TP Évolution Moléculaire UE Evolution L3

Étude d'une famille multigénique: Insulin-Like Protein

Hélène Badouin, Annabelle Haudry, Jean Lobry

Printemps 2021

I Introduction

I.A Objectifs

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

L'ADN des organismes conserve des traces de leur évolution. Une méthode féconde pour étudier l'histoire évolutive de ces organismes consiste donc à étudier ces traces en comparant les séquences d'acides nucléiques ou aminés entre organismes.

Cette comparaison peut s'opérer à plusieurs échelles de la « hiérarchie » moléculaire : à l'échelle de génomes entiers, on peut comparer les génomes d'individus d'espèces différentes ; à l'échelle de protéines, on peut comparer la séquence d'acides aminés (ou des nucléotides de la séquence codante) de protéines homologues entre espèces ; toujours à l'échelle de la protéine, on peut comparer la séquence de différentes sous-unités de la protéine pour en apprendre plus sur l'évolution de la protéine en question et sur les contraintes sélectives (ou non-sélectives) qui pèsent sur elles.

La comparaison des séquences passe par le choix de paramètres quantitatifs susceptibles d'être mesurés afin de formaliser ladite comparaison. **Au cours de ce TP, nous mesurerons deux paramètres quantitatifs : 1) la composition en base G+C ; 2) la vitesse d'évolution.**

Nous nous placerons à l'échelle d'une famille de protéines : **la famille des protéines homologues de l'insuline chez les animaux.** Le terme de « famille » en évolution a la même extension qu'en généalogie humaine : on regroupe en « familles » les gènes que l'on suppose descendre d'un ancêtre commun. [1]

1. Pratiquement, on suppose que deux gènes descendent d'un ancêtre commun si leurs séquences sont suffisamment similaires. La similarité de deux (ou plus) séquences peut s'estimer d'après leur alignement.

I.B Méthodes

Au cours du TP, nous utiliserons le programme Seaview, [2] qui est un logiciel permettant de visualiser, éditer et sauvegarder sous différents formats des alignements de séquences et de reconstruire des arbres phylogénétiques. Nous pourrons utiliser ses fonctionnalités pour estimer des traits moléculaires : le taux de G+C au niveau de différents types de sites d'une séquence, ou encore utiliser la reconstruction phylogénétique pour obtenir des statistiques de taux d'évolution sur différentes régions d'une séquence génomique. Nous verrons que la combinaison des deux approches (reconstruction phylogénétique et analyse de séquences) est complémentaire pour appréhender **la dynamique évolutive d'une famille multigénique**.

2. Gouy, Manolo, Stéphane Guindon, et Olivier Gascuel. « SeaView Version 4 : A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building ». Molecular Biology and Evolution 27, n° 2 (1 février 2010) : 221-24. https://doi.org/10.1093/molbev/msp259.

I.C Avant de débuter …

1. Créez un répertoire de travail. Récupérez tous les fichiers de séquence sur Claroline dans le dossier `Évolution/Resources/TP/TP Évolution Moléculaire/TP_2021/` (télécharger le dossier entièrement depuis `claroline`.)
2. Installer **Seaview**, disponible à cette adresse : `http://doua.prabi.fr/software/seaview`. Choisissez la version correspondant à votre système d'exploitation (GNU/Linux, macOS ou Windows), et suivez les instructions à l'adresse indiquée si besoin.

I.D Conventions

Dans cette fiche de TP, vous identifierez les consignes écrites en rouge et les questions auxquelles il vous faudra répondre en violet.

II Analyse d'une famille multigénique

La superfamille des insulines est constituée de protéines aux activités hormonales variées. Nous étudierons trois 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 codant 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 sur les chromosomes des gènes étudiés sont reportées dans le tableau 1 ci-dessous.

gène	longueur CDS (bp)	Localisation
INS	333	11p15.5
IGF1	462	12p22-23
IGF2	543	11p15.5
REL	558	9p23

TABLE 1: Localisation chromosomique des différents gènes dans le génome humain. (Pour comprendre la notation, voir `https://fr.wikipedia.org/wiki/Chromosome#Chromosomes_chez_les_eucaryotes`)

Commencez par ouvrir le fichier `famille_insuline` avec Seaview :

— `File/Open("famille_insuline.txt")`

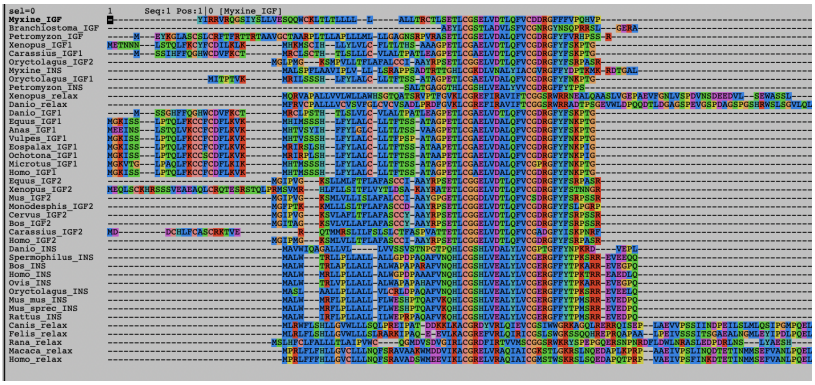

FIGURE 1: Aperçu du fichier `famille_insuline.txt` dans Seaview.

La figure 1 représente l'alignement de certaines séquences de cette famille de gènes, que vous devriez obtenir.

– *Quelles sont les espèces présentes ?*

Vous pouvez voir sur la gauche le nom (latin) des espèces étudiées. Le suffixe est un membre de `{relax,INS,IGF,IGF1,IGF2}` et correspond au type de gène étudié. Essayez de vous représenter même globalement la distribution phylogénétique de ces différentes espèces : laquelle est mammifère, laquelle est artiodactyle, etc.. [3]

3. Vous pouvez pour cela utiliser l'outil `http://lifemap.univ-lyon1.fr/explore.html`, qui permet d'explorer l'arbre du vivant de façon interactive.

