DEMOCRATIC AND POPULAR REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

MOHAMED BOUDIAF UNIVERSITY - M'SILA
FACULTY OF MATHEMATICS AND INFORMATICS
DEPARTEMENT OF COMPUTER SCIENCE

Research Paper

Presented for obtaining: MASTER Degree
Field: Mathematics and Informatics
Branch: Computer Science
Specialty: Advanced Information System

By: Ahmed Majam

Topic

Prediction of Protein Structure Classes Using Support Vector Machine (SVM) Classifier

Publically defended on: 01 / 06 / 2016 Before a Jury composed of:

Mr. A. KHATTAF	**University of M'sila**	**Chairman**
Mr. YAGOUBI Rached	**University of M'sila**	**Supervisor**
Mr. R. MOKHTARI	**University of M'sila**	**Examiner**

Class : 2015 /2016

Contents

Contents I
GENERAL INTRODUCTION 1

CHAPTER I : BIOINFORMATICS

I.1 Introduction 2
I.2 Molecular Biology 2
I.2.1 Cell 2
I.2.1 DNA 3
I.2.2 RNA 4
I.2.3 Protein 4
I.3 Molecular Genetic 8
I.3.1 The Gene 8
I.3.2 The Genetic Code 8
I.4 Bioinformatics 9
I.4.1 What is Bioinformatics 9
I.4.2 Objective of Bioinformatics 10
I.4.3 Some applications of Bioinformatics 10
I.4.4 Top Bioinformatics Challenges 10
I.5 Primary and Secondary Database 11
I.6 Bioinformatics Data Formats 12
I.6.1 Fasta Format 12
I.6.1 Protein DataBank (PDB) Format 13
I.6.2 GenBank Format 13
I.6.4 Swiss-Prot Format 14
I.7 Conclusion 15

CHAPTER II : Bioinformatics for prediction of structural classes of proteins

II.1 Introduction 16
II.2 Protein Structural Classification 16
II.3 Related works 23
II.3.1 Features Extracted From Amino Acid Sequence of Proteins 23
II.3.2 Features Extracted from PSI-BLAST Profiles of Sequence 24
II.3.3 Features Based on Functional Domains of Sequence 24
II.3.4 Features Extracted from Predicted Protein Secondary Structure Sequence 25
II.3.5 Features from both Amino Acid Sequence and Predicted Secondary Structure Sequence 26
II.3.6 Classification Algorithm 28
II.4 Dataset resources 29
II.5Conclusion 30

CHAPTER III : Support Vector Machine (SVM)

III.1 Introduction 31
III.2 The case when the data is Linearly Separable 31
III.3 The Case When the Data Are Linearly Inseparable 36
III.4 Conclusion 39

CHAPTER IV : Results and discusses

IV.1 Introduction 40
IV.2 The Weka 40
IV.2.1 What's in Weka? 41
IV.2.2 How do you use it? 41
IV.2.3 ARFF format 42
IV.3 Results and discusses 45
IV.3.1 Amino Acids Composition 45

IV.3.2 Dipeptide Composition **47**

IV.3.3 Amino Acids Composition and Dipeptide Composition **50**

IV.4 Conclusion **52**

GENERAL CONCLUSION **53**

Bibliography **54**

GENERAL INTRODUCTION

Living creatures bodies constitute billions of cells, each has a different essential function. Protein is an important component of the cell that itself has many functions.

In this research we will highlight the prediction of protein structures classes for its importance in our life. We highlight a collection of state of the art machine learning algorithm and data processing tools which was named the Weka.

This work is divided into four chapters each one discusses an essential part of the research.

In chapter one, we will talk about molecular biology especially the protein and how it is formed. Also, we will define the term of Bioinformatics, its objectives and the challenges it faces. At the end of this chapter we will talk about biological banks and some formats of them.

In chapter two, we will concentrate on protein structural classification. We will discuss the type of feature extracted. We will present a classification of different methods based on features extracted and their classification algorithmic. Finally, we will present the most used dataset which measure the accuracy.

Chapter three is presenting the SVM (support vector machine) and its importance in our study, and present the different kernel. Then, it highlights the data and its two cases that could be found at.

