides.* Pergamon.

Lawrence, J.G., Ochman, H. and Hartl, D.L. 1992. The evolution of insertion sequences within enteric Bacteria. *Genetics* **131** : 9–20.

Lawrence, J.G., Hendrix, R.W. and Casjens, S. 2001. Where are the pseudogenes in bacterial genomes ? *Trends in Microbiology* **9** : 535–540.

Lee, S.Y., Bailey, S.C. and Apirion, D. 1978. Small stable RNAs from *Escherichia coli* : Evidence for the existence of new molecules and for a new ribonucleoprotein particle containing 6S RNA. *Journal of Bacteriology* **133** : 1015–1023.

Levine, R.L., Mosoni, L., Berlett, B.S. and Stadtman, E.R. 1996. Methionine residues as endogenous antioxidants in proteins. *Proceedings of the National Academy of Sciences of the USA* **93** : 15036–15040.

Lewin, B. 2004. *Genes VIII.* Pearson Prentice Hall, Upper Saddle River.

Lewis, P.J., Thaker, S.D. and Errington, J. 2000. Compartimentalization of transcription and translation in *Bacillus subtilis*. *EMBO Journal* **19** : 710–718.

Li, M., Chen, X., Li, X., Ma, B. and Vitanyi, P. 2003. The similarity metric. In *Proceedings of the 14th annual ACM-SIAM Symposium on discrete Algorithms*. 863–872.

Liljenstrom, H. and von Heijne, G. 1987. Translation rate modification by preferential codon usage : Intragenic position effects. *Journal of Theoretical Biology* **124** : 43–55.

Lin, J. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* **37** : 145–151.

Link, A.J., Mock, M.L. and Tirrell, D.A. 2003. Non-canonical amino acids in protein engineering. *Current Opinion in Biotechnology* **14** : 603–609.

Lobanov, A.V., Kryukov, G.V., Hatfield, D.L. and Gladyshev, V.N. 2006. Is there a twenty third amino acid in the genetic code ? *Trends in Genetics* **22** : 357–360.

Lobry, J.R. 1996a. Origin of replication of *Mycoplasma genitalium*. *Science* **272** : 745–646.

Lobry, J.R. and Gautier, C. 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Research* **22** : 3174–3180.

Lobry, J. 1996b. Asymmetric substitution patterns in the two DNA strands of bacteria. *Molecular Biology and Evolution* **13** : 660–665.

Lu, Y. and Freeland, S. 2006. On the evolution of the standard amino-acid alphabet. *Genome Biology* **7**.

Ma, P., Castillo-Davis, C.I., Zhong, W. and Liu, J.S. 2006. A data-driven clustering method for time course gene expression data. *Nucleic Acids Research* **34** : 1261–1269.

Mallick, B., Chakrabarti, J., Sahoo, S., Ghosh, Z. and Das, S. 2005. Identity elements of archaeal tRNA. *DNA Research* **12** : 235–246.

Man, O. and Pilpel, Y. 2007. Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. *Nature Genetics* **39** : 415–421.

Marck, C. and Grosjean, H. 2002. tRNomics : analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA* **8** : 1189–1232.

Marquez, R., Smit, S. and Knight, R. 2005. Do universal codon-usage patterns minimize the effects of mutation and translation error ? *Genome Biology* **6**.

Masashi, F., Hisaaki, M., Susumu, G., Nobuyoshi, E. and Minoru, K. 2007. Mining prokaryotic genomes for unknown amino acids : a stop-codon-based approach. *BMC Bioinformatics* **8** : 225.

Mascarenhas, J., Weber, M.H.W. and Graumann, P.L. 2001. Specific polar localization of ribosomes in *Bacillus subtilis* depends on active transcription. *EMBO Reports* **2** : 685–689.

McInerney, J.O. 1998. Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. *Proceedings of the National Academy of Sciences of the USA* **95** : 10698–10703.

