

FAMILLES MULTIGÉNIQUES

L'étude de l'origine du polymorphisme des gènes, en particulier celle des allèles des gènes qui codent pour l'alpha-antitrypsine ou la G6PD, a révélé l'importance des mutations ponctuelles dans la diversification des génomes au cours de l'évolution. Mais ce mécanisme est loin d'être le seul. Dans le génome de tous les organismes existent des gènes dont les similitudes sont telles qu'elles impliquent une parenté entre eux : on dit que ce sont des gènes homologues paralogues. Ils proviennent tous d'un gène ancestral par duplications successives suivies d'une divergence plus ou moins importante des copies. L'ensemble des gènes ainsi apparentés forme une famille multigénique dont le modèle classique est celui de la famille des gènes qui codent pour les chaînes alpha, bêta, gamma et delta de l'hémoglobine humaine. D'autres exemples sont apportés qui illustrent le fait que la duplication génique fait partie du processus général d'expansion du génome qui va de la répétition d'un triplet au sein d'un gène (cas du gène qui code pour la huntingtine) jusqu'à la polyploïdie.



La famille des globines

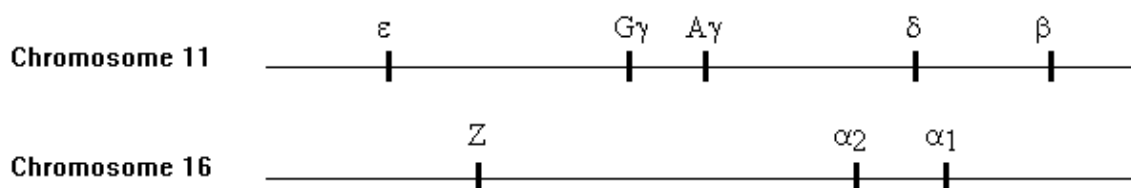
Informations scientifiques

L'élève doit être sensibilisé à l'idée que l'analyse en cours des génomes des espèces actuelles peut fournir des données informatives sur les processus qui, dans le passé, ont contribué à la genèse de nouveaux gènes. Il en est ainsi en ce qui concerne les séquences des gènes qui, dans l'espèce humaine, codent pour les chaînes de l'hémoglobine.

Pour aborder cette étude avec les globines, il est nécessaire d'apporter des informations sur la synthèse des différentes chaînes au cours de la vie de l'individu :

| Hémoglobine | Type de chaînes | Moment de synthèse |
|-----------------|---------------------|------------------------|
| Hb embryonnaire | $Z_2 \epsilon_2$ | avant 9 semaines |
| Hb foétale | $\alpha_2 \gamma_2$ | 9 semaines---naissance |
| Hb A | $\alpha_2 \beta_2$ | après naissance |
| Hb D | $\alpha_2 \delta_2$ | après naissance |

Après la naissance il y a synthèse de 97% Hb A, 2% Hb D et 1% Hb foétale. L'intérêt du changement de chaîne après la naissance pourra être évoqué. Il importe que l'élève saisisse que ces différentes chaînes sont codées par des gènes différents et non par les allèles d'un même gène, ce qui peut être facilité par la localisation chromosomique de ces gènes.



Ce schéma révèle l'existence de deux gènes gamma ($G\gamma$ et $A\gamma$) et de deux gènes alpha (α_1 et α_2). Les deux gènes gamma ne diffèrent, dans leur région codante, qu'au niveau du codon 136 (GGA = Glycine pour $G\gamma$, et GCA = Alanine pour $A\gamma$). Dans la banque, on a retenu un seul gène $G\gamma$. De même, l'identité

des séquences codantes des deux gènes alpha fait qu'on a retenu un seul des deux. De ce fait, le schéma ci-dessus est à simplifier pour les élèves.

Une grande ressemblance dans la structure des chaînes alpha, bêta, gamma et delta a été mise en évidence : elles possèdent :

- 3 exons et 2 introns dans tous les cas ;
- des longueurs de séquences codantes identiques ou très voisines.

