UE Bioinfo - Articles à analyser

① PSI-BLAST : Une Modification du programme BLAST pour construire une recherche par « profil »

Contexte : Le programme BLAST recherche dans les bases de données les séquences présentant des homologies locales avec une séquence « requête » (*Query,* en anglais).

En général, BLAST identifie toute une série de résultats positifs, ensemble de séquences apparentées, constituant une famille. Au sein de ces séquences apparentées, les zones d'homologie ne sont pas distribuées de manière uniforme et aléatoire, mais sont au contraire concentrées sur des régions fonctionnelles importantes.

L'idée est d'apporter une modification à BLAST pour pouvoir tenir compte de ces effets de localisation privilégiée des conservations de séquence dans les alignement.

Dans les scores d'alignement, l'idée est d'accorder plus de « points » aux positions conservées de manière systématique qu'aux positions peu conservée. Ceci implique de s'affranchir des matrice de score.

PSI-BLAST est la variante de BLAST qui permet de réaliser cette opération, en construisant automatiquement des matrices de « profils » de scores.

La seule partie de l'article à analyser est celle qui concerne PSI-BLAST (on ignorera la première partie sur GAPPED-BLAST), à partir de la page 3394 (numérotation des pages de l'article.

Questions à traiter :

1) Présenter de manière synthétique le principe général de la méthode (il n'est pas demandé de traiter en détail le paragraphe sur le calcul des fréquences/scores intitulé *target frequency estimation*). On expliquera en particulier les différences de principe avec le fonctionnement de BLAST présentées dans le cours.
2) Pourquoi le programme nécessite-t-il plusieurs cycles d'itération ?
   Qu'est qui change à chaque cycle ?
3) Discuter les avantages et inconvénients éventuels par rapport à BLAST classique.
4) Quelle information peut-on tirer du profil de score construit par PSI-BLAST et quelles utilisations peut-on en faire ?
5) Tester le programme PSI-BLAST sur le site de l'EBI et comparer les résultats obtenus avec BLAST classique lors de la première itération.
   On utilisera le lien suivant :
   http://www.ebi.ac.uk/Tools/sss/psiblast/
   La séquence à tester est celle de la myoglobine humaine (protéine fixant l'oxygène dans nos muscles et leur donnant leur couleur rouge). La voici en format FASTA, vous pouvez la copier et la coller dans la fenêtre correspondante du site de l'EBI :

```
>sp|P02144|MYG_HUMAN Myoglobin OS=Homo sapiens GN=MB PE=1 SV=2
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE
DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH
PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
```

La base de donnée à chercher est celle des séquences d'origine humaine : **UniProtKB Human** (la base de donnée par défaut vous ne restreignez pas aux seules séquences humaines, le calcul sera beaucoup

plus long et la sortie plus difficile à analyser). Après le premier cycle, si vous faites défiler la page, un bouton apparaît avec marqué « *run next iteration* ». Cliquez pour lancer les itérations successives.

Le résultat de BLAST simple apparait à la première itération. Les résultats PSI-BLAST dans les itérations suivantes.

Il sera demandé de présenter les résultats obtenus et de les commenter. Vous êtes invités à regarder les différents affichages possibles et en particulier l'onglet « *Download PSSM file* » qui affiche la matrice de profil de score.

② Analyse de l'Evolution du génome des vertébrés à partir de l'analyse du génome du Tétrodon (poisson-globe ou *pufferfish*)