McLachlan, A.D. 1971. Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c551. *Journal of Molecular Biology* **61** : 409–424.

McLachlan, A.D., Staden, R. and Boswell, D.R. 1984. A method for measuring the non-random bias of a codon usage table. *Nucleic Acids Research* **12** : 9567–9575.

McLean, M.J., Wolfe, K.H. and Devine, K.M. 1998. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *Journal of Molecular Evolution* **V47** : 691–696.

McNulty, D.E., Claffee, B.A., Huddleston, M.J. and Kane, J.F. 2003. Mistranslational errors associated with the rare arginine codon CGG in *Escherichia coli*. *Protein Expression and Purification* **27** : 365–374.

Médigue, C., Rouxel, T., Vigier, P., Hénaut, A. and Danchin, A. 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. *Journal of Molecular Biology* **222** : 851–856.

Médigue, C., Krin, E., Pascal, G., Barbe, V., Bernsel, A., Bertin, P.N., Cheung, F., Cruveiller, S., D'Amico, S., Duilio, A. et al. 2005. Coping with cold : The genome of the versatile marine Antarctica bacterium *Pseudoalteromonas haloplanktis* TAC125. *Genome Research* **15** : 1325–1335.

Medrano-Soto, A., Moreno-Hagelsieb, G., Vinuesa, P., Christen, J.A. and Collado-Vides, J. 2004. Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes. *Molecular Biology and Evolution* **21** : 1884–1894.

Merkl, R. 2003. A survey of codon and amino acid frequency bias in microbial genomes focusing on translational efficiency. *Journal of Molecular Evolution* **57** : 453–466.

Mézard, M. 2007. Computer science : Where are the exemplars ? *Science* **315** : 949–951.

Miller, E.S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T. and Ruger, W. 2003. Bacteriophage T4 genome. *Microbiology and Molecular Biology Reviews* **67** : 86–156.

Miller, J. and Reznikoff, W. (eds.) 1978. *The Operon*. Cold Spring Harbor Symp Quant Biol.

Miller, S.L. and Urey, H.C. 1959. Origin of life. *Science* **130** : 1622–1624.

Mira, A., Ochman, H. and Moran, N.A. 2001. Deletional bias and the evolution of bacterial genomes. *Trends in Genetics* **17** : 589–596.

Mirkin, E.V. and Mirkin, S.M. 2005. Mechanisms of transcription-replication collisions in bacteria. *Molecular and Cellular Biology* **25** : 888–895.

Mitra, K., Schaffitzel, C., Shaikh, T., Tama, F., Jenni, S., Brooks, C.L., Ban, N. and Frank, J. 2005. Structure of the *E. coli* protein-conducting channel bound to a translating ribosome. *Nature* **438** : 318–324.

Moore, S.D. and Sauer, R.T. 2005. Ribosome rescue : tmRNA tagging activity and capacity in escherichia coli. *Molecular Microbiology* **58** : 456–466.

Moriyama, E. and Powell, J. 1998. Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. *Nucleic Acids Research* **26** : 3188–3193.

Morris, D.W. and DeMoss, J.A. 1965. Role of aminoacyl-transfer ribonucleic acid in the regulation of ribonucleic acid synthesis in *Escherichia coli*. *Journal of Bacteriology* **90** : 1624–1631.

Moszer, I., Rocha, E.P. and Danchin, A. 1999. Codon usage and lateral gene transfer in *Bacillus subtilis*. *Current Opinion in Microbiology* **2** : 524–528.

Musto, H., Naya, H., Zavala, A., Romero, H., Alvarez-Valin, F. and Bernardi, G. 2006. Genomic GC level, optimal growth temperature, and genome size in prokaryotes. *Biochemical and Biophysical Research Communications* **347** : 1–3.

Nagaswamy, U. and Fox, G.E. 2003. RNA ligation and the origin of tRNA. *Origins of Life and Evolution of Biospheres* **33** : 199–209.