– *Quelle est la nature des séquences étudiées ?*

Vous remarquerez que l'alphabet utilisé est celui des **acides aminés** ; il s'agit donc de séquences protéiques. Sont-elles alignées ? Vous remarquez la présence de gaps (tirets) dans les séquences. Or une séquence d'acide aminé ne présente pas de « gaps » biologiques, il n'y a pas de « trous » dans les protéines : un gap dans l'alignement de la séquence d'acide aminé de deux protéines est une représentation des événements biologiques qui les séparent (insertions, délétions, …). Vous avez donc affaire à un alignement de séquences protéiques. [4]

4. Pour la suite de ce TP, dès lors que vous ouvrez un nouveau fichier de séquence, posez-vous les questions suivantes : 1) ADN ou protéine ? ; 2) aligné ou non ?

– *Pourquoi utilise-t-on ce type de séquences ?*

Le recours à ce type de séquences protéiques est fréquent lorsqu'on cherche à faire de la **phylogénie dite « profonde »**, à grande échelle évolutive — c'est-à-dire lorsque l'on cherche à étudier des espèces dont la divergence est très ancienne (figure 9). Pourquoi ? Parce que les séquences protéiques évoluent « moins vite » que les séquences nucléiques, comme vous allez le voir dans la suite de ce TP.

II.A Histoire évolutive de la famille de l'insuline

Une approche couramment employée en évolution moléculaire pour déterminer l'histoire évolutive d'une famille de gène consiste à comparer la reconstruction phylogénétique de la généalogie de cette famille à l'histoire évolutive des espèces — autrement dit à comparer un arbre de gène à un arbre de spéciation. [5]

5. L'approche inverse est également très instructive : on peut vouloir reconstruire l'histoire du processus de spéciation en aggrégeant les histoires évolutives probables du plus grand nombre possible de familles de gènes.

On parle de « reconstruction phylogénétique» car il s'agit de reconstruire l'histoire évolutive probable d'une famille de gène à partir des traces qu'elle conserve de cette évolution dans les séquences. (Vous verrez plus loin que la reconstruction en question peut être totalement erronée, en l'absence d'informations supplémentaires.)

Toutes les positions (colonnes) de l'alignement ne sont pas également propices à cette reconstruction : certaines positions sont trop divergentes pour qu'il soit possible d'en retirer une information sur l'histoire évolutive des séquences. La première étape d'une reconstruction phylogénétique consiste donc à déterminer les positions, ou « sites », de l'alignement qui doivent être considérés comme informatifs. 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. Seaview permet de faire cela de façon automatisée (ou non), via le menu `Sites`.

Cliquez sur `Sites/Create set` (figure 2) : vous obtenez la fenêtre en figure 3. Sélectionnez `Gblocks, ok` et laissez les paramètres par défaut pour la

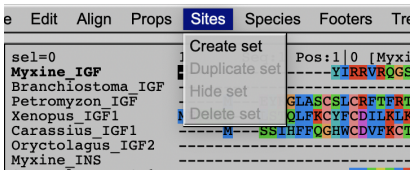

FIGURE 2: Création d'une sélection de sites

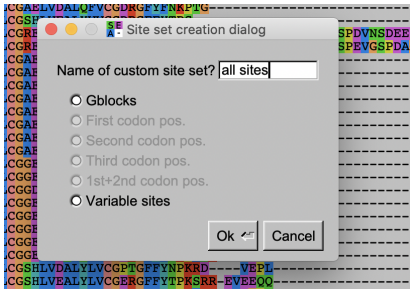

FIGURE 3: site set creation dialog

fenêtre suivante (figure 4). (`Gblocks` est un algorithme de sélection automatique des positions les plus conservées d'un alignement. [6])

Vous verrez apparaître une nouvelle ligne de couleur blanche en bas de l'alignement (figure 5); elle correspond aux sites sélectionnés par l'algorithme de `Gblocks`. Vous pouvez avoir l'intuition de la manière dont fonctionne cet algorithme en observant le type de positions sélectionnées par lui : ce sont des sites contigus et conservés, c'est-à-dire partagés par l'ensemble des séquences sélectionnées. Notez que deux ensembles de sites ont été sélectionnés; nous verrons plus bas à quoi ils correspondent.

6. Castresana, J. « Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis ». Molecular Biology and Evolution 17, n° 4 (1 avril 2000) : 540-52. https://doi.org/10.1093/oxfordjournals.molbev.a026334.

II.B reconstruction de la phylogénie de la famille de l'insuline

On peut désormais construire l'arbre phylogénétiques de la famille complète.

Cliquez sur `Trees/Distance methods` (figure 6). Choisissez `BioNJ` (NJ pour « neighbor joining »), et sélectionnez `Kimura` comme matrice de distance (figure 7). Vous obtenez l'arbre en figure 8.

II.C scénario évolutif global et estimation des dates de duplication

– *Quel scénario évolutif global peut-on proposer au vu de l'arbre phylogénétique obtenu et des dates de divergence (arbre en figure 9) pour ces gènes de la superfamille des insulines ?*

Essayez d'extraire de l'arbre l'information qu'il porte sur l'histoire évolutive de cette famille de gène : Repérez les paralogies s'il y en a, [7] c'est-à-dire les différents événements de duplication depuis l'ancêtre commun aux relaxines et insuline-like protein; repérez aussi les groupes d'orthologues,

7. https://fr.wikipedia.org/wiki/Paralogie

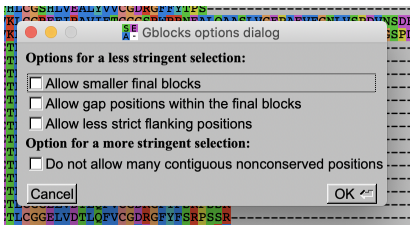

FIGURE 4: Paramètres de `Gblocks`

FIGURE 5: Sites sélectionnés par `Gblocks`.