Structures des gènes codant pour les chaînes alpha, bêta, gamma et delta de l'hémoglobine humaine

| CHAÎNE | EXON 1 | intron 1 | EXON 2 | intron 2 | EXON 3 |
|--------|------------------|-------------------|-------------------|--------------------|---------------------|
| bêta | 1--140 (140)* | 141--270 (130) | 271--494 (224) | 495--1344 (850) | 1345--1606 (262) |
| delta | 1--144 (144) | 145--272 (128) | 273--493 (221) | 494--1382 (889) | 1383--1641 (259) |
| gamma | 1--144 (145) | 145--267 (123) | 268--488 (223) | 489--1376 (887) | 1377--1593 (217) |
| alpha | 1--128 (128) | 129--245 (117) | 246--450 (205) | 451--590 (140) | 591--831 (241) |

* Longueur de la séquence

Utilisation pédagogique

Comparaison des chaînes protéiques

Toutes ces chaînes d'hémoglobine participent au transport du dioxygène, et sont synthétisées dans les mêmes cellules, les érythroblastes. On peut se demander si, à la similitude de fonction de ces chaînes, correspondent des ressemblances dans leur structure et leurs séquences d'acides aminés. Pour obtenir des informations, les élèves doivent noter le nombre d'acides aminés de chaque chaîne, comparer les séquences deux à deux, relever le nombre d'acides aminés identiques ou le pourcentage d'identité (attention le logiciel donne le nombre de différences). Ces comparaisons donnent les résultats suivants concernant le nombre d'acides aminés identiques entre les séquences :

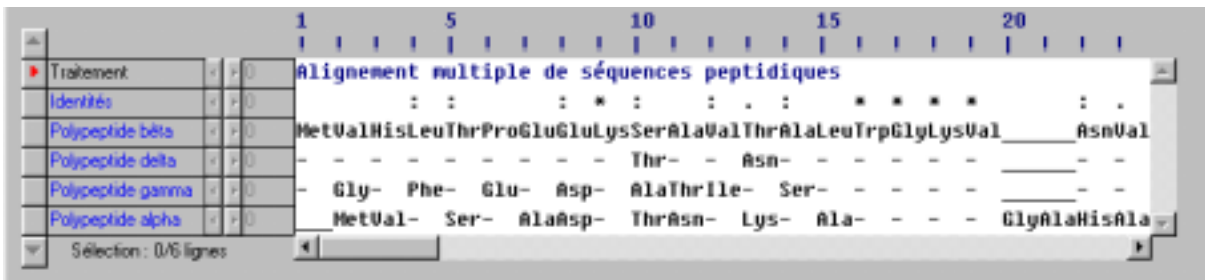
| | | | | |
|------------------------|-------|------------|------------|-----------|
| | delta | | | |
| | bêta | 137 | | |
| | gamma | 106 | 108 | |
| | alpha | 65 | 65 | 60 |
| | | delta | bêta | gamma |
| Longueur des séquences | | 147 | 147 | 147 |
| | | | | alpha |
| | | | | 142 |

On constate que β et δ sont très semblables, que le nombre d'éléments identiques est pratiquement le même entre β et γ et entre δ et γ . La chaîne α montre le plus de différences par rapport aux trois autres.

On peut aussi leur demander de rechercher les sites conservés dans les quatre chaînes : pour cela, ils réalisent une comparaison simultanée (alignement avec discontinuités) des quatre séquences d'hémoglobine. Il est recommandé d'utiliser comme séquence de référence une des séquences les plus longues (bêta, gamma ou delta longues de 444 bases chacune).

Le décompte du nombre de positions identiques dans les quatre chaînes ainsi que les identités entre chacune des séquences et la référence est fourni par l'information (cliquer sur la fenêtre de résultat de l'alignement pour l'activer ; cliquer sur l'icône d'information).

Les résultats de cet alignement multiple peuvent être très légèrement variables suivant la séquence choisie comme référence. Dans l'exemple ci-dessous montrant les 24 premiers acides aminés alignés, c'est la séquence de la globine bêta qui est la référence ; les positions identiques dans les quatre séquences sont indiquées par des étoiles (les points indiquent les positions comportant des acides aminés similaires, acides ou basiques).