Narra, H.P. and Ochman, H. 2006. Of what use is sex to bacteria ? *Current Biology* **16** : R705–R710.

Naya, H., Romero, H., Zavala, A., Alvarez, B. and Musto, H. 2002. Aerobiosis increases the genomic guanine plus cytosine content (gc%) in prokaryotes. *Journal of Molecular Evolution* **55** : 260–264.

Neidhardt, F.C., Bloch, P.L., Pedersen, S. and reeh, S. 1977. Chemical measurement of steady-state levels of ten aminoacyl-transfer ribonucleic acid synthetases in *Escherichia coli*. *Journal of Bacteriology* **129** : 378–387.

Nikolaou, C. and Almirantis, Y. 2005. A study on the correlation of nucleotide skews and the positioning of the origin of replication : different modes of replication in bacterial species. *Nucleic Acids Research* **33** : 6816–6822.

Novembre, J.A. 2002. Accounting for background nucleotide composition when measuring codon usage bias. *Molecular Biology and Evolution* **19** : 1390–1394.

Oelschlaeger, T.A., Dobrindt, U. and Hacker, J. 2002. Pathogenicity islands of uropathogenic e. coli and the evolution of virulence. *International Journal of Antimicrobial Agents* **19** : 517–521.

Ohnishi, M., Tanaka, C., Kuhara, S., Ishii, K., Hattori, M., Kurokawa, K., Yasunaga, T., Makino, K., Shinagawa, H., Murata, T. et al. 1999. Chromosome of the enterohemorrhagic *Escherichia coli* O157 :H7 ; comparative analysis with K-12 MG1655 revealed the acquisition of a large amount of foreign DNAs. *DNA Research* **6** : 361–368.

Olsen, G.J. and Woese, C.R. 1997. Archaeal genomics : An overview. *Cell* **89** : 991–994.

Paddison, P., Abedon, S.T., Dressman, H.K., Gailbreath, K., Tracy, J., Mosser, E., Neitzel, J., Guttman, B. and Kutter, E. 1998. The roles of the bacteriophage T4 *r* genes in lysis inhibition and fine-structure genetics : A new perspective. *Genetics* **148** : 1539–1550.

Parkhill, J., Wren, B.W., Thomson, N.R., Titball, R.W., Holden, M.T.G., Prentice, M.B., Sebaihia, M., James, K.D., Churcher, C., Mungall, K.L. et al. 2001. Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413** : 523–527.

Pascal, G., Médigue, C. and Danchin, A. 2005. Universal biases in protein composition of model prokaryotes. *Proteins : Structure, Function and Genetics* **60** : 27–35.

Pascal, G., Médigue, C. and Danchin, A. 2006. Persistent biases in the amino acid composition of prokaryotic proteins. *Bioessays* **28** : 726–738.

Pedulla, M.L., Ford, M.E., Houtz, J.M., Karthikeyan, T., Wadsworth, C., Lewis, J.A., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N.R. et al. 2003. Origins of highly mosaic mycobacteriophage genomes. *Cell* **113** : 171–182.

Perrière, G. and Thioulouse, J. 2002. Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Research* **30** : 4548–4555.

Plotkin, J.B., Dushoff, J., Desai, M.M. and Fraser, H.B. 2006a. Codon usage and selection on proteins. *Journal of Molecular Evolution* **63** : 635–653.

Plotkin, J.B., Dushoff, J., Desai, M.M. and Fraser, H.B. 2006b. Estimating selection pressures from limited comparative data. *Molecular Biology and Evolution* **23** : 1457–1459.

Plotkin, J.B., Dushoff, J. and Fraser, H.B. 2004. Detecting selection using a single genome sequence of *M. tuberculosis* and *P. falciparum*. *Nature* **428** : 942–945.

Pluhar, W. 2006. AT2-AT3-profiling : A new look at synonymous codon usage. *Journal of Theoretical Biology* **243** : 308–321.