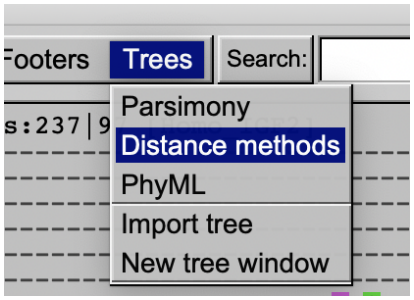

FIGURE 6: Menu `Trees` et ses différentes méthodes de constructions

c'est-à-dire les protéines homologues dont la divergence résulte du processus de spéciation.[8]

8. https://fr.wikipedia.org/wiki/Orthologie

Nous allons prendre l'exemple de la duplication la plus évidente, celle de l'IGF en IGF1 et IGF2. Vous constatez sur l'arbre que vous avez obtenu (figure 8) que Branchiostoma, Myxine et Petromyzon ne possèdent qu'un orthologue de l'IGF ; toutes les espèces de Vertébrés en possèdent 2 (IGF1 et IGF2 : e.g. Xenopus est présent deux fois dans le sous-arbre des IGFs). C'est donc que la duplication de l'ancêtre commun des IGF s'est produite au cours de la divergence qui a conduit aux Vertébrés, et est donc ultérieure à 530 Ma (figure 9).

– *Quels sont les différents événements évolutifs qui permettent de retracer l'histoire de la famille des insulines ?*

III Usage du code génétique et GC% au sein de la famille de l'insuline

Du fait de la dégénérescence du code génétique (figure 10), 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 à la même fréquence par

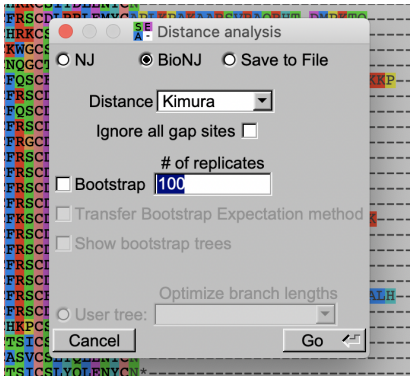

FIGURE 7: Choix de la matrice de distance

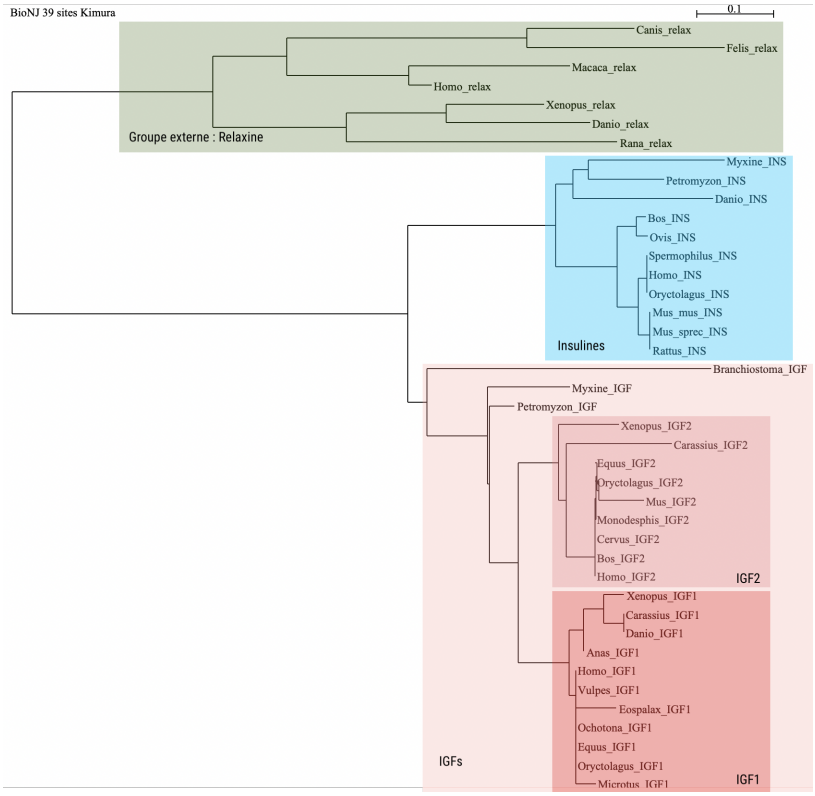

FIGURE 8: Arbre phylogénétique de la famille de l'insuline, coloré par groupes d'orthologues.