L'alignement comprend 149 acides aminés

-> 51 acides aminés identiques (représentés par le signe *)
soit 34,2 % d'identité, et 68,5 % de ressemblance

Bêta protéique

longueur : 147 acides aminés (sans compter les discontinuités)
> référence pour la comparaison

Delta protéique

longueur : 147 acides aminés (sans compter les discontinuités)
-> 137 a.a. identiques à la séquence de référence Bêta protéique,
soit 93,2 % d'identité

Gamma protéique

longueur : 147 acides aminés (sans compter les discontinuités)
-> 108 a.a. identiques à la séquence de référence Bêta protéique,
soit 73,5 % d'identité

Alpha protéique

longueur : 142 acides aminés (sans compter les discontinuités)
-> 65 a.a. identiques à la séquence de référence Bêta protéique,
soit 45,8 % d'identité

le signe - représente les identités

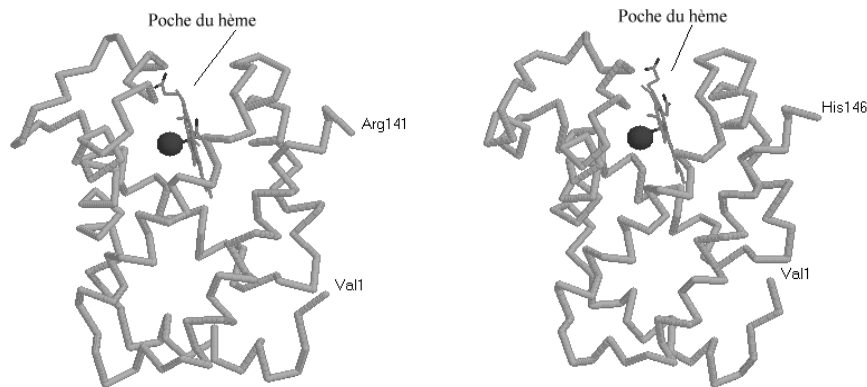
le signe _ représente les discontinuités

le signe . représente la proximité chimique d'acides aminés

La réflexion sur la signification des similitudes constatées conduit à rechercher si elles peuvent être dues au hasard ou si elles témoignent d'une « parenté » entre les chaînes des globines. Les élèves peuvent approcher cette question à partir :

- de **considérations théoriques** : si les 20 acides aminés étaient également utilisés dans les séquences protéiques, la probabilité d'avoir, entre deux séquences non apparentées (constituées au hasard), le même acide aminé à un site déterminé, serait de 5 %. Comme certains acides aminés sont plus fréquemment utilisés que d'autres, on considère que les ressemblances n'indiquent une parenté qu'au delà de 20 %. Or, les ressemblances entre les chaînes de l'hémoglobine sont nettement supérieures à ces valeurs ;
- de la **comparaison de deux polypeptides ayant des fonctions différentes**, par exemple, une chaîne de globine et l'enzyme qui intervient dans la synthèse de l'antigène A du système du groupe sanguin ABO. La comparaison de la séquence de la chaîne de globine avec des segments de même longueur dans la séquence de l'enzyme, montre que le pourcentage des acides aminés identiques ne dépasse pas 20 %.

Alors que les globines bêta, delta et gamma sont assez proches, la globine alpha apparaît comme la plus différente des autres. Cependant, malgré un grand nombre d'acides aminés différents, les globines alpha ont une structure tridimensionnelle semblable. En particulier, la poche accueillant le hème est conservée.



Les globines alpha (à gauche) et bêta (à droite). Visualisation en squelette carboné ; le hème est en bâtonnets et l'oxygène en sphère. Visualition obtenue avec le logiciel RASMOL (Roger Sayle, Glaxo Wellcome Research & Development, Stevenage, Hertfordshire, UK). Coordonnées spatiales provenant de Protein Data bank.

Il importe maintenant de voir si la « parenté » mise en évidence entre les chaînes des globines peut être confirmée au niveau des gènes qui les codent.