Prangishvili, D., Forterre, P. and Garrett, R.A. 2006. Viruses of the Archaea : a unifying view. *Nature Reviews : Microbiology* **4** : 837–848.

Prescott, Harley and Klein 2002. *Microbiologie*. De Boeck Université.

Qin, Y., Polacek, N., Vesper, O., Staub, E., Einfeldt, E., Wilson, D.N. and Nierhaus, K.H. 2006. The highly conserved LepA is a ribosomal elongation factor that back-translocates the ribosome. *Cell* **127** : 721–733.

Raoult, D., Audic, S., Robert, C., Abergel, C. and Renesto, P. 2004. The 1.2-megabase genome sequence of Mimivirus. *Science* **306** : 1344–1350.

Reis, M.d., Savva, R. and Wernisch, L. 2004. Solving the riddle of codon usage preferences : a test for translational selection. *Nucleic Acids Research* **32** : 5036–5044.

Riley, M. 1993. Functions of the gene products of *Escherichia coli*. *Microbiology and Molecular Biology Reviews* **57** : 862–952.

Rocha, E.P.C. 2002. Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes ? *Trends in Microbiology* **10** : 393–395.

Rocha, E.P.C. 2006. The quest for the universals of protein evolution. *Trends in Genetics* **22** : 412–416.

Rocha, E.P.C. and Danchin, A. 2002. Base composition bias might result from competition for metabolic resources. *Trends in Genetics* **18** : 291–294.

Rocha, E.P.C. and Danchin, A. 2003a. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nature Genetics* **34** : 377–378.

Rocha, E.P.C. and Danchin, A. 2003b. Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Research* **31** : 6570–6577.

Rocha, E.P. 2004. Codon usage bias from tRNA's point of view : Redundancy, specialization, and efficient decoding for translation optimization. *Genome Research* **14** : 2279–2286.

Rocha, E.P., Danchin, A. and Viari, A. 1999. Universal replication biases in bacteria. *Molecular Microbiology* **32** : 11–16.

Rocha, E.P., Touchon, M. and Feil, E.J. 2006. Similar compositional biases are caused by very different mutational effects. *Genome Research* **16** : 1537–1547.

Roche, E.D. and Sauer, R.T. 1999. SsrA-mediated peptide tagging caused by rare codons and tRNA scarcity. *EMBO Journal* **18** : 4579–4589.

Rogers, M.J. and Soll, D. 1988. Discrimination between glutaminyl-tRNA synthetase and seryl-tRNA synthetase involves nucleotides in the acceptor helix of tRNA. *Proceedings of the National Academy of Sciences of the USA* **85** : 6627–6631.

Rose, K., Gurewitz, E. and Fox, G.C. 1990. Statistical mechanics and phase transitions in clustering. *Physical Review Letters* **65** : 945.

Roth, V., Lange, T., Braun, M. and Buhmann, J. 2004. Stability-based validation of clustering solutions. *Neural Computation* **16** : 1299–1323.

Saks, M.E., Sampson, J.R. and Abelson, J. 1998. Evolution of a transfer RNA gene through a point mutation in the anticodon. *Science* **279** : 1665–1670.

Saporta, G. 1990. *Probabilités, Analyse des données et Statistique*. Editions Technip.

Sau, K., Gupta, S.K., Sau, S. and Ghosh, T.C. 2005. Synonymous codon usage bias in 16 *Staphylococcus aureus* phages : Implication in phage therapy. *Virus Research* **113** : 123–131.

Sau, K., Gupta, S.K., Sau, S., Mandal, S.C. and Ghosh, T.C. 2006. Factors influencing synonymous codon and amino acid usage biases in Mimivirus. *Biosystems* **85** : 107–113.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* **6** : 461–464.

Sekowska, A. 1999. *Une rencontre du métabolisme du soufre et de l'azote : le métabolisme des polyamines chez* Bacillus subtilis. Ph.D. thesis, Université de Versailles-Saint Quentin en Yvelines.

