REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOHAMED BOUDIAF - M'SILA
FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE

DEPARTEMENT D'INFORMATIQUE

MEMOIRE de fin d'étude

Présenté pour l'obtention du diplôme de MASTER

Domaine : Mathématiques et Informatique

Filière : Informatique

Spécialité : Réseaux.

Par: Benseni Noureddine

SUJET

Adaptation et implémentation de la méthode des réseaux de neurone pour la prédiction des structures secondaires de protéine

Soutenu publiquement le : 01/06/2016 devant le jury composé de :

Nasreddine Amroune	Université de M'sila	Président
Rached YAGOUBI	Université de M'sila	Rapporteur
Ali dabba	Université de M'sila	Examinateur

Promotion : 2015 /2016

Table des matières :

Introduction générale :

Depuis bien longtemps les biologistes ont eu le besoin de faire des calculs sur les données à fin de pouvoir établir des nouvelles lois biologiques .Cependant, avec le développement des ordinateurs puissants et la grande disponibilité des données biologiques (des séquences d'ADN, d'ARN ou de protéines), une nouvelle discipline est apparue appelée la bio-informatique.

La bio-informatique est une discipline récente, qui fait appel aux compétences de la plupart des disciplines scientifiques pour lesquelles il existe déjà des méthodes permettant de résoudre des problèmes analogues. Il y a principalement les mathématiques et l'informatique, mais également dans certains cas la physique ou la chimie. La bio-informatique regroupe donc une partie de chacun de ces domaines, ainsi que la biologie elle-même. La bio-informatique traite différents problèmes parmi lesquelles nous citons : la phylogénie, la recherche de motifs, l'alignement de séquences et la prédiction des structures.

Le problème de la prédiction de structure secondaire d'une protéine à partir de sa seule séquence en acides aminés est un problème très difficile, il existe déjà des méthodes permettant de traiter ce problème parmi lesquelles nous citons : les méthodes statistiques, la méthode des plus proches voisins, la méthode de Chaînes de Markov, la méthode d'apprentissage par réseaux de neurones

Dans ce mémoire Nous avons essayé d'adapté et implémenté la méthode d'apprentissage par réseaux de neurones pour résoudre ce problème.

Notre mémoire est organisé en quatre chapitres. Le premier contient une introduction à la bio-informatique, pour cela nous présentons quelques notions de biologie moléculaire et les différents thèmes de la bio-informatique. Dans Le deuxième chapitre nous allons présenter les principales méthodes pour traiter le

problème de la prédiction de structure secondaire. Le troisième contient une introduction aux réseaux de neurones qui constituent actuellement un des outils les plus efficaces pour le traitement des problèmes de classification. Dans Le dernier chapitre nous allons présenter les outils d'implémentation que nous avons exploitée pour l'algorithme et voir comment adapter le réseau de neurone pour la prédiction des structures secondaires de protéine.

Conclusion générale :

Au cours de ce travail de master, Nous avons essayé d'adapter et implémenter la méthode d'apprentissage par réseaux de neurones pour résoudre Le problème de la prédiction de structure secondaire d'une protéine. Plusieurs méthodes ont été proposées pour la résolution de ce problème comme les méthodes statistiques, la méthode des plus proches voisins, la méthode de Chaînes de Markov mais aucune ne peut le résoudre efficacement dans tous les cas.

Notre implémentation combine trois étapes, Le premier contient la partie de codage des acides aminés sous la forme de vecteurs de 0 et 1. Dans la deuxième étape nous avons adapté le réseau de neurones multicouches du type feedforward pour prédire la structure secondaire. Dans La dernière étape nous avons décodé le résultat des réseaux de neurones sous la forme de lettres H pour l'hélice α E pour le feuillet β C pour le coudes.

Nous avons testé notre implémentation sur différents structures de protéines. En fournissant un ensemble de tests de prédiction prend en compte la relation entre le degré d'apprentissage et la taille de la fenêtre et la qualité des résultats.

Nous avons remarqué que lorsque nous augmentons le degré d'apprentissage et la taille de la fenêtre, ça touche positivement la qualité des résultats, dans notre implémentation le meilleur résultat de prédiction 63% avec une taille de fenêtre 10 et degré d'apprentissage 30.