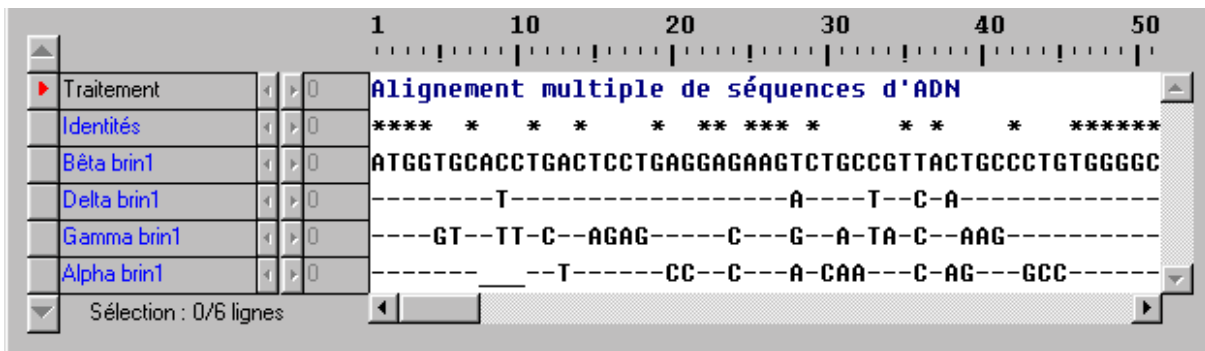
Comparaison des chaînes nucléiques

Le tableau qui résulte de la comparaison des séquences deux par deux est le suivant :

| | | | | | |
|------------------------|-------|------------|------------|------------|-------|
| | delta | | | | |
| | bêta | 411 | | | |
| | gamma | 340 | 339 | | |
| | alpha | 258 | 267 | 238 | |
| | delta | | bêta | gamma | alpha |
| Longueur des séquences | 444 | 444 | 444 | 444 | 429 |

L'alignement multiple avec discontinuité de ces quatre séquences est réalisé en prenant comme référence la séquence bêta.

ATGC Comparer les séquences



Alignement avec discontinuité montrant les 51 premières bases des séquences des globines humaines alpha, bêta, gamma et delta. Le signe * représente les identités, le signe _ les discontinuités.

i

Alignement multiple de séquences d'ADN :

-> 213 bases identiques (représentées par le signe *)
soit 47,3 % d'identité

Bêta
longueur : 444 bases (sans compter les discontinuités)
-> référence pour la comparaison

Delta
longueur : 444 bases (sans compter les discontinuités)
-> 411 bases identiques à la séquence de référence Bêta,
soit 92,6 % d'identité

Gamma
longueur : 444 bases (sans compter les discontinuités)
-> 339 bases identiques à la séquence de référence Bêta,
soit 76,4 % d'identité

Alpha
longueur : 429 bases (sans compter les discontinuités)
-> 264 bases identiques à la séquence de référence Bêta,
soit 61,5 % d'identité

Comme attendu, les similitudes constatées à propos des protéines se retrouvent au niveau des parties codantes des gènes. Notons toutefois que de nombreuses différences au niveau des gènes ne se traduisent pas par des différences entre les polypeptides, à cause du caractère dégénéré du code génétique.

Le bilan de tout ce travail conduit à l'idée qu'il existe dans le génome humain des gènes apparentés car présentant des ressemblances qui ne peuvent s'expliquer par des processus aléatoires. Il reste à rechercher, ce qui ne peut être fait avec le logiciel :

- une explication à cette parenté (notion de duplication des gènes) ;
- une explication aux différences constatées entre ces gènes et à leur degré plus ou moins important (mutations des gènes après leur duplication) ;

- une explication à l'importance des différences constatées au niveau des introns (absence de pression sélective).

Les considérations sur l'importance des différences entre les gènes ou les polypeptides pourra permettre d'établir une généalogie des diverses duplications.