Seligmann, H. and Pollock, D.D. 2004. The ambush hypothesis : Hidden stop codons prevent off-frame gene reading. *DNA and Cell Biology* **23** : 701–705.

Selosse, M.A., Albert, B. and Godelle, B. 2001. Reducing the genome size of organelles favours gene transfer to the nucleus. *Trends in Ecology & Evolution* **16** : 135–141.

Sengupta, S., Xiaoguang, Y. and Higgs, P. 2007. The mechanisms of codon reassignments in mitochondrial genetic codes.

Shabalina, S.A., Ogurtsov, A.Y. and Spiridonov, N.A. 2006. A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Research* **34** : 2428–2437.

Sharp, P.M. 2005. Gene "volatility" is most unlikely to reveal adaptation. *Molecular Biology and Evolution* **22** : 807–809.

Sharp, P.M., Bailes, E., Grocock, R.J., Peden, J.F. and Sockett, R.E. 2005. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Research* **33** : 1141–1153.

Sharp, P.M. and Li, W.H. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of Molecular Evolution* **V24** : 28–38.

Sharp, P.M., Rogers, M.S. and McConnell, D.J. 1985. Selection pressures on codon usage in the complete genome of bacteriophage T7. *Journal of Molecular Evolution* **21** : 150–160.

Sharp, P. and Li, W. 1987. The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research* **15** : 1281–1295.

Sharp, P., Tuohy, T. and Mosurski, K. 1986. Codon usage in yeast : cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research* **14** : 5152–5143.

Shrader, T.E., Tobias, J.W. and Varshavsky, A. 1993. The N-end rule in *Escherichia coli* : cloning and analysis of the leucyl, phenylalanyl-tRNA-protein transferase gene *aat*. *Journal of Bacteriology* **175** : 4364–4374.

Smyth, P. 2000. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing* **10** : 63.

Söll, D. and RajBhandary, U.L. (eds.) 1995. *tRNA : Structure, Biosynthesis and Function*. ASM Press, Washington, D.C.

Solomovici, J., Lesnik, T. and Reiss, C. 1997. Does *Escherichia coli* optimize the economics of the translation process ? *Journal of Theoretical Biology* **185** : 511–521.

Sorensen, M.A., Elf, J., Bouakaz, E., Tenson, T., Sanyal, S., Bjork, G.R. and Ehrenberg, M. 2005. Over expression of a $tRNA^{Leu}$ isoacceptor changes charging pattern of leucine tRNAs and reveals new codon reading. *Journal of Molecular Biology* **354** : 16–24.

Sorensen, M.A., Kurland, C.G. and Pedersen, S. 1989. Codon usage determines translation rate in *Escherichia coli*. *Journal of Molecular Biology* **207** : 365–377.

Sorensen, M.A. and Pedersen, S. 1991. Absolute in vivo translation rates of individual codons in *Escherichia coli* : The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. *Journal of Molecular Biology* **222** : 265–280.

Spanjaard, R.A. and Duin, J.V. 1988. Translation of the sequence AGG-AGG yields 50% ribosomal frameshift. *Proceedings of the National Academy of Sciences of the USA* **85** : 7967–7971.

Sprinzl, M. and Vassilenko, K. 2003. Compilation of tRNA sequences and sequences of tRNA genes.

States, D. and Gish, W. 1994. Combined use of sequence similarity and codon bias for coding region identification. *Journal of Computational Biology* **1** : 39–50.

Sternberg, N. and Hoess, R. 1983. The molecular genetics of bacteriophage p1. *Annual Review of Genetics* **17** : 123–154.

Still, S. and Bialek, W. 2004. How many clusters ? An information-theoretic perspective. *Neural Computation* **16** : 2483–2506.

Still, S., Bialek, W. and Bottou, L. 2004. Geometric clustering using the information bottleneck method. In *Advances in Neural information processing systems*, vol. 16. MIT Press.

