Noura Nouali
https://github.com/nnouali/Projet_blast.git

Rapport projet court

Utilisation de l'algorithme BLAST avec les MSP pour l'Analyse de Séquences et l'Identification de Motifs

Année 2023-2024

Introduction :

BLAST[1] (Basic Local Alignment Search Tool) est un outil en bioinformatique. Il permet d'identifier des régions de similarité locale entre des séquences biologiques. Blast est capable de comparer des séquences de nucléotides ou de protéines à des bases de données de séquences et calcule la signification statistique des correspondances. Ce programme peut être utilisé pour déduire des relations fonctionnelles et évolutives entre les séquences, ainsi que pour aider à identifier les membres de familles de gènes.
Les segments de séquence sont des entités continues de résidus de longueur variable . Pour évaluer la similarité entre deux segments de même longueur, on calcule le score de similarité pour deux segments alignés de la même longueur, soit la somme des valeurs de similarité entre chaque paire de résidus alignés.
Le Maximum-scoring segment pair(MSP) est définie comme étant le meilleur score obtenu parmi tous les couples de fragments possibles que peuvent produire deux séquences. Ce score fournit une mesure de la similarité locale pour n'importe quelle paire de séquences.[1,2]
Ainsi, une paire de segments est localement maximale si son score ne peut pas être amélioré en étendant ou en raccourcissant les deux segments.
Dans ce projet, nous utiliserons l'algorithme de BLAST avec l'approche des MSP (Maximal-scoring Segment Pair) pour l'analyse de séquences et l'identification de motifs.

Matériel et Méthodes :

La création et conception d'une base de données de séquences protéiques a été essentielle dans ce projet.
La base de données a été constituée à partir de séquences protéiques courtes retrouvées dans Uniprot[3] (Universal Protein Resource) qui offre une vaste collection de données sur les protéines, leurs fonctions et leurs interactions.
Nous découpons ensuite les séquences de notre base de données en mots de trois lettres chevauchantes. Chaque mot de trois lettres est associé à une position et à un identifiant décrivant la séquence protéique

Pour la recherche de motifs, nous avons segmenté la séquence d'entrée en mots de trois lettres, en utilisant la position du premier acide aminé trouvé dans chaque mot pour déterminer son emplacement dans la séquence d'entrée (ce qui nous donne tous les mots chevauchants). Nous avons défini un seuil de similarité(Threshold), fixé à 11 afin de filtrer les mots de trois lettres, la taille de la longueur des mots recherchés a été fixée à 3 (w, word_length). Les mots de trois lettres dont le score de similarité est supérieur ou égal au seuil ont été conservés pour l'analyse ultérieure.
Les mots de trois lettres satisfaisant le critère du seuil ont été enregistrés dans un fichier nommé "mots_voisins_score.txt". Les mots voisins ont ensuite été étendus à partir de la séquence requête pour obtenir la séquence complète à analyser.

Nous avons par la suite aligner la séquence obtenue et la séquence trouvée dans la base de donnée à partir du mot voisin. Les scores de similarité sont calculés en utilisant la matrice de similarité BLOSUM62.
On considère les extensions à la fois du côté gauche et droit du mot voisin pour que le calcul des scores soit correct.
De plus, les scores cumulés ont été utilisés pour déterminer les bornes maximales à gauche et à droite du mot voisin. En effet, on cherche à étendre la similitude dans les deux directions le long de chaque séquence, à partir du mot commun, de manière à ce que le score cumulé puisse être amélioré (ce qui nous a permis d'identifier le score cumulé maximal à gauche et à droite).

L'alignement final est obtenue à partir des bornes maximales des scores cumulés. Celui-ci a été utilisé pour recalculer le score de l'alignement final (entre la séquence requête contenant le mot voisin ou d'origine et la séquence de la base de donnée correspondante).
Le score le plus élevé a été identifié comme le plus significatif.

Suite à l'identification de la séquence ayant le score le plus élevé, il est essentiel de calculer la E-value.[4] L'E-value est une mesure statistique qui estime la probabilité qu'un alignement donné entre les séquences se produise par hasard.
Elle se calcule de cette façon:

$$E = K * m * n * \exp(-\lambda * S)$$

où :

- E est la E-value
- K est une constante de réajustement dépendant du contexte de la recherche.
- m est la taille effective de la base de données de recherche.
- n est la taille effective de la séquence requête.
- λ est le taux de décroissance de la distribution des scores.
- S est le score du meilleur alignement trouvé entre la séquence requête et une séquence de la base de données.

Plus laE-value est faible, plus cela indique que l'alignement est très peu susceptible de se produire par hasard et suggère une similitude significative entre les séquences.

Résultats:
Dans cette étude, nous avons utilisé une séquence requête d'un exemple connu[5], "YANCLEHKMGS", et dans notre base de donnée nous avons mis la séquence "DAPCQEHMRGWPNDC" afin de comparer au mieux nos résultats.