La famille des gènes HLA de classe I

Informations scientifiques

Ce système de gènes permet de cibler deux aspects de l'évolution des génomes :

- la duplication génique révélée par la similitude des séquences aux trois locus ;
- le polymorphisme révélé par le grand nombre d'allèles présents dans les populations avec une fréquence élevée à chaque locus.

Ainsi, les trois gènes HLAA, HLAB et HLAC ont la même organisation d'ensemble : 8 exons et 7 introns, comme résumé dans la figure ci-après.

| | HLAA0201 | HLAB2705 | HLACW301 |
|-----------|-----------------|-----------------|-----------------|
| Promoteur | 0-27 | 0-27 | 0-27 |
| Exon 1 | 28-125 | 28-124 | 28-124 |
| Intron 1 | 126-264 | 125-253 | 125-255 |
| Exon 2 | 265-524 | 254-623 | 256-525 |
| Intron 2 | 525-765 | 524-764 | 526-767 |
| Exon 3 | 766-1041 | 765-1040 | 768-1043 |
| Intron 3 | 1042-1640 | 1041-1615 | 1044-1631 |
| Exon 4 | 1641-1926 | 1616-1891 | 1632-1906 |
| Intron 4 | 1927-2015 | 1892-1983 | 1907-2027 |
| Exon 5 | 2016-2130 | 1984-2103 | 2028-2147 |
| Intron 5 | 2131-2568 | 2104-2554 | 2148-2588 |
| Exon 6 | 2569-2604 | 2555-2576 | 2589-2622 |
| Intron 6 | 2605-2746 | 2577-2682 | 2623-2776 |
| Exon 7 | 2747-2794 | 2683-3334 | 2777-2824 |
| Intron 7 | 2795-2963 | | 2825-2988 |
| Exon 8 | 2964-3528 | | 2989-3408 |

L'alignement multiple des trois allèles HLAA0201, HLAB2705 et HLACW301 montre l'importance des sites identiques dans les séquences nucléiques (brin non transcrit). La similitude moyenne entre les gènes HLAA et HLAB est de 82,5%, entre HLAA et HLAC de 83,5%, entre HLAB et HLAC de 87%. La différence moyenne entre les trois gènes est environ trois fois plus importante qu'entre allèles de même locus.

On s'est interrogé sur la signification évolutive de cette famille multigénique. Les trois gènes s'expriment dans la quasi totalité des cellules nucléées de l'organisme en codant pour des polypeptides dont la fonction est de présenter les peptides antigéniques aux lymphocytes T. Par suite du polymorphisme, la plupart des individus sont hétérozygotes pour chacun de ces gènes. Comme il y a codominance, cela fait que chaque personne possède six types d'antigènes HLA de classe I ce qui augmente la capacité à présenter des peptides et sans doute celle de réagir à une multiplicité d'agents pathogènes.

Utilisations pédagogiques

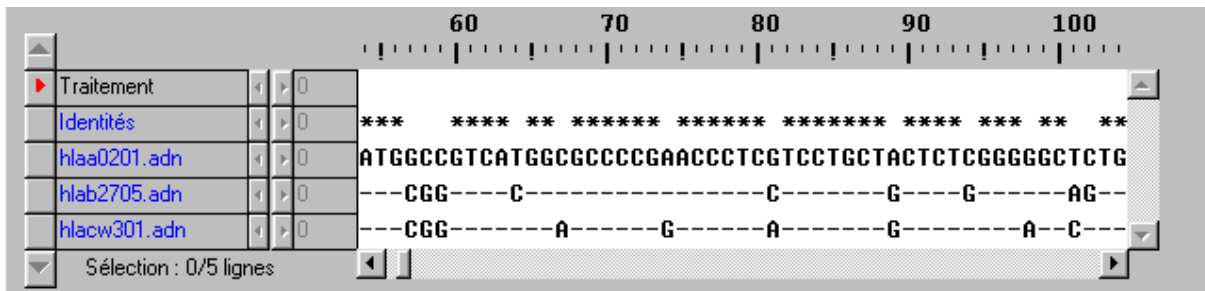
Recherche des similitudes entre les séquences nucléiques des gènes HLAA, HLAB et HLAC