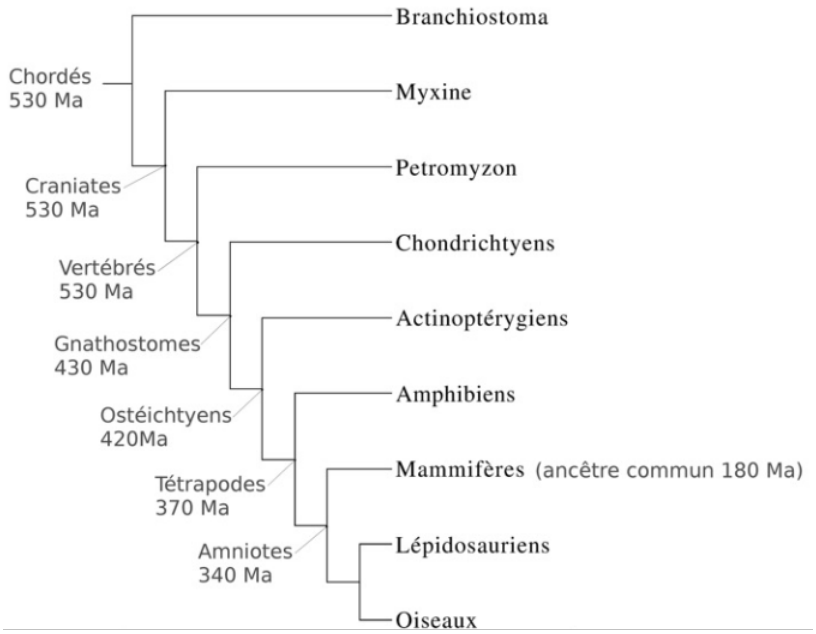

FIGURE 9: Dates approximatives de divergence entre clades

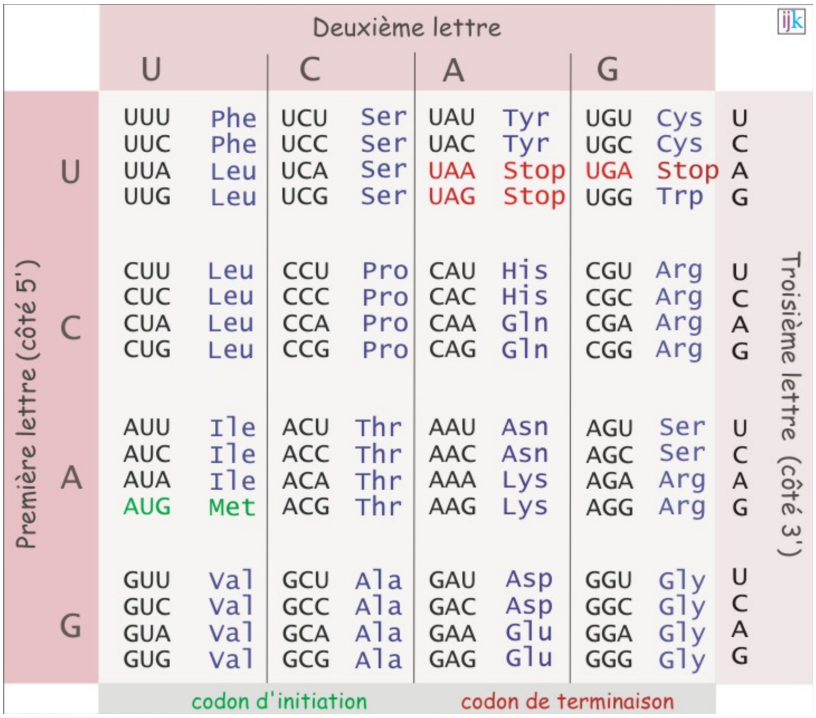

FIGURE 10: Code génétique

chaque espèce (biais d'usage du code), et que cette fréquence peut être influencée à la fois par des phénomènes sélectifs (relation codon/anticodon), et par des mécanismes neutres (biais mutationnels (u = v) et conversion génique biaisée (BGC)). [9] Ces mécanismes neutres influencent globalement ou localement la composition en base du génome ; ils sont souvent estimés à partir de la composition en bases G+C, notée %GC.

Nous allons utiliser Seaview pour mesurer la variabilité des %GC. Cette variabilité peut être quantifiée au sein d'une espèce, en comparant les %GC de différents gènes d'une même famille au sein d'une espèce donnée (le %GC des paralogues). Ou bien elle peut l'être entre espèces, en comparant les %GC de différents orthologues d'une même famille entre espèces.

On parle des %GC puisque la composition en base peut être mesurée à l'échelle d'un gène entier, ou seulement à des positions spécifiques : par exemple les premières, deuxièmes ou troisièmes positions des codons, ou encore séquences introniques vs. séquences codantes (exoniques).

9. Voir cours de D. Mouchiroud et Duret, Laurent, et Nicolas Galtier. « Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes ». Annual Review of Genomics and Human Genetics 10, n° 1 (septembre 2009) : 285-311. https://doi.org/10.1146/annurev-genom-082908-150001.

III.A Variabilité intra-spécifique des taux de G et C

Nous allons commencer par étudier la composition en base à différentes positions de la séquence du **gène de l'insuline** chez l'homme.

Ouvrez le fichier `INS_nonalig` dans Seaview (figure 11). [10] Sélectionnez la séquence humaine en cliquant simplement dessus (figure 12). Dès lors qu'une ou plusieurs séquences sont sélectionnées (noircie), Seaview ne considère plus qu'elles (« aucune sélection » équivaut « à toutes les séquences sont considérées »). On peut donc quantifier la composition en base à l'échelle de ce gène chez l'homme uniquement. Pour cela cliquez sur `Props/Statistics` (figure 13).

Vous obtenez (normalement) le résultat suivant :

10. À quoi avez-vous affaire ? ADN ou protéine ? Aligné ou non ?

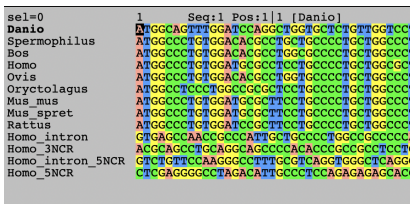

FIGURE 11: Séquences nucléiques de l'insuline

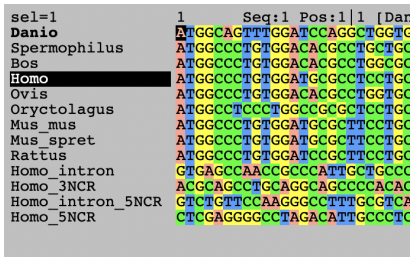

FIGURE 12: Sélection de la séquence humaine

```
BASE COMPOSITION :
     All sites   :   17.1% A    31.8% C    32.7% G    18.3% T
```

Le taux de GC est donc de $31,8 + 32,7 = 64,5\%$.

Répétez l'opération en sélectionnant successivement :

— le premier intron,

— le second intron (dans la région 5' transcrite mais non codante, 5NCR)

— la région 5'NCR

— la région 3'NCR

Nous allons maintenant calculez le taux de G et C aux positions 1, 2 et 3 des codons. Il s'agit de sélectionner des sites spécifiques de la séquence ; ce que Seaview permet de faire via le menu … `Sites` (figure 14).

Une fois les positions 1 des codons sélectionnés, calculez leurs statistiques résumés via `Props/Statistics`. Vous devez obtenir le résultat suivant :

```
BASE COMPOSITION :
     All sites   :   12.6% A    35.1% C    34.2% G    18.0% T
```

Répétez l'opération pour les positions 2 et 3 des codons, de façon à compléter la colonne `INSULINE` du tableau 2.

