Bases de données sur les nucléotides et protéines

1. Introduction

Définitions

Bioanalyse (Bioinformatics) = application d’algorithmes, de modèles statistiques ayant pour objectifs d’interprété, classé et de comprendre les données biologiques

Recherche en bioinformatique (Computational Biology) = développement de modèles mathématiques et d’outils associés pour résoudre des problèmes biologiques.

Attention !

Attention aux résultats des logiciels :

* La qualité des résultats est parfois diminuée au profit de la vitesse
* Certains problèmes admettent un ensemble infini de possibilités -> ce n’est pas toujours la meilleure solution trouvée
* Certains logiciels ne fournissent que des prédictions

Attention aux bases de données :

* Les données ne sont pas toujours fiables
* Les données ne sont pas toujours à jour

Quelques champs d’applications

Bioanalyse de séquences biologiques : ADN, protéines, identification de gênes, ….

Bioanalyse pour les métabolites : Identification, annotation, ….

Bioanalyse structurale : Analyse du pliage des macromolécules biologiques

Bioanalyse des réseaux : Interactions entre les gênes, les métabolites, les protéines, ….

Bioanalyse sur des populations génétiques : modélisation de l’évolution d’une population dans un environnement spécifique

Bases de données – Définition

L’ensemble des données relatives à un domaine, organiser par traitement informatique accessible en ligne. Généralement, les données sont stockées sous format text files (.txt).

Il y a un besoin de développer un logiciel spécifique pour interroger les données contenues dans ces bases de données.

Il existe 3 principaux centres qui récolte et fournis en libre accès des séquences : EBI (UE) , NCBI (USA) , DDBJ (JPN).

Exemple des formats

Séquences ADN/ARN -> EMBL, GenBank, DDBJ

Séquence peptidique -> SwissProt, TrEMBL, …

Les formats les plus utilisés par les bioinformaticiens sont stockés les séquences sont : Fasta et Raw Sequence.

Pour convertir un format en un autre, on a par exemple le site Seqret (EBI)

Fasta

Utilisé par les logiciels d’analyse séquentielles.

|  |  |
| --- | --- |
| >ID, commentaires  Séquence brute | >Human Polycomb 2 homolog (hPc2) mRNA, partial cds ctccggcagcccgaggtcatcctgctagactcagacctggatgaacccatagacttgcgctcggtcaagagccgcagcgaggccggggagccgcccagctccctccaggtgaagcccgagacaccggcgtcggcggcg |

1. Bases de données des nucléotides

Elles ont pour origine le séquençage de l’ADN ou de l’ARN.

Stockage des données : 1 séquences + annotations (fragments, complet, ARNm, …) = 1 enregistrement

TOUTES les séquences (ADN ou ARN) sont écrites avec un « T ».