Compte rendu TP1 GMIN206 - Banques de données biologiques – Anaïs Barray

I. Les banques de données

Q1. PDB collecte et permet de visualiser les formes en 3D des structures macromoléculaires (protéines, acides nucléiques et assemblages complexes).

Q2. Il s'agit de la molécule Cyanomet hemoglobin, composées de l'Hemoglobine alpha chain et Hemoglobine beta chain.

Q3. Des exemples de mots clés des fichiers PDB sont : HEADER, TITLE , COMPND , SOURCE , KEYWDS , EXPDTA , AUTHOR , REVDAT , REMARK , DBREF , SEQRES, SOURCE, HET…

Q4. La base de données Pfam est une collection de familles protéiques. La version actuelle est la 27.0, elle contient actuellement 14831 familles.
Lorsque l'on recherche *hemoglobin*, Pfam affiche 42 résultats uniques.

En utilisant SRS, nous cherchons tout d'abord dans PDB les entrées portant le nom 1o1i, puis nous recoupons cette information avec la banque PFAM contenant hemoglobine.
Nous obtenons deux résultats :
SWISSPFAM:HBA_HUMAN
SWISSPFAM:HBB_HUMAN
La séquence 1o1i fait donc bien partie de la famille des hémoglobines.
Le numéro d'accession est PF00042.12

II. Interrogation de banques de données via GQuerry

1. Recherche simple

Q5. La base de données Nucléotide est une collection de séquences provenant de plusieurs sources, dont GenBank, RefSeq, EMBL, DDBJ, INSDC, TPA et PDB.

Q6. Il y a 10593141 entrées pour les séquences nucléotidiques de l'organisme *homo sapiens*.

Q7. La première page contient les 20 premières séquences nucléiques provenant de l'humain (ARN, ADN etc). Chaque entrée est décrite brièvement par son nom, une description, son numéro d'accession, et des liens vers GenBank...

Q8. La description GenBank des entrées est détaillée.
En écrivant *homo spaiens* on obtient 27 séquences nucléotidiques, Lorsque l'on précise la recherche avec [orgn] nous n'obtenons aucun résultat. Cela est logique car homo spaiens n'est pas recencé comme étant un organisme dans la base de données, les résultats affichés précédemment étant dus à des erreurs de frappes lors de l'entrée de la séquence par les auteurs.

2. Utilisation d'opérateurs booléens

Q9.
*2935 résultats pour trinucleotide repeat ; Il y a forcément écrit trinucleotide repeat l'un à la suite de l'autre dans le nom de l'entrée.

*6087 pour trinucleotide AND repeat ; On observe repeat et trinucleotide dans l'entrée, mais pas forcément l'un à la suite de l'autre.
*5681176 pour trinucleotide OR repeat ; On observe soit repeat, soit trinucleotide, soit les deux dans l'entrée.
*2378 pour trinucleotide NOT repeat. On n'observe pas le mot repeat dans la page de l'entrée.

Q10.
La différence entre les deux premières requêtes est le fait que l'on recherche les termes de façon dissociée ou non (c'est à dire soit un couple de mots forme un critère de recherche à lui seul, soit les mots séparés forment la recherche).

3. Combinaison de plusieurs requêtes

Q11.
human : 19216717 séquences nucléotidiques
Search details : "Homo sapiens"[Organism] OR human[All Fields]

homo sapiens : 15159327 séquences nucléotidiques

Search details : "Homo sapiens"[Organism] OR homo sapiens[All Fields]

Dans tous les cas, la recherche précise automatiquement l'organisme comme étant homo sapiens, et recherche dans les autres champs soit human, soit homo sapiens, justifiant la différence du nombre d'entrées.

Source : Homo sapiens (human)
Le nom usuel Human apparaît entre parenthèses.

Q12.
Lorsque l'on recherche sans préciser l'organisme, on observe des résultats associés à d'autres autres taxons et donc pas que du génome humain (qui est par exemple l'organisme hote) :
Homo sapiens
uncultured bacterium
Human immunodeficiency virus 1
Macaca mulatta
Pan troglodytes
...

Q13.

Dans tous les cas, l'organisme associé aux recherches human ou homo sapiens est mis par défaut comme étant homo sapiens OU un organisme lié à un des deux termes. Si l'on veut des résultats précis, il faut indiquer [orgn] dans la recherche afin d'être certains d'avoir homo sapiens ou human en organisme, qui dans les deux cas, donnent le même nombre de résultat : 10631878. L'utilisation des champs permet d'afffiner la recherche et d'éviter des confusions.

4. Réduction du nombre de réponses

Q14. Les séquences ne proviennent pas toute de l'humain (car il n'est pas précisé qu'il soit forcément l'organisme, il peut donc s'agir d'autres organismes avec un lien avec homo sapiens, qui serait par exemple un hôte, au niveau du champ *host*).

5. Interrogation des champs

Q15.

La syntaxe est classique (type syntaxe d'interrogation intuitive) avec des champs entre crochets, des opérateurs booléens, etc.

ex : homo sapiens[Organism]
On obtient le même nombre de résultats que précédemment : 10631878

6. Changement du format d'affichage

Q.16
requête : Arabidopsis thaliana[Organism] AND biomol_mrna[Properties] AND dikinase.
26 entrées, dont certaines sont redondantes (même gène, avec quelques différences de séquence).

Q17.
Lorsque l'on fait enregistrer sous, cela enregistre une page web au format html, qui n'est pas pratique.

Q18.
Pour enregistrer la séquence, il y a plusieurs façons à l'aide du SEND, certaines permettent directement de faire des analyses dessus, ou de l'enregistrer à notre clipboard/collection, sinon on peut enregistrer un fichier au format .fasta.

III. SRS

2. Interrogation des champs

Q19.

Requête : ABCB6 dans banque EMBL avec Standard Query Form, Features : Gene ==> 11 entrées

Q20.
La fonction n'est pas toujours précisée, parfois dans la description il est précisé : ATP-binding cassette.
Les gènes qui ont le même nom codent pour des protéines qui ont les même fonctions. (un gène code pour une protéine)

Q21.
Le nom du champ à interroger dans la banque UniProt est Gene Name : ABCB6
Nous trouvons 7 entrées.
Ce nombre est différent de celui obtenus sur EMBL car EMBL recense toutes les entrées dont le

nom de gène est ABCB6 et UniProt recense que les protéines codées par le gène ABCB6, or pas tous les gènes codent pour des protéines (il y a par exemple des transcrits viraux).

3. Liens entre banques, à partir d'une requête

Q22.
On trouve 3 entrées, alors qu'on s'attendait à 7 entrées. Cela est du au fait que la recherche ne se fait pas sur le gene ABCB6 de manière général, mais sur les entrées ABCB6 obtenus par EMBL.

5. Sous-entrées

Q.23
Taille des séquences :

```
AJ289233_2  2529 BP;
BC000559_5  2529 BP;
BC043423_6   906 BP;
BC006634_3  2529 BP;
BC085712_3  2511 BP;
DQ891876_6  2529 BP;
DQ895063_6  2526 BP;
AAFI02000049_10  2037 BP;
```

On retrouve en partie les mêmes résultats que dans la question III.2 : Nous retrouvons les 8 résultats correspondant aux CDS que nous observions avec la requête Gene Name : ABCB6 sur EMBL.