Contexte : Cet article décrit le séquençage du génome du Tétrodon, une espèce de poisson-globe (des poissons qui se gonflent d'eau lorsqu'ils se sentent menacés par des prédateurs). Les auteurs ont obtenu une séquence qui est « ancrée » sur les chromosomes de cet animal, c'est à dire qu'ils ont établi la localisation de chaque élément séquencé et assemblé (*contig* et *scaffold*) sur chaque chromosome. Ceci permet une analyse de l'organisation génomique globale et une comparaison avec d'autres génomes de vertébrés, notamment ceux d'un poisson apparenté, le fugu (*Takifugu*), mais aussi ceux de mammifères, l'homme et la souris. Ils utilisent ensuite ces comparaisons pour tenter de reconstruire une « histoire globale » de l'évolution du génome des vertébrés, de puis la divergence poissons / tétrapodes (oiseaux, reptiles, mammifères, batraciens), il y a 450 millions d'années. Il met en particulier en évidence une duplication génomique globale (*Whole genome duplication* ou WGD en abrégé), intervenue dans la lignée des poissons osseux (poissons téléostéens, *teleost* en anglais).

La partie de l'article à traiter concerne principalement celle sur la duplication génomique (partie intitulée *Genome Duplication*), même si une compréhension superficielle du processus de séquençage et d'assemblage du génome sera demander.

Complément d'information :

Les auteurs utilisent une donnée appelée « taux de mutations silencieuses » et noté K_s. Une mutation silencieuse est une mutation dans l'ADN qui ne change pas la protéine codée par cette zone.

Par exemple, l'acide aminé « lysine » est codé par les codons AAA et AAG qui sont « synonymes ». Une mutation qui changerait AAA en AAG (changement sur la troisième position) est dite silencieuse, parce qu'elle ne modifie pas la protéine. Il n'y a pas donc de pression de sélection évolutive sur ces positions, du fait de cet absence de conséquence. Entre deux gènes dupliqués, les mutations silencieuses s'accumulent donc au cours du temps et leur nombre (ou taux) est une mesure de la durée écoulée depuis cette duplication.

Questions à traiter :

1) Résumer très succinctement le processus d'assemblage et d'annotation du génome. Vous pouvez choisir les points intéressants que vous voulez souligner et/ou que vous avez le mieux compris.
2) Expliquez comment ils montrent l'existence d'une duplication génomique globale dans la lignée des poissons osseux.
3) Quel est le rôle de l'analyse de la synténie dans cette analyse.
4) Comment reconstituent-ils un scénario de l'évolution du génome des vertébrés.
5) Utilisez le serveur http://cinteny.cchmc.org/ pour générer une ou des cartes de synténie entre génomes de vertébrés, à l'échelle génomique globale ou en zoomant à l'échelle de certains chromosomes (par exemple, le chromosome 7 du poulet et le 9 du poisson-zèbre *zebrafish*).

③ GENSCAN, l'outil informatique de prédiction des gènes dans le génome humain.

Contexte : Cet article décrit la conception de GENSCAN, le principal outil de prédiction de gène utilisé pour l'analyse du génome humain à partir de 1999-2000. Cet outil basé sur une représentation des séquences composant le génome sous forme d'un modèle de Markov caché (HMM) à 27 états (voir la figure 3 de l'article). Il distingue en particulier notamment différents types de régions ayant des propriétés de composition en nucléotides différentes:

Les régions intergéniques (entre les gènes)

Les gènes, qui se décomposent en :

1) Un promoteur P (séquence permettant la transcription des ARN messagers)
2) Une région 5' amont (5' UTR : début de l'ARN messager avant le premier codon)
3) Une alternance d'exons (codants) et d'introns (non-codants, éliminés par épissage dans l'ARN messager final)
4) Une région 3' aval (3'-UTR : fin de l'ARN messager après le codon-stop)
5) Le signal de terminaison de la transcription de l'ARN messager (poly-A signal)

Pour les introns et les exons, le modèle tient compte de la question périodicité d'ordre 3 du code génétique (codons de 3 nucléotides) et introduit donc le concept de « phase » (il y a trois phases pour lire l'ADN, en fonction du nucléotide sur lequel on commence un codon).

Le modèles traite de manière spécifique et fine les jonctions exon-intron également appelées *donor splice site*. Ces jonctions se caractérisent par un motif conservé (voir Figure 2) que le programme essaie d'identifier au moyen de règles adaptées, basées sur une classification de sous-catégories de ce motif conservé.