– *Quel constat peut-on faire en comparant le %GC de la troisième position par rapport aux deux premières positions du gène de l'insuline ?*

Complétez maintenant les autres colonnes du tableau 2. Pour cela répétez les différentes analyses de cette partie pour le gène IGF2, IGF1 et REL en

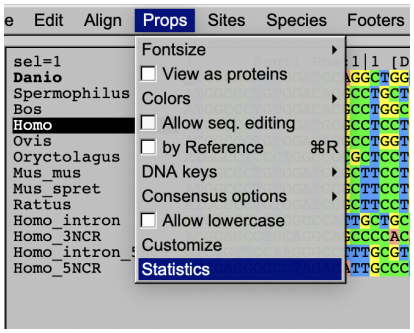

FIGURE 13: Calcul des statistiques résumés des séquences sélectionnées.

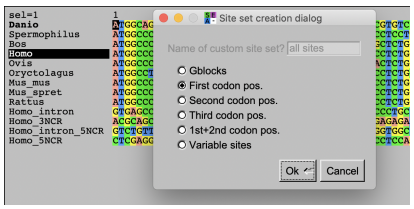

FIGURE 14: Sélection des premières positions des codons

Sites	INSULINE	IGF2	IGF1	RELAXINE
CDS total				
GC1				
GC2				
GC3				
GC Intron				
1				
2				
GC NCR				
5'				
3'				

TABLE 2: Statistiques résumées de %GC pour l'insuline.

ouvrant le fichier `X_nonalig`, où X correspond au nom du gène, toujours en se focalisant sur la séquence humaine.

- *Comment interprétez-vous les relations entre le %GC du codant et du non codant (voir figure 15) ?*
- *Pouvez- vous mettre ceci en relation avec la structuration en base GC des génomes de Vertébrés (structuration en Isochores) et l'existence d'un biais mutationnel local ?* [11]
- *Les différents gènes de la superfamille se caractérisent-ils tous par le même biais d'usage du code génétique ?*
- *Comment expliquez-vous la similarité de biais en GC entre l'Insuline et l'Insulin Growth Factor 2 chez l'homme ?*

11. Indice : confrontez avec le tableau 1 plus haut.

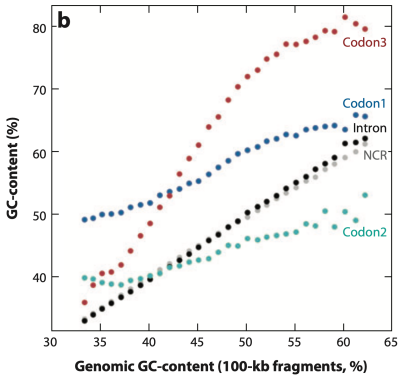

FIGURE 15: Distribution de la composition en GC des différents compartiments d'un gène en fonction de la composition en GC de la région (régions de 100kb classées par ordre croissant) dans le génome humain (Duret et Galtier, 2009).

III.B Variabilité inter-spécifique des taux de G et C

Pour chaque orthologue (REL, INS, IGF1 et IGF2), calculez le taux de G et C aux troisièmes positions des codons uniquement pour chaque espèce de façon à compléter le tableau 3 ci-dessous. Pour cela utilisez le fichier `X_mase`, où `X` correspond au nom du gène. (Pensez à bien sélectionner l'espèce et à

	INS	IGF2	IGF1	REL
Poisson				
Amphibien				
Souris				
Bos/Equus				
homme				

TABLE 3: Comparaison du %GC3 entre espèces

utiliser le menu `Sites` pour sélectionner la position 3 des codons.) Si une espèce n'est pas présente dans l'alignement, choisissez une espèce voisine.

- *Que remarquez-vous ?*
- *Comment pouvez-vous expliquer ce patron interspécifique (voir figure 16) ?*

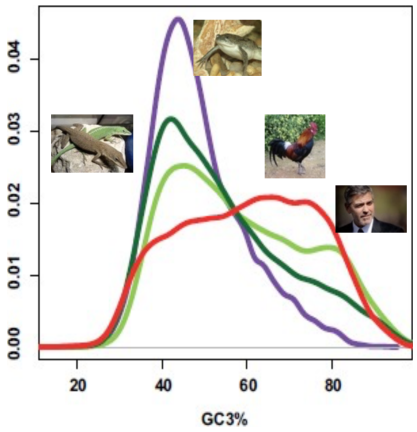

FIGURE 16: Distribution de la fréquence en GC3 de l'ensemble des gènes codant chez 4 espèces (homme rouge, Poulet vert clair, Xenope mauve, Lézard vert foncé) Figuet, Emeric, Marion Ballenghien, Jonathan Romiguier, et Nicolas Galtier. « Biased Gene Conversion and GC-Content Evolution in the Coding Sequences of Reptiles and Vertebrates ». Genome Biology and Evolution 7, n° 1 (1 janvier 2015) : 240-50. https://doi.org/10.1093/gbe/evu277.

III.C Conclusion

L'objectif de cette partie était de vous faire comprendre que l'évolution des séquences dépend largement de processus indifférents à la fonction exercée par les protéines dans l'organisme. Il convient donc de les prendre en compte dans les inférences que l'on tire de l'évolution de ces séquences.

La sélection naturelle n'opère vraiment que sur les positions du génome ou des gènes qui contribuent à la fitness de l'organisme. C'est ce que nous allons toucher du doigt dans la partie suivante.

IV Étude de taux d'évolution au sein de la superfamille des insulines

IV.A Introduction

Les taux d'évolution [12] 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 (voir la fiche UniProt de l'Insuline) :

12. Voir https://fr.wikipedia.org/wiki/vitessedévolution.

```
    Pep.      Chaine                 Chaine
    signal      B        Pep. C        A
H2N--------+-------RR----------KR--------COOH
     24AA     30AA        31AA        21AA
```

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. [13] Nous allons voir dans quelle mesure les contraintes exercées par la sélection naturelle sur la fonction de ces différentes chaînes affectent leur vitesse d'évolution.