Stoletzki, N. and Eyre-Walker, A. 2007. Synonymous codon usage in *Escherichia coli* : Selection for translational accuracy. *Molecular Biology and Evolution* **24** : 374–381.

Stoletzki, N., Welch, J., Hermisson, J. and Eyre-Walker, A. 2005. A dissection of volatility in yeast. *Molecular Biology and Evolution* **22** : 2022–2026.

Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Stat. Soc. Ser. B. (Stat. Method.)* **36** : 111.

Sueoka, N. 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *Journal of Molecular Evolution* **40** : 318–325.

Sueoka, N. 1962. On the genetic basis of variation and heterogeneity of dna base composition. *Proceedings of the National Academy of Sciences of the USA* **48** : 582–592.

Sueoka, N. 1988. Directional mutation pressure and neutral molecular evolution. *Proceedings of the National Academy of Sciences of the USA* **85** : 2653–2657.

Sueoka, N. 1992. Directional mutation pressure, selective constraints, and genetic equilibria. *Journal of Molecular Evolution* **34** : 95–114.

Sueoka, N. and Kano-Sueoka, T. 1964. A specific modification of leucyl-srna of escherichia coli after phage t2 infection. *Proceedings of the National Academy of Sciences* **52** : 1535–1540.

Suttle, C.A. 2005. Viruses in the sea. *Nature* **437** : 356–361.

Tang, Y.C., Chang, H.C., Roeben, A., Wischnewski, D., Wischnewski, N., Kerner, M.J., Hartl, F.U. and Hayer-Hartl, M. 2006. Structural features of the GroEL-GroES nano-cage required for rapid folding of encapsulated protein. *Cell* **125** : 903–914.

Tatusov, R., Fedorova, N., Jackson, J., Jacobs, A., Kiryutin, B., Koonin, E., Krylov, D., Mazumder, R., Mekhedov, S., Nikolskaya, A. et al. 2003. The COG database : an updated version includes eukaryotes. *BMC Bioinformatics* **4** : 41.

Tatusov, R.L., Galperin, M.Y., Natale, D.A. and Koonin, E.V. 2000. The COG database : a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research* **28** : 33–36.

Taylor, W.R. 2006. A molecular model for the origin of protein translation in an RNA world. *Journal of Theoretical Biology* **243** : 393–406.

Tekaia, F., Yeramian, E. and Dujon, B. 2002. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends : a global picture with correspondence analysis. *Gene* **297** : 51–60.

Thanaraj, T.A. and Argos, P. 1996a. Protein secondary structural types are differentially coded on messenger RNA. *Protein Science* **5** : 1973–1983.

Thanaraj, T.A. and Argos, P. 1996b. Ribosome-mediated translational pause and protein domain organization. *Protein Science* **5** : 1594–1612.

Thomas, L.K., Dix, D.B. and Thompson, R.C. 1988. Codon choice and gene expression : Synonymous codons differ in their ability to direct aminoacylated-transfer RNA binding to ribosomes in vitro. *Proceedings of the National Academy of Sciences of the USA* **85** : 4242–4246.

Tibshirani, R., Walther, G. and Hastie, T. 2001. Estimating the number of clusters in a dataset via the Gap statistic. *J. Roy. Stat. Soc. Ser. B. (Stat. Method.)* **63** : 411.

Tishby, N., Pereira, F.C. and Bialek, W. 1999. The information bottleneck method. In B. Hajek and R.S. Sreenivas (eds.), *Proceedings of the 37th annual Allerton Conference.*

Trifonov, E.N. 2004. The triplet code from first principles. *Journal of Biomolecular Structure and Dynamics* **22** : 1–12.

Tritton, D. 1977. *Physical fluid dynamics.* ELBS.

van Valen, L. 1973. A new evolutionary law. *Evolutionary Theory* **1** : 1–30.

