PRÉDICTION DU SITE DE CLIVAGE DU PEPTIDE SIGNAL EN UTILISANT L'APPRENTISSAGE SUPERVISÉ

16 mai 2024

INF442

Ruflin TSASSE KIAMPI, Ouattara ABOUBAKAR

ÉCOLE POLYTECHNIQUE

IP PARIS

TABLE DES MATIÈRES

1 INTRODUCTION

Le tri des protéines subcellulaires, qui désigne les processus de transport des protéines vers leur destination finale au sein d'une cellule, est crucial pour le bon fonctionnement de la vie cellulaire. Dans de nombreux cas, ce tri repose sur des « signaux » identifiables dans la structure primaire des protéines. Par exemple, le ciblage vers les voies sécrétoires, les mitochondries et les chloroplastes dépend généralement d'une séquence N-terminale spécifique ou d'un peptide ciblant, reconnus par les récepteurs à la surface de l'organelle appropriée. Une fois ciblées, des mécanismes de translocation membranaire assurent la livraison des protéines à l'intérieur de l'organelle. Le clivage des protéines est donc un processus crucial dans la biologie cellulaire, assurant leur livraison vers leurs destinations appropriées. Celà comprend l'identification des séquences de peptides signal et de leurs sites de clivage, ce qui est d'un grand intérêt pour la conception de médicaments et la compréhension des processus biologiques.

Les travaux de prédiction de sites de clivage ne datent pas d'aujourd'hui. En effet, ils débutent sur les peptides de signaux sécrétoire (Von Heijne,1983 et 1986 [1]) avec l'application des techniques modernes d'apprentissage automatique, telle que les réseaux de neurones (NN).

Dans ce projet, il sera question pour nous de prédire les sites de clivage des protéines en utilisant un modèle *Support Vector Machine* (SVM) qui est parfaitement adapté aux taches de reconnaissance de formes où des quantités relativement importantes de données sont présentes.

2
APPROCHE

Nous disposons de trois versions du peptide signal et donc trois varietés d'organismes, eucaryotes, Gram-positif et Gram-négatif procaryotes. Chaque organisme est identifié par sa sequence et son annotation qui est en fait l'encodage des différents acides aminés (signal peptide S et mature protein M) qui la constitue et la position de son site de clivage.

201 séquences eucaryotes, 28 Gram-positif procaryotes et 53 Gram-négatif procaryotes dont ont connais leurs sites de clivage ont été sélectionnées pour l'évaluation du modèle. Il a été montré dans l'article [1] que les sites de clivage peuvent etre caractériser par leurs voisins les plus proches qui sont définis à partir des p acides aminés avant le site et des q acides aminés après.

2.0.1 • Encodage des données :

Pour une manipulation plus aisée des données, nous encodons les lettres représentant les acides aminés par des vecteurs de taille 26 avec la valeur 1 sur le rang d'apparition de l'acide dans l'alphabet. Ainsi, l'encodage d'une séquence de protéine s'obtient en concatenant les vecteurs encodés des différents acides aminés qui la constitue. Une approche directe utilisant les machines à vecteurs de support (SVM) et quelques noyaux nous permettra de faire une prédiction des différents sites. La taille des voisins proches (p + q) est optimisée afin d'avoir une meilleure performance du modèle.

2.0.2 • Labelisation :

Les labels 0, 1 sont obtenus en selectionnant de proche en proche un voisin de longueur p + q et en vérifiant si la position courante est sur un site C (annotation de la protéine) ou non. Cette méthode fait considérablement accroitre la taille entière des données.

2.0.3 • Implémentation du modèle SVM :

Le SVM a été implémenté en utilisant la librairie Scikit-Learn de Python. Pour l'apprentissage des classificateurs SVM sur des données d'entrées, différentes tailles d'acides aminés ont été implémenté. Chaque fenêtre représente une caractéristique spécifique ; Soit elle représente une fenêtre de clivage, si le site de clivage (C) se trouve à sa position correspondande, soit elle représente une fenêtre de non-clivage dans toutes les autres conditions. Le site de clivage réel se situe entre le résidu central et le résidu C-terminal qui le suit dans la séquence. Les classificateurs prédisent que le résidu central est soit un site de clivage, soit un site de non-clivage pour une configuration de fenêtre particulière. Chaque acide aminé était représenté en utilisant 20 (acides aminés standard) + 6 positions d'encodage binaire (encodage clairement conventionnel). Les 20 acides aminés ont été encodés comme suit : A = 10000000000000000000000000, G = 00000010000000000000000000, etc.

2.0.4 • Evaluation et mesures de performances :

Une technique de validation hold-out a été utilisée pour évaluer les performances de notre meilleur classificateur. L'ensemble de données des sites de clivage et de non-clivage a été divisé de manière aléatoire en deux sous-ensembles contenant des ratios de sites de clivage et de non-clivage. Les classificateurs ont été entraînés sur un ensemble et les performances ont été évaluées sur l'ensemble restant en jouant sur le type de noyau à implémenter. En identifiant le meilleur noyau et en évaluant les performances via les paramètres d'évaluation des modèles de classification (*accuracy, f-score, precision, recall*).