Après avoir appliqué notre méthode, nous avons obtenu un meilleur score d'alignement de 41.0.

```
Request sequence    : APCLEHKMG
                      :||| |||  |
Sequence database   :  APCQEHKRG
```

Distribution des Scores

Voici la distribution des scores d'alignement que nous avons obtenue grâce à notre approche :

- Score : 32.0, Séquence : ANCLEHKMG
- Score : 38.0, Séquence : ANCLEHKRG
- Score : 39.0, Séquence : ANCQEHKMG
- Score : 41.0, Séquence : APCLEHKMG

Nous retrouvons bien la sequence “ANCQEHKMG” de l’exemple avec un score de 39

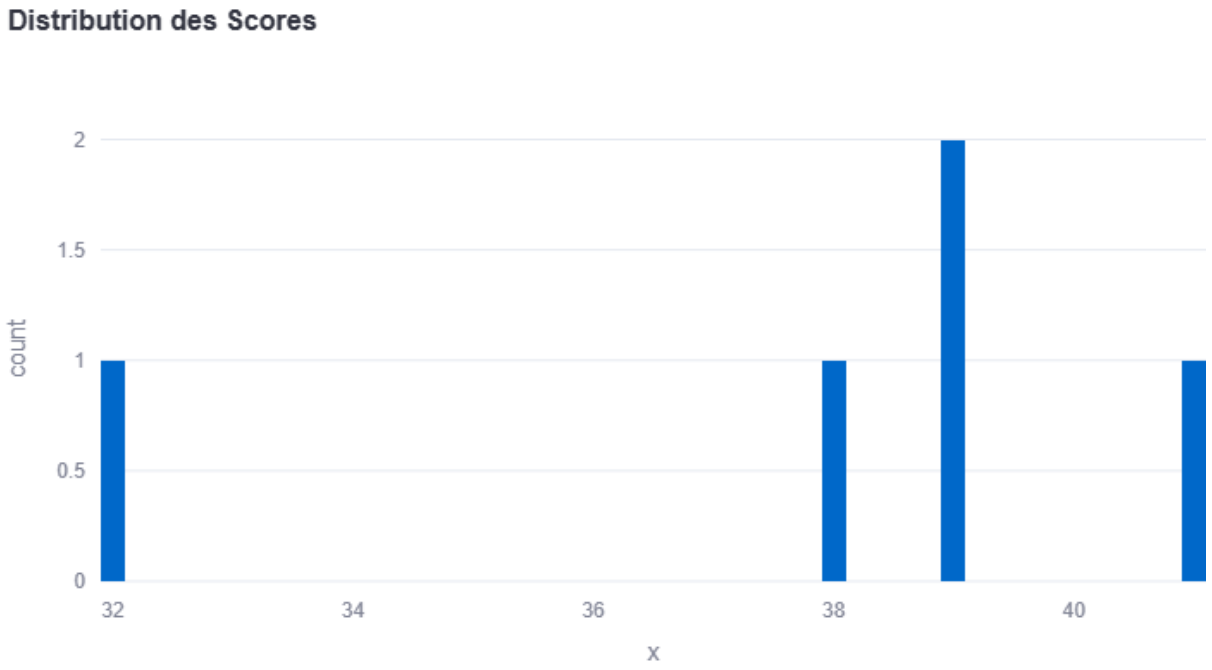

Figure1: Distribution des scores des alignements trouvés.

Les E-values que nous avons calculées indiquent la probabilité de trouver des correspondances aussi significatives que celles que nous avons observées par hasard.
Dans notre cas, toutes les E-values sont petites, ce qui signifie que les correspondances que nous avons trouvées sont significatives.

- E-value : 0.0001049486150495286, Séquence : ANCLEHKMG
- E-value : 2.1146424776249918e-05, Séquence : ANCLEHKRG
- E-value : 1.6191248008942767e-05, Séquence : ANCQEHKMG
- E-value : 9.492204187622978e-06, Séquence : APCLEHKMG

Ces E-values confirment la pertinence de nos alignements et suggèrent que les correspondances que nous avons identifiées ne sont pas le résultat du hasard, renforçant ainsi la validité de notre approche pour l'analyse de séquences et l'identification de motifs.

Conclusion :
Dans ce projet nous avons réalisé un outil similaire à Blast utilisant avec l'approche des MSP pour analyser des séquences biologiques et identifier des motifs significatifs.

Les résultats de notre analyse ont révélé des correspondances significatives et des motifs conservés entre ces séquences. Le meilleur alignement obtenu avait un score de 41.0, ce qui indique une similarité locale forte entre les deux séquences.
De surcroit, la distribution des scores a montré plusieurs correspondances potentielles, suggérant la présence de motifs conservés.
Les valeurs de signification statistique (E-values) que nous avons calculées pour nos correspondances étaient toutes inférieures à des seuils significatifs.
Ces E-values renforcent la validité de notre méthode pour l'analyse de séquences et l'identification de motifs.

Références:

1. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
2. Pertsemlidis, A. & Fondon, J. W. Having a BLAST with bioinformatics (and avoiding BLASTphemy). *Genome Biol.* **2**, reviews2002.1-reviews2002.10 (2001).
3. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
4. Durand, D. BLAST: Target frequencies and information content. (2015).
5. Blast Algorithm | PPT. https://www.slideshare.net/Fardin6600/blast-algorithm.