Vetsigian, K., Woese, C. and Goldenfeld, N. 2006. Collective evolution and the genetic code. *Proceedings of the National Academy of Sciences of the USA* **103** : 10696–10701.

Vogl, C., Sanchez-Cabo, F., Stocker, G., Hubbard, S., Wolkenhauer, O. and Trajanoski, Z. 2005. A fully bayesian model to cluster gene-expression profiles. *Bioinformatics* **21** : ii130–136.

Wachtershauser, G. 1988. Before enzymes and templates : theory of surface metabolism. *Microbiology and Molecular Biology Reviews* **52** : 452–484.

Wächtershäuser, G. 2007. On the chemistry and evolution of the pioneer organism. *Chemistry & Biodiversity* **4** : 584–602.

Wang, H.C., Susko, E. and Roger, A.J. 2006. On the correlation between genomic G+C content and optimal growth temperature in prokaryotes : Data quality and confounding factors. *Biochemical and Biophysical Research Communications* **342** : 681–684.

Webb, J.S., Givskov, M. and Kjelleberg, S. 2003. Bacterial biofilms : prokaryotic adventures in multicellularity. *Current Opinion in Microbiology* **6** : 578–585.

Weinbauer, M.G. 2004. Ecology of prokaryotic viruses. *FEMS Microbiology Reviews* **28** : 127–181.

Welch, R.A., Burland, V., Plunkett, G., I., Redford, P., Roesch, P., Rasko, D., Buckles, E.L., Liou, S.R., Boutin, A., Hackett, J. et al. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proceedings of the National Academy of Sciences of the USA* **99** : 17020–17024.

Whitman, W.B., Coleman, D.C. and Wiebe, W.J. 1998. Prokaryotes : The unseen majority. *Proceedings of the National Academy of Sciences of the USA* **95** : 6578–6583.

Willenbrock, H., Friis, C., Juncker, A. and Ussery, D. 2006. An environmental signature for 323 microbial genomes based on codon adaptation indices. *Genome Biology* **7**.

Withers, M., Wernisch, L. and Reis, M.d. 2006. Archaeology and evolution of transfer RNA genes in the *Escherichia coli* genome. *RNA* **12** : 933–942.

Withey, J.H. and Friedman, D.I. 2003. A salvage pathway for protein synthesis : tmRNA and trans-translation. *Annual Review of Microbiology* **57** : 101–123.

Woese, C.R. 1987. Bacterial evolution. *Microbiology and Molecular Biology Reviews* **51** : 221–271.

Wommack, K.E. and Colwell, R.R. 2000. Virioplankton : Viruses in aquatic ecosystems. *Microbiology and Molecular Biology Reviews* **64** : 69–114.

Wong, J.T.F. 1975. A co-evolution theory of the genetic code. *Proceedings of the National Academy of Sciences, USA* **72** : 1909–1912.

Wong, J.T.F. 2005. Coevolution theory of the genetic code at age thirty. *Bioessays* **27** : 416–425.

Wright, F. 1990. The "effective number of codons" used in a gene. *Gene* **87** : 23–29.

Xia, X. 1996. Maximizing transcription efficiency causes codon usage bias. *Genetics* **144** : 1309–1320.

Xia, X. 1998. How optimized is the translational machinery in *Escherichia coli*, *Salmonella typhimurium* and *Saccharomyces cerevisiae* ? *Genetics* **149** : 37–44.

Yarus, M. 1982. Translational efficiency of transfer RNA's : uses of an extended anticodon. *Science* **218** : 646–652.

Yarus, M. 2002. Primordial genetics : Phenotype of the ribocyte. *Annual Review of Genetics* **36** : 125–151.

Yegian, C.D. and Stent, G.S. 1969. An unusual condition of leucine transfer RNA appearing during leucine starvation of *Escherichia coli*. *Journal of Molecular Biology* **39** : 45–58.