In chapter four, we will get to know the Weka, what it contains and how we use it. Finally, we will discuss the results which we got through the experiments we are going to make.

Prediction of protein structural classification is widely used in biological laboratories for its huge importance determining protein's functions.

GENERAL CONCLUSION

In this project of master, a soft computing technique like SVM was used to construct classification models for protein structural class prediction. We got to know the SVM, and we used it for the prediction of protein structure classes. Using SVM for prediction seems to be the most popular among all soft computing techniques to solve the protein structural classification problem.

There were old methods of classification of proteins which depended on only features extracted from Amino Acids sequences that has 20 features, and there was another method of classification of protein which relied on features extracted from Dipeptide composition that has 400 features.

Accuracy tests have used 25PDB dataset which is the most popular. We have made some tests on four different kernel kinds of SVM which were linear, polynomial, RBT, and sigmoid. To enhance the results, we have used different cost values in three different tests which were Amino Acids composition, Dipeptide composition, and a gathering of Amino Acids and Dipeptide composition. The results of these tests were similar to each other. We notice that the RBF kernel when (cost=3) scores the highest accuracy's percentages.

In my opinion, the good accuracy's percentages are very important regardless the time it may takes.

Bibliography

[1] Jin Xiong "ESSENTIAL BIOINFORMATICS ",Cambridge university press,2006

[2] What is Molecular Biology?," http://www.news-medical.net/life-sciences/What-is-Molecular-Biology.aspx", [Accessed 30/4/2016] .

[3] D. Robert et B. Vian, "´El´ements de biologie cellulaire 3e ´edition", Doin, 2004

[4] Oliver Brandenberg, Zephaniah Dhlamini, Alessandra Sensi, Kakoli Ghosh, Andrea Sonnino"Introduction to Molecular Biology and Genetic Engineering " Rome, 2011.

[5] Georgia C. Lauritzen "What is protein?" Utah State University, 1992.

[7] Patricio Jeraldo "The Genetic Code" May 5, 2006

[8] C. Burge, "Bioinformaticists Will Be Busy Bees," Genome Technology, No. 17, January, 2002.

[9] Andreas D. Baxevanis, B.F. Francis Ouellette "A Practical Guide to the Analysis of Genes and Proteins, Second Edition", 2001.

[11] Xiaohua Hu, Yi Pan "KNOWLEDGE DISCOVERY IN BIOINFORMATICS, Techniques, Methods, and Applications" John Wiley & Sons, 2007.

[13] Mr. Sundeep Singh Nanuwa, " Investigation into the role of sequence-driven-features and amino acid indices for the prediction of structural classes of proteins ", A thesis, De Montfort University, April 2013.

[14] H. Nakashima, K. Nishikawa, T. Ooi, The folding type of a protein is relevant to the amino acid composition, J. Biochem. 99 (1986) 153–162.

[15] K.C. Chou, A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space, Proteins 21 (1995) 319–344.

[16] W. Kabsch, C. Sander, Dictionary of protein secondary structures: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.

[17] F. Eisenhaber, C. Frommel, P. Argos, Prediction of secondary structural content of proteins from their amino acid composition alone, II the paradox with secondary structural class, Proteins 25 (1996) 169–179.

[18] H.M. Berman, et al., The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.

[19] A. Murzin, S. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of protein database for the investigation of sequence and structures, J. Mol. Biol. 247 (1995) 536–540.

[20] K.C. Chou, C.T. Zhang, Prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.

[21] K.C. Chou, G.M. Maggiora, Domain structural class prediction, Protein Eng. 11 (1998) 523–538.