- *Pensez-vous que la sélection naturelle a plutôt tendance à ralentir ou accélérer l'évolution d'une région fonctionnelle ?*

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. Encore une fois, nous allons utiliser Seaview pour comparer les vitesses d'évolution :

1. à l'échelle inter-spécifique : entre les différents orthologues de l'insuline.
2. à l'échelle intra-génique : entre les différentes chaînes des orthologues de l'insuline.
3. à l'échelle inter-génique : entre les différents paralogues de l'insuline chez l'homme

13. Confrontez avec l'alignement de la famille de l'insuline réalisé au début du TP (figure 1). Qu'en concluez-vous ?

IV.B variabilité inter-spécifique des taux d'évolution

1. Ouvrez le fichier `INS_mase` qui contient le gène codant l'insuline chez l'homme et plusieurs autres espèces.
2. Traduisez les séquences nucléiques en protéine via l'onglet `Props/View as proteins` (figure 17).
3. Effectuez l'alignement des séquences protéiques (figure 18). [14]
4. Revenez aux séquences nucléiques en décochant `Props/View as proteins` (figure 19).
5. Effectuez la reconstruction phylogénétique avec la méthode des distances (Ka) (figure 20). [15]

14. Avez-vous une idée de la raison pour laquelle il est souhaitable de réaliser l'alignement des séquences protéiques plutôt que directement nucléiques ?

15. La matrice de distance Ka repose sur l'estimation de la distance génétique entre deux séquences en ne se concentrant que sur les substitutions non-synonymes. À votre avis, pourquoi utilisons-nous cette matrice de distance ici ? N'hésitez-pas à consulter cet article fondamental de l'évolution moléculaire : McDonald, J. H., et M. Kreitman. « Adaptive Protein Evolution at the Adh Locus in Drosophila ». Nature 351, n° 6328 (20 juin 1991) : 652-54. https://doi.org/10.1038/351652a0.

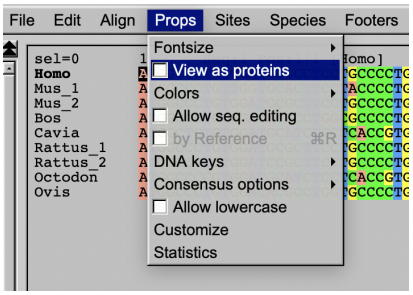

FIGURE 17: Traduction *in silico* des séquences codantes

À ce stade, vous devriez obtenir l'arbre en figure 21. Comparez-le avec l'arbre de spéciation (figure 22). Vous faites face à un problème récurrent de renconstruction phylogénétique appelé « **problème d'attraction des longues branches** » (LBA). [16] Spécifiquement, les longues branches correspondant à Cavia et Octodon (respectivement le Cochon d'Inde et le Dègue

16. https://fr.wikipedia.org/wiki/Attraction_des_longues_branches

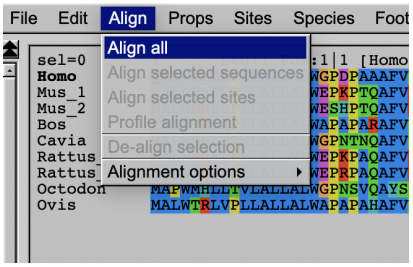

FIGURE 18: Alignement des séquences protéiques

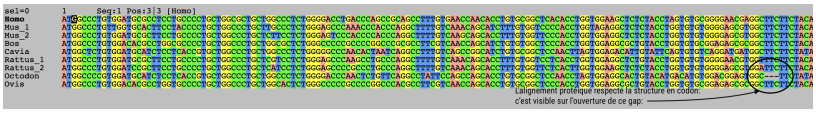

FIGURE 19: Retour aux séquences nucléotidiques alignées

du Chili, deux rongeurs d'Amérique du Sud) entraînent leur regroupement, et au clustering de l'homme avec les artiodactyles (Bos et Ovis).

Connaissant l'arbre de spéciation par ailleurs, il nous faut re-raciner l'arbre en choisissant comme groupe externe l'ancêtre commun des deux artiodactyles présents. Utilisez pour cela la fonction `Re-root` de la fenêtre d'arbre dans Seaview (figure 23).

Vous devriez obtenir l'arbre en figure 24 ; comparez-le avec l'arbre de spéciation (figure 22) pour vérifier que le racinement soit correct : normalement, les artiodactyles sont en groupe externe, et les rongeurs sont regroupés dans un même clade.

1. Accélération de l'évolution de l'insuline chez les rongeurs du « Nouveau Monde »
   - *Comment pouvez-vous expliquer les longueurs de branches élevées chez Cavia et Octodon ?*

   (Les 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.)

   - *Comment pouvons-nous tester ces hypothèses ?*

   Si c'est la sélection naturelle qui est responsable de l'accélération de l'évolution de l'insuline chez Cavia et Octodon, son action devrait être la plus forte aux positions fonctionnelles. Il nous faut donc comparer la vitesse d'évolution des deux chaînes A et B avec les deux chaînes non-

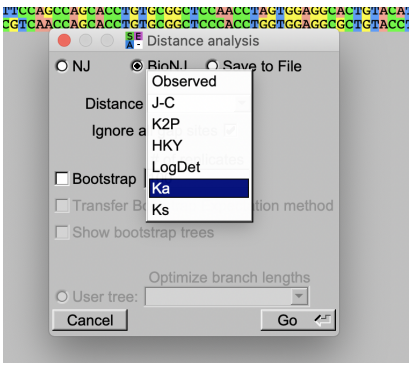

FIGURE 20: Choix de la matrice de distance Ka

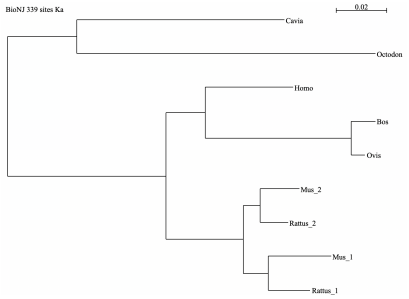

FIGURE 21: Arbre obtenu d'après les positions non-synonymes (matrice de distance Ka) pour l'insuline.