Yusupova, G., Jenner, L., Rees, B., Moras, D. and Yusupov, M. 2006. Structural basis for messenger RNA movement on the ribosome. *Nature* **444** : 391–394.

Zhang, J. 2005. On the evolution of codon volatility. *Genetics* **169** : 495–501.

Zhao, K.N., Liu, W.J. and Frazer, I.H. 2003. Codon usage bias and A+T content variation in human papillomavirus genomes. *Virus Research* **98** : 95–104.

Zwieb, C., Wower, I. and Wower, J. 1999. Comparative sequence analysis of tmRNA. *Nucleic Acids Research* **27** : 2063–2071.

Résumé/Abstract

Cette thèse regroupe des travaux concernant le biais d'usage de codons et son rôle chez les bactéries et leurs phages, en particulier sur les processus de traduction et l'organisation des génomes bactériens. Après une introduction portant sur i) la traduction chez les procaryotes, et ii) les techniques de classification et leurs liens avec la théorie de l'information, un nouvel algorithme de partition d'un ensemble de gènes en fonction de leur usage de codons est présenté. Son application aux génomes d'*E. coli* et de *B. subtilis* permet de mettre en évidence plusieurs phénomènes. Le génome de ces organismes se décompose respectivement en 4 et 5 groupes de gènes ayant des usages de codons distincts. Les gènes du même groupe tendent à partager des fonctions similaires, et sont organisés sur le chromosome en domaines cohérents d'une longueur de 10 à 15 gènes. Cette organisation non triviale pourrait permettre une régulation de la vitesse de traduction des gènes en fonction de leur similarité avec leur environnement génétique.

Dans la seconde partie le biais de codons et le contenu en ARN de transfert de bactériophages sont analysés, comparativement à ceux de leurs hôtes. L'étude statistique montre que le contenu en ARNt des phages n'est pas aléatoire, mais biaisé en faveur d'ARNt complémentaires aux codons fréquents dans le génome du phage. Un modèle d'équation maîtresse montre que cette distribution des ARNt au sein des génomes de phages pourrait être le résultat de deux processus : l'acquisition aléatoire par le phage d'ARNt, parmi ceux de l'hôte, et la perte préférentielle des ARNt correspondants à des codons moins utilisés par le phage que par son hôte. Un tel mécanisme permettrait au phage de s'adapter en ne conservant au final que les ARNt présents en quantité insuffisante chez son hôte pendant l'infection. Finalement, on observe plus d'ARNt chez les phages lytiques que chez les tempérés, laissant supposer que les processus de traduction sont soumis à une plus forte pression de sélection chez eux.

This thesis contains some works about the codon bias and its role in bacteria and phages, particularly about regulation of translation and chromosome organization in bacteria. After an introduction describing i) translation processes in prokaryotes, and ii) bases of classification and information theories, a new clustering algorithm designed to classify a set of genes according to their codon usage is presented. Its application to the genomes of *E.coli* and *B. subtilis* puts forward multiple phenomena. Their genomes are respectively composed of 4 and 5 groups of genes sharing the same codon usage. The genes of the same group tend to have similar function, and are organized in coherent domains 10 to 15 genes long on the chromosome. This non-trivial organisation could be used to regulate the translation speed of genes depending on their similarity with their genetic context.

In the second part, the codon bias and tRNA content of phages are analyzed, relative to those of their hosts. Statistical tests show that tRNA content in phage genomes is not random, but biased towards the tRNA cognate to the frequent codons in the phage genome. A master equation model shows that this tRNA distribution could be the result of two processes : random acquisition of tRNA among those of the host, and preferential loss of tRNA cognate to codons used less in the phage genome than inside its host. Such a selection could be adaptative by allowing the phage to keep only the tRNAs insufficiently represented inside its host. Eventually, more tRNAs are observed among lytic phages than among temperate ones, which lead to the hypothesis that the selective pressure acting on translation is more important to them.