[22] A. Andreeva, D. Howorth, S. Brenner, T. Hubbard, C. Chothia, A. Murzin, SCOP database in 2004: refinements integrate structure and sequence family data, Nucleic Acid Res. 32 (2004)

D226–D229.
[27] Zerrin Isik, Berrin Yanikoglu, and Ugur Sezerman. "Protein structural class determination using support vector machines." In Computer and Information Sciences-ISCIS 2004, pp. 82-89. Springer Berlin Heidelberg, 2004.
[28] Zhou, Guo-Ping. "An intriguing controversy over protein structural class prediction." Journal of Protein Chemistry 17, no. 8 (1998): 729-738.
[29] Wu, Li, Qi Dai, Bin Han, Lei Zhu, and Lihua Li. "Prediction of protein structural class using a combined representation of protein-sequence information and support vector machine." In Bioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference on, pp. 101-106. IEEE, 2010.
[30] Nair, Achuth Sankar S., and T. Mahalakshmi. "Visualization of genomic data using internucleotidedistance signals." Proceedings of IEEE Genomic Signal Processing (2005): 11-13.
[31] Afreixo, Vera, Carlos AC Bastos, Armando J. Pinho, Sara P. Garcia, and Paulo JSG Ferreira. "Genome analysis with inter-nucleotide distances." Bioinformatics 25, no. 23 (2009): 3064-3070.
[32] Zhang, T-L., and Y-S. Ding. "Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes." Amino Acids 33, no. 4 (2007): 623- 629.
[33] Tang, Fa-ming, Zhong-dong Wang, and Mian-yun Chen. "On multiclass classification methods for support vector machines." Control and Decision 20, no. 7 (2005): 746.
[34] Ding, Yong-Sheng, Tong-Liang Zhang, and Kuo-Chen Chou. "Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network." Protein and peptide letters 14, no. 8 (2007): 811-815.
[35] Chou, Kuo-Chen. "Prediction of protein cellular attributes using pseudo-amino acid composition." Proteins: Structure, Function, and Bioinformatics 43, no. 3 (2001): 246-255.
[36] Abe, Shigeo. "Fuzzy LP-SVMs for multiclass problems." In Proceedings of European Symposium on Artificial Neural Networks (ESANN'2004) Bruges, Belgium, pp. 429 - 434. 2004.
[37] Chou, Kuo-Chen. "A key driving force in determination of protein structural classes." Biochemical and biophysical research communications 264, no. 1 (1999): 216-224.
[38] Klein, Petr, and Charles Delisi. "Prediction of protein structural class from the amino acid sequence." Biopolymers 25, no. 9 (1986): 1659-1672.
[39] Bu, Wei-Shu, Zhi-Ping Feng, Ziding Zhang, and Chun-Ting Zhang. "Prediction of protein (domain) structural classes based on amino-acid index." European Journal of Biochemistry 266, no. 3 (1999): 1043-1049.
[40] Chou, Kuo-Chen. "A key driving force in determination of protein structural classes." Biochemical and biophysical research communications 264, no. 1 (1999): 216-224.

[41] Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic acids research 25, no. 17 (1997): 3389-3402.

[42] Liu, Taigang, Xingbo Geng, Xiaoqi Zheng, Rensuo Li, and Jun Wang. "Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles." Amino acids 42, no. 6 (2012): 2243-2249.

[43] Mizianty, Marcin, and Lukasz Kurgan. "Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences" BMC bioinformatics 10.1 (2009): 414.

[44] Dong, Qiwen, Shuigeng Zhou, and Jihong Guan. "A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation." Bioinformatics 25, no. 20 (2009): 2655-2662.

[45] Guo, Yanzhi, Lezheng Yu, Zhining Wen, and Menglong Li. "Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences." Nucleic acids research 36, no. 9 (2008): 3025-3030.

[46] Guo, Yanzhi, Menglong Li, Minchun Lu, Zhining Wen, and Zhongtian Huang. "Predicting G-protein coupled receptors–G-protein coupling specificity based on autocross-covariance transform." Proteins: structure, function, and bioinformatics 65, no. 1 (2006): 55-60.

[47] Wu, Jiang, Meng-Long Li, Le-Zheng Yu, and Chao Wang. "An ensemble classifier of support vector machines used to predict protein structural classes by fusing auto covariance and pseudo-amino acid composition." The Protein Journal 29, no. 1 (2010): 62-67.

[48] Chou, Kuo-Chen, and Yu-Dong Cai. "Predicting protein structural class by functional domain composition." Biochemical and biophysical research communications 321, no. 4 (2004): 1007-1009.

