Data Mining Tool for Biological Sequences

publié par Huiqing Liu et Limsoon Wong en 2003

Résumé : la publication fait office de tutoriel expliquant les pratiques de data mining liés à la reconnaissance de génomes dans une séquence d’ADN.

Elle montre les trois étapes nécessaire à l’exécution du processus :

1. Générer des fonctionnalités candidates pour les séquence en utilisant “k-gram”
2. Sélectionner les fonctionnalités utiles (entropie, gain d’information…)
3. intégrer les fonctionnalités choisies dans un algorithme de prise de décision

Le problème majeur du minage d’ADN est que les données sont très mal connues et ne semblent pas avoir de logique.

Le papier utilise l’exemple de la reconnaissance de “Translation Initiation Sites” (TIS) dans l’ADN afin de mieux comprendre les acides aminés résultant du génome après être passé par l’ARN messager.

*Note : la définition de TIS et son utilité est évoquée dans la publication mais n’est pas vraiment claire. Il peut être nécessaire de faire plus de recherche pour mieux comprendre le mécanisme.*

Pour se faire, les auteurs utilisent le dataset de Pedersen et Nielsen et étudient les résultats obtenus par plusieurs publications utilisant elles même le même dataset

Les auteurs indiquent cependant que l’obtention de données d’ADN est un processus complexe. En effet, obtenir un séquence complète sans la détruire ou la contaminer est compliqué.

Compréhension des données et sélection des attributs.

Lors de leur première recherche, Pedersen et Nielsen ont mis au point un réseau de neurones permettant d’évaluer l’utilité de chaque nucléotide. (notés A C G ou T).

Cet algorithme est appelé NetStart 1.0 et est disponible à l’adresse suivante

<http://www.cbs.dtu.dk/services/NetStart/>

Cet algorithme a notamment permis de comprendre que la position -3 d’un séquence est cruciale pour reconnaître un ATG étant TIS d’un non TIS.

Ensuite, il a été définit que la reconnaissance d’un TIS revient à analyser la seconde occurrence des nucléotides ATG d’une séquence. C’est sur cette propriété que le reste de la publication va s’appuyer.

Ensuite les auteurs expliquent la structure d’une séquence et comment elle divisée en fonction du second ATG.

Les nucléotides apparaissant avant le second ATG sont notés comme “upstream” et les autres comme “downstream”.  
Si une séquence contient moins de 100 nucleotides upstream ou downstream, les nucléotides manquants sont remplacés par des bits de bourrage (nommé “don’t care symbols”).

A partir de cette représentation, il est possible d’utiliser l’algorithme k-grams pour analyser les nucleotides downstream et upstream séparément et calculer la fréquence d’apparitions des différentes séquences de nucléotides (séquences de 3, 2 et 1)

Plusieurs variations de k-grams peuvent être utilisé pour arriver à ce résultat.

A la fin de ce processus, on arrive avec un “feature vector”: tableau listant le nombre d’occurence de chaque sous-ensemble d’une séquence.

D’autre algorithmes sont aussi rapidement présentés.

L’étape suivante consiste à tenter donner une classe aux différents éléments. Ces classes sont :

A : TIS

B : non-TIS

Une fois arrivé à ce point, on dispose d’un ensemble de données au nombre d’attributs réduits et aux valeurs simplifiées. Ainsi il est possible d’y appliquer des techniques classique de data mining (C4.5, SVM etc)

Enfin, la publication effectue un comparaison des différents résultats obtenus par d’autres chercheurs sur ce sujet.

Conclusion

Cette publication montre que la principale difficulté de l’ADN mining est de réussir à comprendre les séquences d’ADN et de sélectionner les attributs utiles à la classification de sites ATG.

Une fois cela fait, les données interprétées peuvent être utilisées dans des algorithmes classique comme n’importe quel autre type de données

L’article fait un bon état de l’art en matière d’ADN mining et expose les différentes alternatives possible à la résolution d’un problème donné (classification TIS). Cependant, la notion de TIS, site ATG etc est compliquée et demande de plus amples recherches pour être comprises et ainsi pour comprendre les problématiques rencontrées.

Par contre, si les difficultés et les enjeux sont compris, je pense que cela permet d’avoir une bonne connaissance global du domaine d’études.