L'étude du polymorphisme a familiarisé les élèves avec les gènes du système HLA. La recherche d'une parenté éventuelle entre ces gènes peut être posée à partir de l'identité des fonctions qu'exercent les polypeptides qu'ils codent (notion envisagée lors de l'étude des mécanismes immunitaires). Le seul moyen de tester cette hypothèse est de comparer les trois séquences codantes des allèles de ces gènes (HLAA0201, HLAB2705, HLACW301). Puisqu'elles n'ont pas la même longueur, on peut utiliser la modalité " **Alignement avec discontinuités** ". La comparaison de ces séquences deux à deux permet de dresser le tableau suivant :

| | | | |
|------------------------|------------|------------|----------|
| HLAA0201 | | | |
| HLAB2705 | | 976 | |
| HLACW301 | 952 | 993 | |
| | HLAA0201 | HLAB2705 | HLACW301 |
| Longueur des séquences | 1098 | 1089 | 1101 |

Plus long, l'alignement simultané des trois séquences donne les résultats suivants :

Comparer les séquences



 L'alignement comprend 1101 bases
 -> 919 bases identiques (représentées par le signe *)
 soit 83,5 % d'identité

HLAA0201 codant
 longueur : 1098 bases (sans compter les discontinuités)
 -> référence pour la comparaison

HLAB2705 codant
 longueur : 1089 bases (sans compter les discontinuités)
 -> 974 bases identiques à la séquence de référence HLAA0201 codant,
 soit 89,4 % d'identité

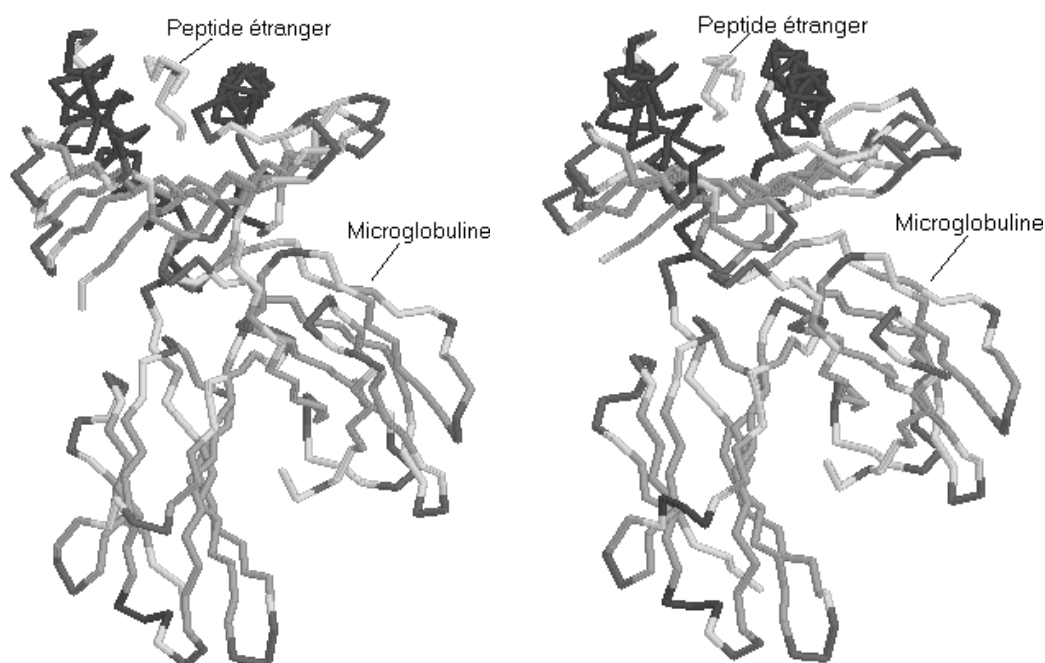
HLACW301 codant
 longueur : 1101 bases (sans compter les discontinuités)
 -> 952 bases identiques à la séquence de référence HLAA0201 codant,
 soit 86,5 % d'identité

Recherche des similitudes entre les séquences protéiques des gènes HLA A, HLAB et HLAC

