UE Evolution

TP 5 « Evolution Moléculaire »

Etude d’une famille multigénique : Insulin like protein

Dominique Mouchiroud, Marc Bailly-Bechet, Céline Brochier, Annabelle Haudry

Objectifs :

- Analyse de deux traits moléculaires : composition en base et taux d’évolution
- Manipulation des outils pour l’étude de l’usage du code génétique (calcul des fréquences en GC (G+C), calcul des fréquences en acides aminés)
- Manipulation des outils pour le calcul des taux d’évolution (modèle de Kimura, Ka, Ks)

Durant le TP, nous utiliserons le **programme Seaview**, qui est un logiciel permettant à la fois de visualiser, éditer et sauvegarder sous différents formats des alignements, de reconstruire des arbres et d’obtenir des statistiques de taux d’évolution sur différentes régions d’une séquence génomique. Nous utiliserons également **le package « seqinr** » de R permettant la manipulation et l’analyse de séquences nucléiques et protéiques. Nous verrons que la combinaison des 2 approches, de reconstruction phylogénétique et d'analyse des séquences, sont complémentaires pour appréhender la dynamique évolutive d’une famille multigénique.

Avant de débuter,

- Créer un répertoire de travail.
- Télécharger le logiciel seaview :
    - Aller sur la plateforme du *pbil*
    - *Rechercher seaview* – Télécharger la version windows
    - Récupérer le dossier *seaview.exe* dans téléchargement et le mettre dans votre répertoire
    - Cliquer sur l’icône pour décompresser le dossier.

- Pour installer le package *seqinr* dans votre répertoire
    - Lancer la version de R la plus récente
    - Installer seqinr, en acceptant les paramètres par défaut*: install.packages (« seqinr »)*
    - Charger seqinr *: library (« seqinr »)*

  Suite à l'installation, la commande *library(seqinr)* ne devrait plus donner d'erreur.

- Récupérer les fichiers de séquences sur Spiral à l'adresse « TP_Evolution_Moléculaire_DM». Vous n'analyserez qu'une partie de ces séquences, votre enseignant de TP définira au début de la séance lesquelles.

I. Analyse d'une famille multigénique

La superfamille des insulines est constituée de protéines aux activités hormonales variées. Nous étudierons 3 membres de cette superfamille : les gènes codant la protéine Insuline « INS » et les gènes codant les « Insulin-like growth factor » de type 1 (IGF1) et de type 2 (IGF2). Les gènes de la relaxine « REL », qui appartiennent à la même superfamille, permettront de raciner l'arbre de la superfamille des insulines dans les analyses phylogénétiques qui vont suivre. Chez l'homme, la taille des parties codantes et la localisation des gènes étudiés sont reportées dans le tableau ci-dessous.

Nom du gène	Longueur CDS pb	Localisation
INS	333	11p15.5
IGF1	462	12p22-23
IGF2	543	11p15.5
REL	558	9p23

- Ouvrir le fichier « famille_insuline » avec Seaview. Ce fichier contient des séquences déjà alignées. *Identifier les espèces présentes. Quelle est la nature des séquences étudiées ? Pourquoi utilise-t-on ce type de séquences ?*

Le menu « *Sites* » de Seaview permet de sélectionner semi-automatiquement certains sites particuliers de l'alignement avec la commande « *Create set* ». Dans le cas d'un alignement protéique, on peut s'intéresser à la phylogénie en n'étudiant que les sites conservés, pour éviter des biais de reconstruction phylogénétique dus aux sites très variables. La sélection *« Gblocks »* détermine automatiquement quels sont les sites conservés dans l'alignement.

- Construire l'arbre phylogénétique de la famille complète avec Seaview avec la méthode des distances sur les sites conservés. Pour cela, naviguer dans Seaview dans le menu *Tree*, sous-menu *distance*.
- *Rappeler la différence entre méthode de distance et méthode de parcimonie.*
- *Quel scénario évolutif global peut-on proposer au vu de l'arbre phylogénétique ci-dessus, pour ces gènes de la superfamille des insulines ?*
- *Sauver le fichier avec les sites sélectionnés avec la commande « save as » sous le format mase.*

II Usage du code génétique et %GC au sein de la famille multigénique

Du fait de la dégénérescence du code génétique (voir ci-dessous), de nombreux acides aminés peuvent être codés dans la séquence d'ADN par plusieurs codons dits « codons synonymes ». Le choix des codons synonymes propre à un gène ou un génome représente *l'usage du code*. Généralement, les codons synonymes ne se distinguent que par leur troisième base : la première et

surtout la deuxième base du codon sont déterminées par l'acide aminé codé. On sait que les différents codons synonymes ne sont pas toujours employés de la même manière par chaque espèce, et qu'ils peuvent être influencés à la fois par des phénomènes sélectifs (relation codon/anticodon), et par les biais mutationnels (u ≠ v) qui influencent la composition globale ou locale en bases du génome ; ces biais mutationnels sont souvent estimés à partir de la composition en bases G+C, notée %GC.

