Ministry of Higher Education and
Scientific Research
University of Mosul
College of Computer Science and
Mathematics
Department of Mathematics

Prediction of Secondary Structure of Protein by Combining Hidden Markov Model and Support Vector Machine

**A Thesis Submitted to the Council of the College of
Computer Science and Mathematics
University of Mosul
as a Partial Fulfillment of Requirements
for the Degree of Master of Science
in
Mathematics/Computational**

By

Ahmed H. Qasim

Supervised by

Dr. Omar S. Qasim **Dr. Fatima M. Hasan**

2022 A.D. **1444 A.H.**

Abstract

The study of the 3D structures of protein is a part of bioinformatics, and it is not possible to know the three structures of a protein without knowing the secondary structure. Many researches and studies have been submitted for the purpose of predicting the secondary structure of the protein (PSSP) by studding the distribution of the amino acids in the primary structure, in addition to extracting the secondary structure by laboratory methods, moreover the methods used to predict protein secondary structure are using artificial intelligence algorithms and statistical systems such as SVM and HMMs.

In this thesis, we have presented a proposed algorithm to predict protein secondary structure by combining HMM and SVM, and this algorithm is named **Combining Hidden Support CHS.** A special dataset was used for this algorithm called intersect-Pieces dataset, as well as the well-known RS126 protein datest is used for comparing CHS algorithm results with other research.

CHS is characterized by new features that outperform both SVM and HMM such as great flexibility in dealing with protein sequences, unaffected by defects that may appear in both HMM and SVM, and the highest prediction accuracy was obtained for PSSP reaching 93,8%, and about 95 % With the well-known dataset RS126, CHS has shown a significant superiority over traditional algorithms by accurately predicting the secondary structure of a protein

وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم الرياضيات

التنبؤ بالبنية الثانوية للبروتين من خلال الدمج بين نموذج ماركوف المخفي والة المتجه الداعم

رسالة مقدمة
الى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
كجزء من متطلبات نيل شهادة ماجستير علوم في
الرياضيات/حاسوبية

من قبل

احمد حسين قاسم عبداللطيف

بإشراف

أ.د عمر صابر قاسم

م.د فاطمة محمود حسن

1444 هـ

2022 م

الخلاصة

تضمنت الرسالة دراسة تحليلية لكل من متسلسلات البروتين والاحماض الامينية، ودراسة الجانب النظري لنماذج ماركوف المخفي والة المتجه الداهم وتقديم الحلول المقترحة لبعض المعوقات التي قد تظهر عند استخدام هذه الخوارزميات للتنبؤ بالبنية الثانوية للبروتين.

تم تقديم خوارزمية مقترحة للتنبؤ بالبنية الثانوية للبروتين من خلال الدمج بين نماذج ماركوف المخفي Hidden Markov Model (HMM) والة المتجه الداعم Support Vector Machine (SVM)، وقد سميت هذه الخوارزمية .Combining Hidden Support (CHS) مع استخدام مجموعة بيانات خاصة مكونة من 9078 متسلسلة بروتين في هذه الرسالة لغرض التدريب والاختبار, بالإضافة الى استخدام مجموعة البيانات المعروفة RS126 لغرض مقارنة النتائج مع الدراسات المشابهة.

وقد أظهرت الخوارزمية المقترحة CHS مرونة عالية في التعامل مع متسلسلات البروتين، غير متأثرا بالمشاكل التي تأثر على HMM وSVM، وقد تم الحصول على دقة تنبؤ بالبنية الثانوية للبروتين حوالي 94%, مما يظهر تفوقًا كبيرًا على الخوارزميات التقليدية من خلال التنبؤ بالبنية الثانوية للبروتين.