Un travail similaire peut être réalisé avec les séquences des polypeptides codés par ces allèles, ce qui donne les résultats suivants :

| | | | |
|------------------------|------------|------------|----------|
| HLAA0201 | | | |
| HLAB2705 | | 304 | |
| HLACW301 | 291 | 305 | |
| | HLAA0201 | HLAB2705 | HLACW301 |
| Longueur des séquences | 365 | 362 | 366 |

Malgré les différences dans leurs séquences, HLA A02 (à gauche) et HLAB27 (à droite) ont des structures spatiales similaires, avec une corbeille de présentation de peptides antigéniques.



Les deux molécules, flanquées sur leur droite de la microglobuline, sont montrées en squelette carboné avec le logiciel RASMOL (Roger Sayle, Glaxo Wellcome Research & Development, Stevenage, Hertfordshire, UK). Coordonnées tridimensionnelles provenant de Protein Data Bank.

L'existence d'une famille multigénique étant corroborée, il reste à faire la synthèse sur l'évolution de cette partie du génome : duplication de gènes et nombreuses mutations à chaque locus.

La famille multigénique des hormones GH, HPRL et HLP

L'hormone de croissance (GH), la prolactine (HPRL) et l'hormone lactogène placentaire (HLP) ont des actions différentes mais apparentées. Ainsi, l'hormone lactogène placentaire stimule la synthèse protéique et a aussi une action sur la glande mammaire nettement inférieure toutefois à celle de la prolactine. Ces hormones sont codées par des gènes situés sur des chromosomes différents (chromosome 6 pour le gène de la prolactine, et chromosome 17 pour ceux de l'hormone de croissance et de l'hormone lactogène placentaire).

La comparaison des séquences codantes de GH et HLP révèle une similitude de l'ordre de 92%. La similitude de leurs séquences d'acides aminés est de 85%. Le gène HLP existe chez tous les Mammifères placentaires mais non chez d'autres Vertébrés. Une duplication du gène de l'hormone de croissance a donc probablement eu lieu il y a plus de 85 à 100 millions d'années, lorsque de nombreuses lignées de Mammifères se sont établies. Ces gènes paralogues sont restés associés sur le même chromosome.

La similitude de la séquence codante du gène qui code pour la prolactine avec celles des gènes GH et HLP est beaucoup moins parlante puisqu'elle ne dépasse pas 35%. Néanmoins, la structure globale identique des gènes, la longueur de même ordre des polypeptides, la présence d'acides aminés communs à certaines positions font que les spécialistes admettent que le gène HPRL fait partie de la même famille multigénique que GH et HLP.

Tous les Vertébrés possèdent les gènes GH et HPRL (mais la prolactine joue des rôles différents...) ; il est donc possible qu'un gène ancestral unique ait été dupliqué avant que ne divergent les Amphibiens et les Poissons il y a environ 400 millions d'années. Ensuite, les deux gènes ont évolué indépendamment dans chaque espèce. Leur localisation chromosomique différente, de même que celle des gènes de la famille FSH, TSH, LH, HCG, révèle la souplesse du génome au cours de l'évolution.

Utilisations pédagogiques

Toutes les actions possibles reposent sur la comparaison de séquences permettant la recherche de similitudes, témoignages d'une parenté. Il est possible de partir d'une réflexion sur les diverses hormones hypophysaires et placentaires de la banque, d'**imaginer des parentés éventuelles à partir de leurs fonctions** et de **les tester à l'aide du logiciel**.

Dans un premier temps, on peut se limiter aux hormones LH, HCG, GH, HLP et HPRL. Les élèves doivent arriver aisément à conclure à l'existence de deux familles multigéniques (LH et HCG d'une part, GH et HLP d'autre part).