Comme perspective, nous intentons d'hybrider cette approche avec la méthode de plus proche voisin ainsi qu'utiliser la méthode de classification de protéine afin d'améliorer les résultats.

BIBIIOGRAPHIE :

[1] J.MULLER "Analyse du cytosquelette par des approches bio-informatiques à haut débit de génomique comparative et de transcriptomique.". Thèse de doctorat l'Université Louis Pasteur Strasbourg 1,2006.

[2] Ramachandran ET Sasisekharan. "Conformation of polypeptides and proteins 1968".

[3] Watson E T Crick. "A structure for deoxyribose nucleic acid." F.H.C, 1953.

[4] Alain Raisonnier. "Structures Biologiques." Université Paris-VI, 2010.

[5] V. Derrien, "Heuristiques pour la résolution du problème d'alignement multiple", Thèse de doctorat, Université d'Angers, 2008.

[6] Isabelle SOURY-LAVERGNE NAVIZET, "MODÉLISATION ET ANALYSE DES PROPRIÉTÉS MÉCANIQUES DES PROTÉINES", Thèse de doctoratl'UNIVERSITÉ PARIS 6, 2004.

[7] Azaquar, http://www.azaquar.com/iaa/chimie/ca_images/ca_proteine3.gif, consulté le 15/2/2016.

[8] D Robert et B Vian. "Element de biologie cellulaire." 2008.

[9] I.Ruczinski. " Protein Structure Prediction: Secondary Structure. Department of Biostatistics," Johns Hopkins University, 2008.

[10] G.Chakroun "Prédiction de la structure d'une protéine "2004.

[11] C. Gibas and P. Jambeck. "Introduction à la bioinformatique. O'Reilly", 2002.

[12]Wikipedia,https://fr.wikipedia.org/wiki/Prot%C3%A9ine#Pr.C3.A9diction_de_structure_et_simulation consulté le 18/3/2016.

[13]Wikipedia,https://fr.wikipedia.org/wiki/Alignement_de_s%C3%A9quences. Consulté le 18/3/2016.

[14] Cock PJ., Fields CJ., Goto N., Heuer ML. & Rice PM., "The Sanger FASTA file format for sequences with quality scores, and the Solexa/Illumina FASTA variants. ", Nucleic Acids

Research,vol.38,no6,2010,p.176771(ISSN 13624962, PMID 20015970,DOI 10.1093/nar/gkp 1137).

[15]William R. Pearson, "Documentation des versions 3.x de la suite de programmes FASTA ", sur Center for Biological Sequence analysis .

[16]Wikipédia, https://fr.wikipedia.org/wiki/GenBank , consulté le 22/4/2016.

[17]Wikipédia, https://fr.wikipedia.org/wiki/Protein_Data_Bank#Le_format_PDB, consulté le 22/4/2016.

[18] B. Rost. Prediction in id: secondary structure, membrane helices, and accessibility. Methods Biochem Anal, 44:559–87, 2003.

[19] Chou ET Fasman. "Prediction of protein conformation. Biochemistry". Springer, 1974.

[20] M. Zaki ET C.Bystroff." Protein Structure Prediction." Oxford, 2008.

[21] Blaise Gassend, Charles W. O'Donnell, William Thies, Andrew Lee, Marten van Dijk et Srinivas Devadas. "Secondary Structure Prediction of All-Helical Proteins Using Hidden Markov Support Vector Machines. Computer Science and Artificial Intelligence Laboratory (CSAIL) " 2005.

[22] Juliette Martin. "Prédiction de la structure locale des protéines par des modèles de chaine de Markov Caché. Université" Paris VII - Denis Diderot, 2005.

[23] B. Bergeron. "Bioinformatics Computing. Prentice Hall PTR", 2002.

[24] B.Messabih et Hafida Bouziane, Belhadri Messabih et Abdellah Chouarfia. Prédiction de la Structure des Protéines par Apprentissage Automatique. SETIT 2009, 2009.

[25] Wikipédia ,https://fr.wikipedia.org/wiki/Structure_secondaire, le 22/4/2016.le 25/4/2016.

[26]Wikipedia,https://fr.wikipedia.org/wiki/R%C3%A9seau_de_neurones_artificiels, Consulte le 12/05/2016.

