Ain Shams University
Faculty of Computer and Information Sciences
Information Systems Department

Mining Structural Patterns for Automatic Protein Function Prediction

A thesis submitted in partial fulfillment of the requirements for
the degree of PhD in Computer and Information Sciences
To
Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University

By

Huda Amin Maghawry Amin

M.Sc. in Computer and Information Sciences,
Assistant Lecturer at Information Systems Departments
Faculty of Computer and Information Sciences, Ain Shams University

Under Supervision of

Prof. Dr. Mostafa Gadal-Haqq M. Mostafa

Computer Science Department
Faculty of Computer and Information Sciences, Ain Shams University

Prof. Dr. Mohamed Hashem Abdel Aziz

Information Systems Department
Faculty of Computer and Information Sciences, Ain Shams University

Prof. Dr. Tarek Fouad Gharib

Information Systems Department
Faculty of Computer and Information Sciences, Ain Shams University

Cairo - 2014

Acknowledgement

First and foremost, thanks to Allah, the most merciful, the most graceful, for His great help throughout this work.

I would like to express my deep and sincere gratitude to Prof. Dr. Mostafa Gadal-Haqq. His leadership, patience, understanding, encouragement, immense knowledge and personal guidance have provided a significant basis for this work. His detailed review and constructive comments have been of great value for me. I am gratefully thankful.

Also, I would like to express my deep and sincere gratitude to Prof. Dr. Tarek Gharib. I deeply appreciate his continuous supportive supervision. His advices, enthusiasm, patience and understanding were very motivational for me. I am gratefully thankful.

I wish also to express my warm and sincere thanks to Prof. Dr. Mohamed Hashem for his help and support.

My sincere and special gratitude from my deep heart is due to my great and beloved parents who supported me throughout my life with all their love and faith. They were always around at times I thought that it is impossible to continue, they helped me to keep things in perspective. Without their encouragement and understanding it would have been impossible for me to finish this work.

Last but not the least; I would like to thank my dear and beloved husband. His continuous and unconditional love, support, patience, sacrifice and understanding will remain my inspiration throughout my life. Also, my little and lovely son with his sweet smile, I owe him lots and lots of fun hours. Words would never say how grateful I am to both of them.

Abstract

Protein function prediction is a challenging problem in bioinformatics. It has a great impact on the areas of diseases treatment and drug industry. Structure-based proteins representation plays an important role in proteins function prediction process. Three aspects of protein functions prediction have been considered: Predicting enzymes family and superfamily, classifying enzymes versus non-enzymes proteins and discriminating DNA-binding and non DNA-binding proteins. The thesis presents a modification to an existing protein representation approach which utilizes distance patterns between protein residues and a maximum cutoff. Also, the thesis presents a new protein structure representation for efficient protein function prediction. The new representation is based on three-dimensional patterns of protein residues. It utilizes atoms coordinates of protein residues, including the angles and distance patterns. This thesis also presents a study of the proposed protein representation and other protein-derived sequence, psychochemical and structure features to enhance the prediction of DNA-binding proteins and their classes.

Summary

One of the challenging problems in bioinformatics is the prediction of protein function. Protein function is the main key that can be used to classify different proteins. The analysis of proteins and their functions is an important research area. Such analysis affects many applications like clarification of the living body mechanism, treatment of diseases and drug industry. Protein function can be inferred experimentally with very small throughput and high cost or computationally with very high throughput and lower cost. Many of protein sequences and structures are available but with no knowledge about their function. Therefore, methods for protein function prediction are highly and continuously required. Computational methods are based on protein sequences or structures. Protein functions are highly related to their structures. Therefore, structure-based proteins representation plays an important role in the prediction process.

This thesis presents a modification to an existing protein representation approach which utilizes distance patterns between protein residues and a maximum cutoff. The proposed modified representation considers the whole protein instead of using cutoff. Comparative analysis was done to evaluate the proposed representation method and the existing method. The aspect of protein function considered is based on enzyme activity. The results show that the proposed representation outperforms the existing representation with a prediction accuracy of 90.12% and 80.27% for superfamily and family level, respectively, with accuracy improvement of about 5% in average.

This thesis also presents a new structure-based protein representation for efficient protein function prediction. The new representation is based on three-dimensional patterns of protein residues. It utilizes atoms coordinates of protein residues, including the angles and distance patterns. The proposed representation uses protein structure only with no need to any sequence information. Besides, it does not need any prior alignment process. The aspects of protein functions considered using different datasets: Predicting enzymes family and superfamily and classifying enzymes versus non-enzymes proteins. The prediction accuracy of the proposed representation using various classification methods outperforms a recently introduced representation that is based only on the distance patterns. The results show that the proposed representation achieved prediction accuracy of 98.3% in predicting superfamily, 91% in predicting family and 79.25% in predicting enzyme proteins, with improvement of about 10% on average.

Finally, the thesis presents a study of different protein-derived sequence, psychochemical and structure features. The objective was to enhance the prediction of DNA-binding proteins and classes. This is achieved through finding efficient protein representations that able to predict whether a protein is DNA-binding protein and analyzing how well protein-derived representations predict each of DNA-binding protein classes. The protein features achieved accuracy improvement of about 7% on average for the prediction of DNA-binding proteins. The proposed representation when combined with other features achieved improvement in accuracy about 7% and 12% on average for the prediction of DNA-binding proteins and DNA-binding protein classes, respectively.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

2SAAC	Two-segment Amino Acid Composition
AAC	Amino Acid Composition
AACD	Amino Acid Composition Distribution
ANN	Artificial Neural Network
APGM	Approximate Graph Mining
BLAST	Basic Local Alignment Search Tool
CAFA	Critical Assessment of protein Function Annotation
CDF	Cumulative Distribution Function
CE	Combinatorial Extension
CFS	Correlation-based Featured Subset
CSM	Cutoff Scanning Matrix
DALI	Distance-matrix ALIgnment
DC	Dipeptide Composition
DNA	Deoxyribonucleic Acid
DWT	Discrete Wavelet Transform
EC	Enzyme Commission
ETA	Evolutionary Trace Annotation
FAIR	Finding All Internal Repeats
FASTA	Fast-All
FN	False Negative
FP	False Positive
GAAP	Gapped Amino Acid Pair Composition
GO	Gene Ontology
GRAVY	Grand Average of Hydropathicity Index
KNN	K-Nearest Neighbor
MCC	Mathew Correlation Coefficient
MCSM	Modified Cutoff Scanning Matrix
MD	Moment Descriptor

MIPS	Munich Information Center for Protein Sequences
mRNA	messenger Ribonucleic Acid
MSE	Multi-Scale Energy
NMR	Nuclear Magnetic Resonance
OIT	Optimal and Information Theoretic
PCA	Principle Components Analysis
PcAA	Pair-coupled Amino Acid Composition
PDB	Protein Data Bank
PDF	Probability Distribution Function
PPI	Protein-protein Interactions
PSAS	Pairwise Sequence Alignments Scores
PseAAC	Pseudo Amino Acid Composition
PSI-BLAST	Position-Specific Iterative Basic Local Alignment Search Tool
PSM	Protein Structure Matrix
PSM-C	Protein Structure Matrix using Cutoff
RBF	Radial Basis Function
ROC	Receiver Operating Characteristics
SALSA	Structurally Aligned Local Sites of Activity
SCOP	Structural Classification of Proteins
SSAP	Sequence Structure Alignment Program
SSMBS	Sequentially Separated Motifs in Biological Sequences
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TN	True Negative
TOPS	Topology-based Protein Structure
TP	True Positive