[49] Ahmadi Adl, Amin, Abbas Nowzari-Dalini, Bin Xue, Vladimir N. Uversky, and Xiaoning Qian. "Accurate prediction of protein structural classes using functional domains and predicted secondary structure sequences." Journal of Biomolecular Structure and Dynamics 29, no. 6 (2012): 1127-1137.

[50] Apweiler, Rolf, Terri K. Attwood, Amos Bairoch, E. Birney, M. Biswas, P. Bucher, L. Cerutti et al. "The InterPro database, an integrated documentation resource for protein families, domains and functional sites." Nucleic acids research 29, no. 1 (2001): 37-40.

[51] Hall, Mark A. "Correlation-based feature selection for machine learning." PhD diss., The University of Waikato, 1999.

[52] Liu, Tian, and Cangzhi Jia. "A high-accuracy protein structural class prediction algorithm using predicted secondary structural information." Journal of theoretical biology 267, no.3 (2010): 272-275.

[53] Zhang, Shengli, Shuyan Ding, and Tianming Wang. "High-accuracy prediction of protein

structural class for low-similarity sequences based on predicted secondary structure." Biochimie 93, no. 4 (2011): 710-714.

[54] Ding, Shuyan, Shengli Zhang, Yang Li, and Tianming Wang. "A novel protein structural classes prediction method based on predicted secondary structure." Biochimie 94, no. 5 (2012): 1166-1171.

[55] Jones, David T. "Protein secondary structure prediction based on position-specific scoring matrices." Journal of molecular biology 292, no. 2 (1999): 195-202.

[56] Lin, Kuang, Victor A. Simossis, Willam R. Taylor, and Jaap Heringa. "A simple and fast secondary structure prediction method using hidden neural networks." Bioinformatics 21, no. 2 (2005): 152-159.

[57] Kurgan, Lukasz A., Tuo Zhang, Hua Zhang, Shiyi Shen, and Jishou Ruan. "Secondary structure-based assignment of the protein structural classes." Amino Acids 35, no. 3 (2008): 551-564.

[58] Kurgan, Lukasz, Krzysztof Cios, and Ke Chen. "SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences." BMC bioinformatics 9, no. 1 (2008): 226.

[59] Mohammad, Tabrez Anwar Shamim, and Hampapathalu Adimurthy Nagarajaram. "Svmbased method for protein structural class prediction using secondary structural content and structural information of amino acids." Journal of Bioinformatics and Computational biology 9, no. 4 (2011): 489-502.

[60] Cheng, Jianlin, Arlo Z. Randall, Michael J. Sweredoski, and Pierre Baldi. "SCRATCH: a protein structure and structural feature prediction server." Nucleic acids research 33, no. suppl 2 (2005): W72-W76.

[61] Kurgan, Lukasz, and Ke Chen. "Prediction of protein structural class for the twilight zone sequences." Biochemical and biophysical research communications 357, no. 2 (2007): 453-460.

[62] Yang, Jian-Yi, Zhen-Ling Peng, and Xin Chen. "Prediction of protein structural classes for low-homology sequences based on predicted secondary structure." BMC bioinformatics 11, no. Suppl 1 (2010): S9.

[63] Hastie, Trevor, and Robert Tibshirani. "Classification by pairwise coupling." The annals of statistics 26, no. 2 (1998): 451-471.

[64] Cai, Yu-Dong, and Guo-Ping Zhou. "Prediction of protein structural classes by neural network." Biochimie 82, no. 8 (2000): 783-785.

[65] Chandonia, John-Marc, and Martin Karplus. "Neural networks for secondary structure and structural class predictions." Protein Science 4, no. 2 (1995): 275-285.

[66] Syeda Nadia Firdaus, " PROTEIN STRUCTURAL CLASS PREDICTION USING PREDICTED SECONDARY STRUCTURE AND HYDROPATHY PROFILE", A thesis, (Master), Toronto- Canada, 2013.

[67] Lukasz A. Kurgan∗, Leila Homaeian, "Prediction of structural classes for protein sequences and domains—Impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy", A thesis, Pattern Recognition 39 (2006) 2323 – 2343

[68] Jiawei Han , Micheline Kamber , "Data Mining Concepts and Techniques " , second edition , Elsevier Inc., 2006.