Enfin, il utilise une modélisation particulière de la longueur de certains de ces éléments (introns, exons initiaux, internes et terminaux) pour être plus réaliste.

Questions à traiter :

L'analyse devra se centrer principalement sur la partie « *Methods* » qui commence à la page numérotée 85. Cet article est complexe et je vous demande de vous focaliser sur ce que vous en avez compris.

1) Décrire de manière synthétique le principe du modèle et du programme, en essayant d'expliquer les choix des auteurs, lorsque ça vous parait compréhensible
2) Quelles sont les particularités de GENSCAN par rapport aux autres programmes antérieurs cités dans l'article ?
3) Comment les auteurs ont paramétré leur modèle (probabilités de transition…) ?
4) A votre avis, pourquoi modéliser les longueurs des segments ?
5) Quel est l'intérêt de la modélisation fine des jonctions exon-intron (*donor splice site*) et plus généralement des signaux (*Signal models*) ?

6) A votre avis, quel est l'intérêt d'utiliser un modèle de Markov d'ordre 5 pour les exons codants (plutôt que 4, par exemple ?)

④ Analyse bioinformatique de l'évolution moléculaire d'une famille de protéines.

Contexte : la Dihydro-uridine (DHU ou D) est une base nucléique modifiée, présente dans certains ARN (comme les ARN de transfert), dérivée de l'uridine (U). La conversion du U en D dans ces ARN est réalisée par des enzymes (protéines) spécifiques, appelées **dihydro-uridine synthases** ou DUS.

Ces enzymes qui catalysent toute la même réaction chimique de base forment une grande famille qui varie en fonction des espèces, de la nature de l'ARN dans lequel est introduite la modification et de la position précise de la conversion U->D dans cet ARN.

L'article présente une analyse bioinformatique exhaustive de toute cette famille, afin de comprendre son organisation et son évolution.

La catégorisation initiale des séquences repose sur un programme d'analyse de graphe, appelé CLANS qui effectue une projection géométrique d'un graphe sur l'espace euclidien (représentation 2D), en essayant d'adapter la longueur des arêtes du graphe au poids de cette arête (ici, la eValue de BLAST), voir figure 2 de l'article.

Questions à traiter :

1) Décrivez de manière synthétique la stratégie utilisée par les auteurs
2) Quels types d'analyses sont effectuées par les auteurs, avec quels outils informatiques ?
3) Quel est à votre avis l'intérêt de la catégorisation initiale des séquences de DUS
4) Quelles correlations observent-ils avec la phylogénie des espèces vivantes ?
5) Expliquez le scénario hypothétique qu'ils proposent pour l'évolution de cette famille d'enzyme

⑤ La base de donnée des COG : les clusters de gènes orthologues

Contexte : Les COG, *cluster of orthologous groups of genes*, sont des ensembles de gènes provenant chacun d'une espèce différente, et tous orthologues entre eux. On rappelle que deux gènes sont orthologues entre eux si ils présentent des homologies de séquence et codent des protéines remplissant **la même fonction** dans les deux espèces dont ils sont issus.

Ces clusters sont construits à partir de comparaisons de séquences entre génomes complets et permettent une classification par groupes.

L'un des apports de l'article à analyser est d'une part l'augmentation de la taille de la base de données des génomes de bactéries et d'autre part son extension aux génomes des organismes supérieurs eucaryotes (les COG eucaryotes sont appelés KOG)

Questions à traiter :

1) Comment est construite la banque de COG ? Avec quels critères sont définis les COG et les orthologues ?
2) Quels conclusions évolutives tirent les auteurs de l'analyse ?
3) Quelles différences qualitatives observent-il entre COG et KOG ?
4) Qu'est ce que l'analyse de pattern phylétique mentionnée par les auteurs ? Quelles en sont les utilisations possibles ?