

RECHERCHE DANS UNE BASE DE DONNEES

✕ Utilisation de méthodes heuristiques

œ Rappel : l'alignement d'une séquence donnée avec toutes les séquences de la base de données ne peut s'effectuer en un temps acceptable.

↳ Développement de nouveaux algorithmes qui même s'ils ne garantissent pas de trouver l'alignement optimal sont très performants et efficaces pour des comparaisons de séquences

œ Deux des méthodes les plus connues sont **BLAST** et **FASTA**.

RECHERCHE DANS UNE BASE DE DONNEES

✕ Utilisation de méthodes heuristiques

‣ **BLAST : Basic Local Alignment Search Tool**

Alstchul, Gish, Miller, Myers and Lipman 1990

✕ Comparaison de deux séquences et recherche de toutes les paires de segments **similaires** dont le score excède un certain seuil de similarité S .

RECHERCHE DANS UNE BASE DE DONNEES

✕ **BLAST : Basic Local Alignment Search Tool (suite)**

✕ Quelques définitions:

- Paires de segments dont le score excède le seuil = high scoring segment pairs HSP
- Un segment est une **sous-séquence contigue** d'une des deux séquences.
- Une paire de segments correspond à deux segments de même longueur dans chacune des séquences.

RECHERCHE DANS UNE BASE DE DONNEES

✕ **BLAST : Basic Local Alignment Search Tool (suite)**

✕ - L'alignement des segments est défini sans gap.

↳ Le score global S de la paire de segments est donc simplement la somme des scores de substitution.

✕ Ceux dont le score global est maximal sont dénommés Maximum Segment Pair (MSP)

RECHERCHE DANS UNE BASE DE DONNEES

✕ **BLAST : Basic Local Alignment Search Tool (suite)**

✕ Identification des HSP et MSP :

Soit w un paramètre de longueur et T un paramètre de seuil:

↻ recherche de tous les mots de longueur w dans la base de séquences qui s'alignent avec les mots de la séquence requête avec un score d'alignement $> T$

↳ localisation de toutes les graines de similarité entre la séquence requête et la séquence de la base de données. Chaque réponse est appelé un « *hit* »

RECHERCHE DANS UNE BASE DE DONNEES

✂ **BLAST : Basic Local Alignment Search Tool (suite)**

↻ Chaque *hit* est étendu de façon à trouver s'il est contenu dans une paire de segments dont le score est supérieur à S .

↻ Une série d'alignements locaux sans gaps est donc obtenue

RECHERCHE DANS UNE BASE DE DONNEES

✂ **BLAST : Basic Local Alignment Search Tool (suite)**

⌘ *Actuellement il est possible d'améliorer les alignements en sélectionnant les réponses l'étape précédente et en effectuant un alignement local avec gap par programmation dynamique*

⌘ *En général $w = 3$ à 5 pour les acides aminés et pour les acides nucléiques.*

RECHERCHE DANS UNE BASE DE DONNEES

✂ FASTA (Lipman et Pearson 1985)

⌘ *Principe: FASTA compare une chaîne de caractère issue de la séquence requête et une chaîne de caractère dans la base de données de séquences. Il fait l'hypothèse que dans un alignement, on peut s'attendre à trouver des segments dans lesquels il existe une totale identité.*

RECHERCHE DANS UNE BASE DE DONNEES

< FASTA (Lipman et Pearson 1985)

œ *Paramètres : ktup nombre entier*

Etape 1:

œ *On recherche toutes les sous-chaînes de longueur ktup identiques entre les deux chaînes.*

œ *De telles sous-chaînes sont appelées hot spots*

œ *Des hots spots consécutifs sont situés sur les diagonales de la matrice de programmation dynamique.*

RECHERCHE DANS UNE BASE DE DONNEES

< FASTA (Lipman et Pearson 1985)

Etape 2:

∞ On va chercher les N meilleurs chemins diagonaux ($N=10$ en général) i.e des segments de haute densité situés sur la même diagonale mais qui peuvent être séparés par des espaces.

↪ Le score de la diagonale est la somme des différents scores de hot spots et espaces en utilisant une matrice de scores type PAM ou

RECHERCHE DANS UNE BASE DE DONNEES

FASTA (Lipman et Pearson 1985)

ape 2 uite)

Les meilleurs segments diagonaux en termes de score sont conservés. Chaque région est donc un alignement local sans gap.

Score = $init1$

ape 3

S'il existe plusieurs régions initiales dont les scores sont supérieurs à un seuil donné, on examine des bandes (largeur 16 si $ktup = 2$, 32 si $ktup = 1$) autour de la diagonale de façon à vérifier si les différents segments peuvent être rejoints sans des gaps (introduction d'un coût de gap très élevé ~ 20)

RECHERCHE DANS UNE BASE DE DONNEES

FASTA (Lipman et Pearson 1985)

ape 4

œ Pour les séquences dont les scores sont supérieurs à un seuil donné, on construit l'alignement optimal de la séquence requête avec la séquence de la base de données. (score opt

marques:

œ Estimation statistique : Quand les 60000 premiers scores ont été calculés, on normalise les scores de similarité en utilisant une estimation des paramètres statistiques de la distribution de la valeur extrême. On calcule des valeurs Z (scores normalisés de moyenne nulle et de variance 1). Le calcul est répété pour les séquences de la base de données après élimination des

RECHERCHE DANS UNE BASE DE DONNEES

PSIBLAST : Utilise des matrices position-spécifiques

Principe:

- ✘ **Requête sur une banque de séquences**
- ✘ **Récupère les séquences avec $Evalue < 10^{-2}$**
- ✘ **Séquences trop similaires % Identité $> 98\%$ éliminées**
- ✘ **Alignement multiple**
- ✘ **Construction d'une matrice de fréquences position spécifique (rapport des log-vraisemblances) : PROFIL**
- ✘ **Utilisation du profil pour lancer une nouvelle requête et récupérer de nouvelles séquences.**
- ✘ **Et on itère jusqu'à ce qu'aucune ne puisse être**