- Lancer R. Lire les séquences nucléiques du gène sur lequel vous travaillez, qui sont au format fasta, à l'aide des commandes :
- *library(seqinr)*
- *data<-read.fasta(«http://pbil.univ-lyon1.fr/members/mbailly/TP/Evolution/nomgene_nonalig.txt»)*

Dans la commande précédente, *nomgene* est bien entendu remplacé par le nom du gène sur lequel vous devez travailler, qui vous a été indiqué par votre enseignant.

The Standard Genetic Code

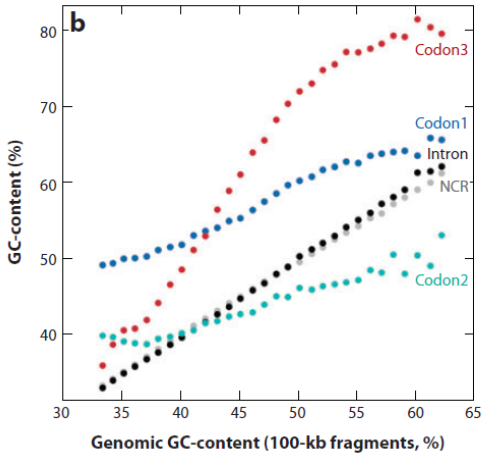
	T			C			A			G		
T	TTT	Phe	F	TCT	Ser	S	TAT	Tyr	Y	TGT	Cys	C
	TTC			TCC			TAC			TGC		
	TTA	Leu	L	TCA			TAA	STOP		TGA	STOP	
	TTG			TCG			TAG			TGG	Trp	W
C	CTT	Leu	L	CCT	Pro	P	CAT	His	H	CGT	Arg	R
	CTC			CCC			CAC			CGC		
	CTA			CCA			CAA	Gln	Q	CGA		
	CTG			CCG			CAG			CGG		
A	ATT	Ile	I	ACT	Thr	T	AAT	Asn	N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

Noter que pour les analyses de séquence qui vont être effectuées sous R, il n'est pas nécessaire que les séquences soient alignées. L'objet *data* est une liste. Pour savoir quelles séquences sont contenues dans *data*, on peut employer *names(data)* ; pour accéder aux différentes séquences que cette liste contient, on peut employer l'opérateur *$*, comme par exemple *data$Homo* pour avoir accès à la séquence humaine. Attention, R est sensible aux majuscules et aux minuscules !

Utiliser R pour étudier la composition en GC total et aux 3 différentes positions des codons avec les fonctions *GC*, *GC1*, *GC2* et *GC3*. On rappelle que pour obtenir de l'aide sur une fonction sous R, il suffit de taper ?*nomfonction.*

- *Pour **la séquence humaine**, analyser la composition en GC à chaque position des codons pour la partie codante du gène et pour les séquences non codantes (intron, 5NCR, 3NCR).*

Que remarquez-vous ? Mettez vos observations en relation avec ce qui a été dit sur le biais d'usage du code génétique. Comment interprétez-vous ces résultats en relation avec la structuration en base GC des génomes de Vertébrés (structuration en Isochores) et l'existence d'un biais mutationnel local ?Vous pouvez vous aider de la figure ci-dessous.



Distribution de la composition en GC des différents compartiments d'un gène en fonction de la composition en GC de la région (régions de 100kb classées par ordre croissant) dans le génomes humain (Duret et al 2009).

- **Mise en commun** : *les différents gènes de la superfamille se caractérisent-ils tous par le même biais, au niveau du %GC3 ? Quel constat peut-on faire en comparant le %GC de la troisième position par rapport aux deux premières positions des codons du gène ?*
- *Comment expliquez-vous la similarité de biais en GC entre l'Insuline et l'Insulin Growth Factor 2 chez l'homme ?*

- Pour les **séquences non humaines**, *analyser la composition en GC à la troisième base des codons des gènes orthologues.* Pour éviter de taper de trop nombreuses commandes, on peut utiliser *lapply()*, par exemple:
- *lapply(data,GC3)*
- *Que remarquez-vous ? Comment pouvez-vous expliquer ce patron interspécifique ?*

III Etude des taux d'évolution au sein de la superfamille des insulines

Les taux d'évolution présentent une forte variabilité : intragénique, intergénique et interspécifique. L'étude des taux d'évolution au sein de la superfamille des insulines va nous permettre d'approcher cette variabilité à ces trois niveaux d'organisation en considérant plus particulièrement le taux d'évolution dans la branche humaine. Nous prendrons le gène codant l'insuline pour illustrer ces dynamiques.