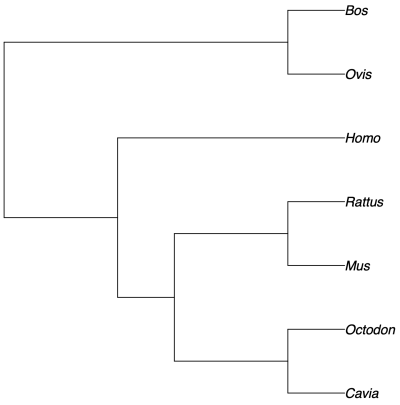

FIGURE 22: Arbre de spéciation entre les différentes espèces étudiées

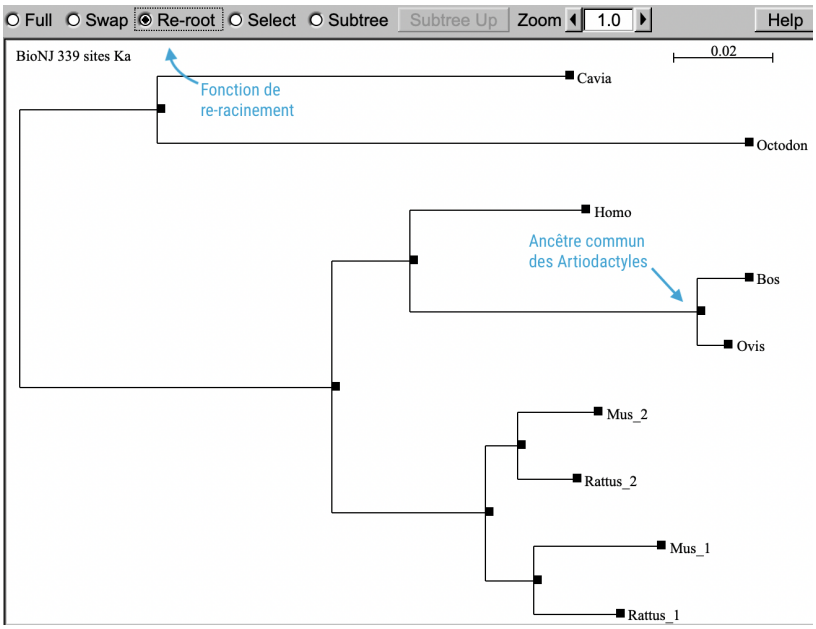

FIGURE 23: Re-racinement de l'arbre par les artiodactyles

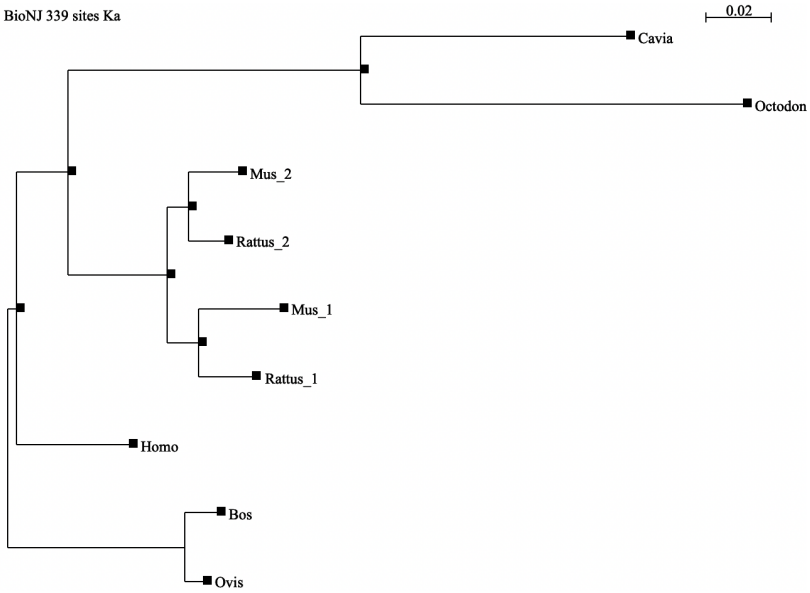

FIGURE 24: Arbre de la famille de l'insuline correctement raciné.

fonctionnelles du peptide C et signal qui sont excisées au cours de la maturation de la protéine.

Pour cela deux fichiers sont à votre disposition : `INS_AB_mase` contient une sélection pré-enregistrée des sites correspondant aux acides aminés des chaînes A et B ; `INS_var_mase` contient une sélection des chaînes variables (peptide signal et C).

Ouvrez le fichier `INS_AB_mase` et sélectionnez les sites des chaînes A et B via le menu `Sites` (figure 25).

FIGURE 25: Sélection des chaînes A et B

Re-générer l'arbre en utilisant la méthode des distances (Ka), re-racinez l'arbre par l'ancêtre commun des artiodactyles et affichez les longueurs de branches (cochez `Br. length`).

Conservez cette fenêtre, et répétez la procédure avec le fichier `INS_var_mase`. Placez les deux fenêtres côte à côte pour comparez les longueurs de branches (figure 26).

– *Est-ce que cela confirme votre hypothèse ?*

IV.C Variabilité intragénique des taux d'évolution

La partie précédente sur l'accélération de la vitesse d'évolution chez Cavia et Octodon a dû vous faire prendre conscience du fait que toutes les chaînes d'une protéine ne sont pas également contraintes par l'action de la sélection naturelle. Cette partie montrera que ces différences de contraintes

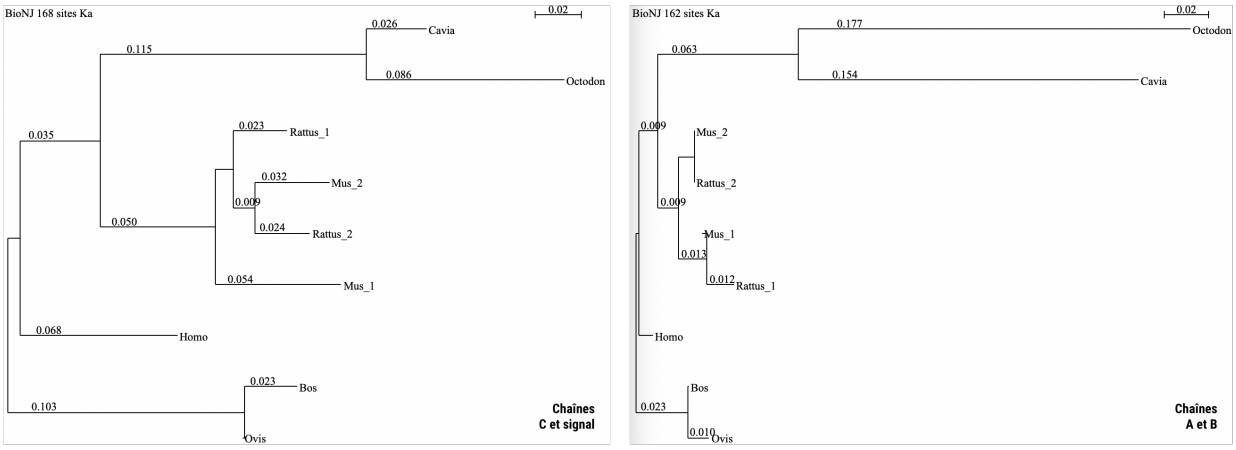

FIGURE 26: Comparaison de la vitesse d'évolution des régions fonctionnelles et non-fonctionnelles de l'insuline

