تصنيف ميادين البروتينات البنيوية المبني على نموذج ماركوفي مخفي

طارق أبوشنب

بحث مقدم لنيل درجة الماجستير في العلوم

[الهندسة الكهربائية وهندسة الحاسبات / هندسة الحاسبات]

إشراف

د. رامي الحموز

د. عدنان منيش

كلية الهندسة

جامعة الملك عبد العزيز

جدة-المملكة العربية السعودية

شعبان ١٤٣٨ هـ - مايو ٢٠١٧ م

البروتين مركب عضوي معقّد التركيب ذو وزن جزيئي عالٍ يتكون من أحماض أمينية مرتبطة مع بعضها بواسطة رابطة ببتيدية. الرابطة الببتيدية هي الرابطة التي تكون البروتينات وهي رابطة تساهمية كيميائية تنشيء بين جزئين، عندما تتفاعل مجموعة الكربوكسيل -COOH لأحد الأحماض الأمينية مع مجموعة أمين NH2 لحمض أميني آخر، وينتج عن هذا التفاعل تكون جزيء من الماء H2O ورابطة ببتيدية.في الآونة الاخيرة تم اكتشاف العديد من النطاقات داخل البروتين و نطاق بروتين أو كما يعرف بمجالات البروتين (بالإنجليزية: protein domain) هو جزء مكتمل من سلسلة بروتين معين ذات تركيب مجسم (في ثلاثة أبعاد) قد تنشأ ، وتكون لها وظيفة حيوية ، ويمكنها البقاء منفصلة عن بقية سلسلة البروتين. تختلف المجالات (النطاقات) في طولها بين ٢٥ حمض أميني إلى ٥٠٠ حمض أميني يعد نطاق PDZ واحداً من أكثر مناطق مماثلة البروتينات شيوعاً ويلعب دوراً رئيسياً في العديد من الأمراض. اشتق اسم PDZ من الأحرف الأولى لثلاث بروتينات تم التعرف عليها في هذا النطاق وهي: PSD-95 (بروتين مشارك في كثافة ما بعد المشبكي) و DLG (بروتين ذبابة الفاكهة/الأقراص الكبيرة) و ZO-1 (بروتين مشارك في إصلاح النسيج الطلائي). نقطة الطفرات في السلسلة البدائية للحمض الأميني من نطاق PDZ يمكنها أن تغير من وظيفة النطاق بواسطة التأثير فعلى سبيل المثال مصب الفسفرة وهي عملية محورية في بيولوجيا الخلية. يتم تصنيف نطاقات PDZ الى ثلاثة فئات (فئة١- فئة٢- فئة٣). يعتبر تصنيف هذة النطاقات في المختبر عمل شاقا عمليا ويحتاج الى وقت طويل. لذا كان الهدف من هذه الأطروحة هو ايجاد طريقة لتصنيف هذه النطاقات باستخدام الخوارزميات المناسبة والتي تم استخدامها في بناء نظام تعلم الآلة.

الفصل الأول من هذه الأطروحة عبارة عن مقدمة لموضوع الرسالة وطريقة توزيعها وطرحها وأهدافها. فقد تم في هذا الفصل تقديم فكرة مبسطه عن البروتين وتفاعلات الأحماض الأمينية داخل البروتين واثرها على التكوين الوظيفي في الخلية. كما تم شرح نطاقات البروتين ونطاق PDZ والذي بنيت عليه هذه الدراسة. كما تم في هذا الفصل تقديم فكره مبسطة عن نظم تعلم الآلة واهميتها حديثا في ايجاد حللول للمشاكل الاخرى المشابهة.

الفصل الثاني من هذه الرسالة يقدم فكرة عامه عن الدراسات السابقة في هذا المجال. فقد تم في هذا الفصل مناقشة الخوارزميات المختلفة والتي تم استخدامها في حل مشاكل مشابهة للمشكلة التي تم طرحها في هذه الرسالة. فعلى سبيل المثال تم مناقشة دراسة سابقة للتنبؤ بالأمراض السراطنية داخل الخية Naive Bayes . عن طريق استخدام خورازميات تسمى

في الفصل الثالث من هذه الرسالة تم شرح بشكل مفصل الخوارزمية التي تم استخدامها في هذه الأطروحة لحل مشكلة تصنيف نطاقات PDZ وهي خوارزمية ماركوف المخفي.

الفصل الرابع في هذا الفصل تم شرح نموذج الماركوف المخفي والذي تم بنائه لحل المشكلة المطروحة في هذه الرسالة. بعد بناء النموذج تم في هذا الفصل ايضا شرح النتائج التي توصلنا اليها ودقة النموذج في اعطاء نتائج صحيحة لتصنيف نطاقات PDZ كما تم مقارنة دقة النموذج مع النتائج الاخرى في دراسات سابقة.

الفصل الخامس والأخير تم استعراض الخلاصة التي حصلنا عليها في هذه الدراسة وتوضيح العنصر المستهدف للتحسين وشرح مبسط لتسلسل العمل الذي تم انهائه في هذه الأطروحة.

Classification of Structural Protein Domain Based on Hidden Markov Model

Tarek AbuShanab

A thesis submitted in partial fulfillment of the requirements for degree of Master of Science [Electrical and Computer Engineering / Computer Engineering]

Supervised by
Dr. Rami Al-Hmouz
Dr. Adnan Memic

ACULTY OF ENGINEERING
KING ABDULAZIZ UNIVERSITY
JEDDAH-SAUDI ARABIA
SHABAN 1438H – May 2017G

PDZ is an acronym consolidating the main letters of three proteins — post-synaptic density protein (PSD95), Drosophila disc large tumor suppressor (Dlg1), and Zonula occludens-1 protein (zo-1) [3-5].The PDZ domain consists of a sequence of amino acids usually between 80-90 amino-acids and found in the signaling proteins of microscopic organisms, yeast, plants, viruses, and animals. They have been presented to act as key players ranging from cystic fibrosis to cancer. The classification of PDZ domain in laboratory based on the chemical characteristic is a very difficult and high cost task.

Thus our aim in this study is to find an algorithm to classify PDZ domain as Class I or II based on a given sequence of amino acid. We use the hidden markov model to classify the PDZ domain

sequence as it has been used for solving many problems similar to our problem for example, forecast of protein-coding districts in genome successions, demonstrating groups of related DNA or protein taxonomies.

Hidden markov model was constructed by using interaction dataset. Our dataset consisted of 78 sequences of Class I, and 37 sequences of Class II domains. We split our dataset into training and test sets. In the training phase, we used 90% of the dataset, while the remaining 10% was considered as a testing set.

The HMM model consist of two important matrices, emission matrix and transmission matrix. The emission matrix represents the probability of occurrences of each amino acid in both classes of PDZ domain while the transition matrix represents the probability of transmission between classes. We assumed that our transmission matrix is (0.5, 0.2 and 0.1). Assuming that our initial matrix is 0.5 the HMM has been calculated for both classes based on all transmission matrix that we have assumed. The minimum classification error was (0.35). We tested the decision at each state and reported the final decision based on voting process and found that the classification rate has been improved. The final improvement has been made base on the multiplication of all state probabilities. Here, the decision of class 1 is taken when by multiplication of state probabilities of class 1 is greater than class 2 otherwise the decision will be class 2

We compared the classification results with respect to three different approaches [21-23] with our HMM method. We found that, the HMM method is computationally more effective than the other three classifiers for our problem. We predicted the classes of PDZ domain with accuracies of (83.25%). With this highly enquiring result, this study could be an important step in the automated prediction of PDZ domain classes.