The screenshot shows a software window titled "Alignement multiple de séquences peptidiques". On the left, there is a list of sequences: "Lh.pro" and "hcgb.pro". The main area displays the alignment of these two sequences. The LH beta sequence is: MetGluMetLeuGlnGlyLeuLeuLeuLeuLeuLeuSerMetGlyGlyAlaTrpAlaSerArgGluPro. The HCG beta sequence is: Phe-Thr-Lys. Asterisks are placed under the first 15 residues of the LH beta sequence, indicating a high degree of identity with the HCG beta sequence. The alignment is shown with vertical lines above the sequences, and a selection bar at the bottom indicates "Sélection : 0/4 lignes".

Début de la comparaison des séquences de LH bêta et HCG

Les informations obtenues sont :



-> 394 bases identiques (représentées par le signe *)
soit 79,0 % d'identité

LH β codant

longueur : 426 bases (sans compter les discontinuités)

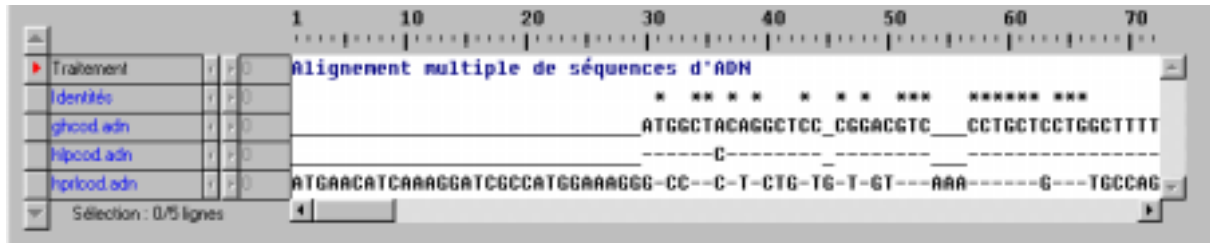
-> référence pour la comparaison

HCG β codant

longueur : 498 bases (sans compter les discontinuités)

-> 394 bases identiques à la séquence de référence LH β codant,
soit 79,1 % d'identité

En revanche, il est difficile de conclure pour le gène de la prolactine. On peut toutefois constater que ce gène présente un peu plus de similitude avec GH et HLP qu'avec les hormones de l'autre famille.



L'alignement comprend 695 bases

-> 291 bases identiques (représentées par le signe *)
soit 41,9 % d'identité

GH codant

longueur : 654 bases (sans compter les discontinuités)

-> référence pour la comparaison

HLP codant

longueur : 654 bases (sans compter les discontinuités)

-> 706 bases identiques à la séquence de référence GH codant,
soit 92,7 % d'identité

HPRL codant

longueur : 684 bases (sans compter les discontinuités)

-> 306 bases identiques à la séquence de référence GH codant,
soit 44,7 % d'identité

le signe - représente les identités

le signe _ représente les discontinuités

Il est possible de réaliser l'alignement multiple des chaînes bêta de FSH, LH, TSH et HCG et repérer l'existence de sites identiques remarquables pour les quatre chaînes (en particulier les cystéines).

| | | 95 | 100 | 105 | 110 | 115 |
|------------|--|-----|-----|-----|-----|-----|
| Traitement | | | | | | |
| Identités | | * | - | - | ; | ; |
| hcg.pro | | Cys | Pro | Arg | Gly | Val |
| lh.pro | | - | - | Asp | - | - |
| lsh.pro | | Ala | His | His | Ala | Asp |
| tsh.pro | | - | Leu | His | - | - |

Alignement multiple des séquences peptidiques des hormones HCG, LH, FSH et TSH. Quelques-unes des cystéines alignées sont visibles en positions 94, 110, 112 et 115.

Cela conduit à l'idée que la recherche de similitudes repose sur des analyses plus fines que sur le simple dénombrement de sites identiques. Ce sont des analyses de ce type qui amènent les spécialistes à l'idée que le gène HPRL fait partie de la même famille multigénique que GH et HLP. Il reste à envisager la signification des différences de similitude entre gènes faisant partie d'une même famille multigénique pour amorcer une histoire des duplications géniques successives.

En conclusion, cette activité doit permettre une sensibilisation à l'importance évolutive des duplications géniques et à l'idée qu'elles peuvent aboutir à des gènes codant pour des polypeptides aux fonctions différentes.