[27] Philippe POINCOT " Classification et recherche d'information bibliographique par l'utilisation des cartes auto-organisatrices, applications en astronomie "Thèse de Doctorat université louis pasteur 1999.

[28] L.Personnaz, I.Rivals, "Réseaux de neurones formels pour la modélisation, la commande, et la classification", CNRS éditions, collection Sciences et Techniques de l'Ingénieur, 2003.

[29] G.Z winngelsten, " Diagnostic des défaillances : théorie et pratique pour les systèmes industriels ", Ed. Hennes Pane, 1995.

[30]developpez,http://alp.developpez.com/tutoriels/intelligence-artificielle/reseaux-de-neurones/. Consulte le 12/05/2016.

[31]T.A.Freman, D.M.Skapura,"Neural networks: algorithm, applications and programming techniques ", CNS, Computation and neural systems series, 1992.

[32] Marc Parizeau, "*Réseaux de Neurones* (Le perceptron multicouche et son algorithme de retropropagation des erreurs) ", Université Laval, Laval, 2004.

[33]UniversitéLaval3, http://www.grappa.univ-lille3.fr/polys/apprentissage/sortie005.html. Consulte le 12/05/2016.

[34] Wikipedia, https://en.wikipedia.org/wiki/Unsupervised_learning. Consulte le 12/05/2016.

[35] java, http://www.java.com/fr/download/faq/whatis_java.xml. Consulte le 15/05/2016.

[36] Wikipédia, https://fr.wikipedia.org/wiki/NetBeans, Consulte le 15/05/2016.

ملخص :

التنبؤ بالبنية الثانوية للبروتين مرحلة هامة في الطريق نحو معرفة البنية ثلاثية الأبعاد للبروتين و وظيفته.وفي حين ان التقنيات التجريبية التي تقوم بالتنبؤ بالبنية الثانوية للبروتين (الأشعة السينية، الرنين المغناطيسي النووي) أصبحت مكلفة جدا وبطيئة جدا في كثير من الأحيان النتائج مشكوك فيها مع حجم المعلومات التي لا تزال تنمو.أصبح من المنطقي التوجه لاستخدام أساليب التعلم الآلي .هاذ العمل هو عبارة نهج قائم على التعلم الآلي عن طريق الشبكات العصبية متعددة الطبقات للتنبؤ بالبنية الثانوية للبروتين انطلاقا من تسلسل الاحماض الأمنية ، وقمنا في هاذا العمل بمحاولة تطبيق وتكيف الشبكات العصبية لغرض الوصول لنسبة تنبؤمقبولة .

الكلمات المفتاحية: التنبؤ بالبنية الثانوية ، البروتين ، التعلم الآلي ،الأحماض الأمنية، الشبكات العصبية.

Résumé :

La prédiction des structures secondaire des protéines est une étape importante sur le chemin pour définir sa structure tridimensionnelle et sa fonction. Tandis que les techniques expérimentales qui jouent le rôle de prédire la structure secondaire (diffraction des rayons X, Résonnance Magnétique Nucléaire) sont devenues très couteuses et trop lentes aux résultats souvent douteux avec un volume d'information qui ne cesse de croitre. Devenir une approche logique le recours vers les méthodes d'apprentissage automatique. Ce travail est une approche fondée sur l'apprentissage automatique à travers les réseaux neuronaux multicouches pour prédire des structures secondaire des protéines à partir de la séquence des acides aminés, et nous avons essayé dans ce travail d'implémenter et adapter les réseaux de neurones afin d'atteindre un pourcentage de prédiction assez acceptable.

Mots clés : Prédiction des structures secondaire, protéines, apprentissage automatique, acide aminées, les réseaux de neurones.

Abstract:

The prediction of secondary structure of proteins is an important step to define its tridimensional structure and its function. While the experimental technics that do the prediction of secondary structure of the protein (diffraction of X-ray, Nuclear magnetic) became very expensive and extremely slow, often the results are very dubious with the huge quantity of information that are still growing. It became very logical to take the attention to the automatic learning approaches. This work is an approach based on the automatic learning through the multi-layer neural network in order to predict the secondary structures of proteins starting from the sequence of amino acids, we have tried to implement and to adapt the neural network for obtaining good percentage of result.

Key Words: Secondary structure prediction, Protein, the automatic learning, amino acids, neural network.