[69] Eibe Frank, Ian H. Witten '' Data Mining Practical Machine Learning Tools and Techniques'', Elsevier Inc., 2005.

[6] Protein structure "http://www.particlesciences.com/news/technical-briefs/2009/protein-structure.html " [Accessed 18/5/2016].

[10] http://www.biostat.jhsph.edu/~iruczins/teaching/260.655/links/pdbformat [Accessed 5/2/2016]

[12] Bioinformatics: Sequence File Formats, http://www.algosome.com/articles/bioinformatics-sequence-file-formats.html [Accessed 18/5/2016]

[23]"Information on 2hbc", http://www.rcsb.org/pdb/explore/explore.do?structureId=2hbc [Accessed 21/5/2016].

[24]"Information on 1ku8", http://www.rcsb.org/pdb/explore/explore.do?structureId=1ku8 [Accessed 21/5/2016].

[25]"Information on 2bnh", http://www.rcsb.org/pdb/explore/explore.do?structureId=2BNH [Accessed 21/5/2016].

[26]"Information on 1pya" http://www.rcsb.org/pdb/explore/explore.do?structureId=1pya [Accessed 21/5/2016].

ملخص

تتحكم البروتينات في جميع الوظائف البيولوجية للكائنات الحية. تتألف بنية البروتين من أربع فئات رئيسية، وكل فئة تؤدي وظيفة مختلفة وفقاً لطبيعتها. ونظراً للاستكشاف الكبير من سلاسل البروتين في بنوك المعلومات، يعتبر تحديد فئات بنية البروتين شيئاً صعباً من خلال الطرق التقليدية فيما يتعلق بالتكلفة والوقت. وبالنظر إلى أهمية فئات بنية البروتين، من المستحسن تطوير نموذج حاسوبي لمعرفة وتمييز تلك الفئات بدقة عالية. سنستخدم ثلاثة أنظمة استخراج للخصائص من الأحماض الأمينية لمعرفة المعلومات من تسلسل البروتين. يتم تقييم آداء النموذج المقترح باستخدام قاعدة بيانات البروتين. نسبة نجاح النموذج المقترح هو 54.085%.

كلمات مفتاحية: المعلومات الحيوية، تصنيف فئات بنية البروتين،SVM

Abstract

Proteins control all biological functions in living species. Protein structure is comprised of four major classes including all-α class, all-β class, α/β and α+β. Each class performs a different function according to their nature. Owing to the large exploration of protein sequences in the databanks, the identification of protein structure classes is difficult through conventional methods with respect to cost and time. Looking at the importance of protein structure classes, it is thus highly desirable to develop a computational model for discriminating protein structure classes with high accuracy. For this purpose, we propose a Support Vector Machine. Three features extraction schemes named Amino Acid Composition, Dipeptide Composition and a combination of them both are used to explore valuable information from protein sequences. The performance of the proposed model is assessed using the common dataset 25PDB. The success percentage of the proposed model is 54.085 %.

Key words: Bioinformatics, Support Vector Machine (SVM), Structural classes of protein.

Résumé

Les protéines contrôlent toutes les fonctions biologiques des espèces vivantes. La structure des protéines est composée de quatre grandes classes incluant all-α classe, all-β classe, α / β classe et α + β classe. Chaque classe a une fonction différente en fonction de leur nature. En raison de la grande exploration des séquences de protéines dans les banques de données, l'identification des classes de structure de protéines est difficile par les méthodes conventionnelles en ce qui concerne le coût et le temps. En regardant l'importance des classes de structures de protéines, il est donc hautement souhaitable de développer un modèle de calcul pour distinguer les classes de structure de protéines avec une grande précision. A cet effet, nous utilisons (SVM) Support Vector Machine. Trois systèmes d'extraction des caractéristiques du nom de composition d'acides aminés, de dipeptide Composition et une combinaison de tous les deux sont utilisées pour explorer des informations précieuses à partir de séquences de protéines. La performance du modèle proposé est évaluée en utilisant le 25PDB ensemble de données commun. Le pourcentage de réussite du modèle proposé est 54,085%.

Mots clé: Bioinformatiques, Support Vector Machine (SVM), les classes des structures de protéine.