3
RÉSULTAT ET DISCUSSION

Après avoir recherché le meilleur noyau pour notre modèle, avec les valeurs de p = 13 et q = 2, nous avons constaté que le noyau rbf était le plus performant parmi ceux testés. Nous l'avons donc utilisé pour entraîner notre modèle. En utilisant ce noyau, nous avons obtenu les résultats suivants :

Metrique	**Temps d'entraînement**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Valeur	3.88	0.774	0.739	0.489	0.588

Table 1 – Résultats de l'entraînement du modèle avec noyau RBF

Ces résultats indiquent que notre modèle présente une précision relativement élevée, mais une sensibilité et une spécificité plus faibles. Cela suggère qu'il peut encore être amélioré pour une meilleure prédiction sur l'ensemble des données.

concernant les predictions sur de nouvelle donnée, notre modèle de prédiction n'est pas homogène sur tous les types d'organismes. Il parvient à prédire avec une précision remarquable de 90% pour les organismes de type GRAM+ procaryote, tandis qu'il atteint seulement environ 69% pour les organismes de type Eucaryote et GRAM- procaryote. Ces résultats suggèrent que notre modèle est plus adapté à la prédiction sur les organismes de type GRAM+ procaryote, mais nécessite des améliorations pour les autres types d'organismes.

	Euk	**Gneg**	**Gpos**
Accuracy	0.696	0.692	0.907
Precision	0.712	0.715	0.940
Recall	0.313	0.289	0.778
F1-score	0.435	0.411	0.851

Table 2 – Résultats de l'évaluation par classe

Quant aux resultats de l'optimisation des paramètre p et q, notre algorithme n'a malheureusement pas terminé de tourner. Nous presenterons donc les resultats de ce dernier lors de la soutennce. Vous trouverez en annexe le code qui a été ajouté pour faire l'optimisation de ces paramètre.

4 CONCLUSION

Une nouvelle méthode basée sur SVM est proposée pour la discrimination des protéines sécrétées en utilisant un encodage approprié de ces protéines. Le modèle décrit l'information évolutive de 20 acides aminés dans une séquence protéique. Les variables p=13 et q=2 intègrent les effets taille des plus proches voisins des séquence. Les résultats de prédiction montrent que les performances du modèle ont été améliorées en choisissant un noyau approprié (ici RBF). Lorsqu'elle est appliquée sur les ensembles de test, notre méthode a obtenu un résultat de prédiction acceptable avec des précisions de 69,6 %, 69,2 %, et 90,7 % pour les trois types de protéines sécrétées, respectivement. Comme piste d'amélioration du modèle, on pourra optimiser suivant les valeurs des paramètres des différents noyaux en occurrence C et gamma. Par conséquent, nous espérons que notre méthode sera utile pour discriminer différents types de protéines sécrétées en l'absence de données expérimentales et élucider la fonction biologique de nouvelles protéines sécrétées découvertes.

5 BIBLIOGRAPHIE

[1] Gunnar von Heijne. A new method for predicting signal sequence cleavage sites. *Acids Research*,14(11) :4683–4690, 1986.

[2] Jean-Philippe Vert. Support Vector Machines prediction of signal peptide cleavage site using a new class of kernels for strings. In : *Proceedings of the 7th Pacific Symposium on Biocomputing*, pp. 649–660, 2002.

[3] https ://en.wikipedia.org/wiki/Additive_smoothing.

[4] https ://en.wikipedia.org/wiki/Substitution_matrix.

6 ANNEXE

Listing 1 – Optimisation des paramètres

```
def optimize_parameters(df, kernels, p_values, q_values):
    best_accuracy = 0
    best_params = {'kernel': None, 'p': None, 'q': None}

    for p in p_values:
        for q in q_values:
            features, labels = extract_features(df, p, q)
            X_train, X_test, y_train, y_test = train_test_split(features, labels,
            test_size=0.1, random_state=42)

            for kernel in kernels:
                accuracy, _, _, _= train_and_evaluate(X_train, X_test, y_train,
                y_test, kernel)

                if accuracy > best_accuracy:
                    best_accuracy = accuracy
                    best_params = {'kernel': kernel, 'p': p, 'q': q}

    return best_params, best_accuracy
```

Listing 2 – Execution de la fonction optimisation

```
best_params, best_accuracy = optimize_parameters(df_train, kernels, p_values=p_plage,
q_values=q_plage)

print("Meilleur kernel:", best_params['kernel'])
print("Meilleur p:", best_params['p'])
print("Meilleur q:", best_params['q'])
print("Meilleure precision:", best_accuracy)
```