se retrouvent également aux différentes positions des codons. Nous allons mesurer les vitesses d'évolution aux positions 1, 2 et 3 des codons des quatre protéines de la superfamille {REL, INS, IGF1 et IGF2}.

Commençons par l'insuline :

1. Ouvrez le fichier `INS_mase` ;
2. sélectionnez toutes les séquences sauf celles de Cavia et Octodon ;
3. traduisez les séquences nucléiques en protéines ;
4. alignez les séquences protéiques ;
5. revenez aux séquences nucléiques en décochant `Props/View as proteins` ;
6. créez via le menu `Sites` une sélection des positions 1 des codons ;
7. reconstruisez la phylogénie de l'insuline en utilisant la matrice de distance K2P (matrice de Kimura à 2 paramètres) ;
8. Racinez l'arbre par l'ancêtre commun des artiodactyles si besoin ;
9. Notez la longueur de la branche séparant la Souris de l'homme.
10. faites de même pour les positions 2 et 3.

   – *Comparez les taux d'évolution entre les trois positions du codon (figure 27).*
   – *Que constatez-vous ?*

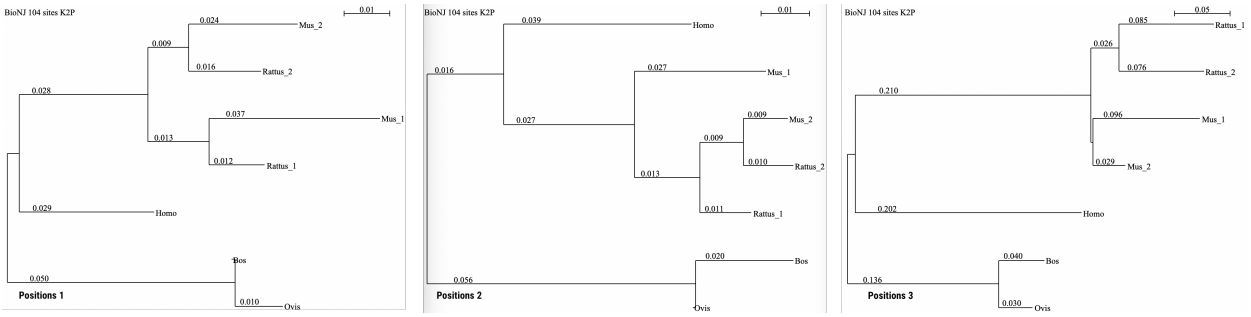

FIGURE 27: Comparaison de la vitesse d'évolution aux différentes positions des codons de la séquenced de l'insuline.

	1ère position	2ème position	3ème position
INS			
IGF2			
IGF1			
REL			

TABLE 4: Comparaison des taux d'évolution aux différentes positions des codons entre les différents gènes.

IV.D Variabilité des taux d'évolution au sein de la famille

Répétez les 10 étapes de la partie précédente pour les trois autres gènes, de façon à compléter le tableau ci-dessous (tableau 4) :

- *Comparez les résultats obtenus sur les différents gènes.*
- *Mettez en relation avec ce qui a été observé précédemment concernant le biais d'usage du code génétique.*

V Conclusion

Ce TP avait pour but de vous faire appréhender la manière dont on peut tenter d'inférer l'histoire évolutive d'une famille de gène à partir des traces que leur évolution laisse dans les séquences d'acides nucléiques, et par voie de conséquence dans les séquences d'acides aminés.

Ce TP nous a permis d'aborder des notions vraiment essentielles de l'évolution moléculaire :

1. les séquences fonctionnelles et les séquences non-fonctionnelles ne sont pas contraintes de la même façon par la sélection naturelle. C'est la raison pour laquelle on a souvent recours aux séquences protéiques des gènes pour faire de la phylogénie à grande échelle évolutive. (À des échelles évolutives plus courtes, par exemple celles en jeu dans le monitoring de populations sauvages en écologie, on peut utiliser des marqueurs évoluant plus vite — donc moins contraints, qui sont donc plus informatifs de l'histoire évolutive récente des populations, par exemple les microsatellites).
2. Au sein des séquences fonctionnelles, on peut distinguer différents niveaux de contraintes : Les chaînes peptidiques qui ne sont pas présentes dans la protéine maturée sont moins contraintes que les chaînes qui le sont (A et B chez l'insuline). De plus, les positions des codons 1 et 2 sont plus contraintes que les troisièmes positions, du fait de la redondance du code génétique.
3. Ces contraintes peuvent se mesurer de différentes façons. Nous avons abordé dans ce TP deux traits moléculaires : la composition en base et la vitesse d'évolution.
4. La vitesse d'évolution peut se mesurer sur un arbre phylogénétique obtenu en tentant de reconstruire l'histoire évolutive d'une famille de gènes. Cette reconstruction peut se faire via différentes méthodes (nous avons essentiellement abordé les méthodes de distance dans ce TP : Kimura, Ka, K2P, etc.).
5. D'autres forces que des contraintes sélectives peuvent influencer l'évolution des séquences. Nous avons abordé ce point à travers l'étude de la structuration en isochore de la composition en base : rappelez-vous

que les deux gènes Insuline et IGF2 ont des compositions en base et un usage du code similaires ; ce que nous avons attribué à leur proximité chromosomique.