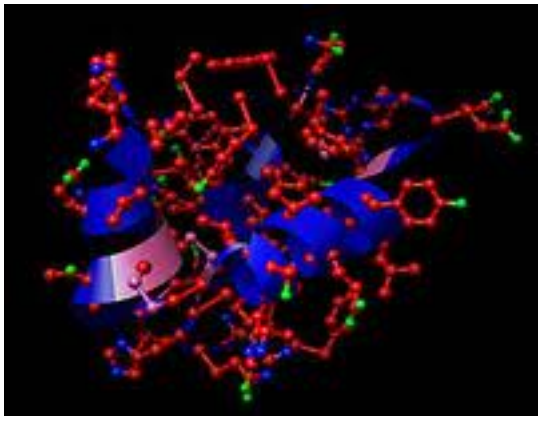
L'insuline chez l'homme présente la structuration suivante :

Peptide signal Chaine B Peptide C Chaine A

H2N----------+==========RR----------KR=======COOH

24 AA 30AA 31 AA 21 AA

Lors de la maturation de la protéine, le peptide signal et le peptide C sont éliminés. La protéine fonctionnelle est constituée des chaînes A et B reliées par deux ponts disulfures.



Une manière d'étudier les taux d'évolution consiste à établir une phylogénie du gène et à utiliser les longueurs des branches de l'arbre comme un indicateur du taux d'évolution du gène pour une lignée donnée. Ces longueurs de branches peuvent être estimées dans Seaview soit en utilisant l'échelle donnée en haut dans les arbres représentés, soit en cochant l'option *Br. Length* dans la fenêtre d'arbre.

A) Variabilité interspécifique des taux d'évolution

- Ouvrir le fichier INS_mase qui contient l'alignement du gène codant l'insuline chez l'homme avec l'insuline de plusieurs autres espèces. Faire la phylogénie globale de ce gène avec la méthode de distance *(modèle de Kimura, dit K2P).* Les arbres obtenus devront donc être racinés manuellement en employant la fonction *Re-root* de Seaview.
- *Discuter l'arbre obtenu. Quelles hypothèses pouvez-vous formuler pour expliquer les longueurs de branche élevées au sein des rongeurs ? Comment pouvez-vous tester ces hypothèses ? Faites les tests.*

Pistes : utilisation de la partie contrainte et non contrainte du gène codant, comparaison des distances Ka et Ks, calcul du rapport Ka/Ks...

B) Variabilité intragénique des taux d'évolution

- *Quels sont les attendus en termes de différence de taux d'évolution au sein du gène codant l'insuline ?*
- *En quoi la structure particulière de la protéine (figure ci-dessus) explique la variation des taux d'évolution intragénique ? Comparer les taux d'évolution Ka et Ks.*
- *A l'aide des possibilités de sélection du menu « Sites », effectuer la même analyse séparément pour les trois positions des codons dans la séquence (modèle K2P), et mettre ces résultats en relation avec ce qui a été dit pour le taux de GC aux différentes positions des codons.*

C) Variabilité des taux d'évolution au sein de la famille

Ouvrir le fichier _mase correspondant à votre gène d'étude avec Seaview.

Construire l'arbre phylogénétique correspondant. Noter la longueur de la branche humaine (modèle Ka).

- *A l'aide des possibilités de sélection du menu « Sites », effectuer la même analyse séparément pour les trois positions des codons dans la séquence (modèle K2P). Comparer les résultats obtenus entre eux. Que constate-t-on ? Ces résultats sont-ils similaires à ceux que vous avez obtenus pour l'insuline ?*

Mise en commun : *comparer les résultats obtenus sur les différents gènes. Les mettre en relation avec ce qui a été trouvé précédemment concernant le biais d'usage du code génétique.*

C) Synthèse des résultats concernant la structure et le taux d'évolution au sein de la superfamille des insulines

- L'évolution au sein d'une famille multigénique
- L'usage du code génétique et la composition en GC aux différentes positions des codons
- La variabilité des taux d'évolution entre gènes et au sein d'un gène.

D) Pour en savoir plus !

Pour compléter l'histoire de la superfamille des insulines, on recherche les séquences de la famille disponibles dans les banques de données pour des organismes phylogénétiquement plus éloignés que les Vertébrés, comme les Chordés. Voici quelques organismes clés : la lamproie (*petromyzon marinus*) à la base des Vertébrés, la myxine (*myxine glutinosa*) à la base des craniates, ou amphioxus (*Branchiostoma floridae*) à la base des Chordés.

- Aller sur http://pbil.univ-lyon1.fr, lien *WWW-QUERY*, et effectuer une recherche concernant les séquences disponibles dans Genbank pour la famille des insulines, chez ces espèces.
- *Species or taxon : mettre le nom de l'espèce*
- *Keyword :*insulin**

- *Quelles séquences obtient-on ? Comment interpréter ce résultat par rapport au scénario précédent ?*

Pour étudier l'impact de la structuration en base sur la composition en acides aminés des protéines codées,

- *Calculer la composition en acides aminés de deux protéines avec le contraste en GC3 le plus important. Quels sont vos attendus sous l'hypothèse d'un biais mutationnel ?*
- Une commande utile *: table(aaa(translate(data$nom de l'espèce)))*
- *Commenter vos résultats*.