Journal of Quality Measurement and Analysis
Jurnal Pengukuran Kualiti dan Analisis

JQMA 8(2) 2012, 9-20

FIS-PNN: A HYBRID COMPUTATIONAL METHOD FOR PROTEIN-PROTEIN INTERACTIONS PREDICTION USING THE SECONDARY STRUCTURE INFORMATION

(FIS-PNN: Suatu Kaedah Pengkomputeran Hibrid bagi Ramalan Interaksi Protein-Protein Menggunakan Maklumat Struktur Sekunder)

SAKHINAH ABU BAKAR[1], JAVID TAHERI[2] & ALBERT Y ZOMAYA[2]

ABSTRACT

The study of protein-protein interactions (PPI) is an active area of research in biology because it mediates most of the biological functions in any organism. This work is inspired by the fact that proteins with similar secondary structures mostly share very similar three-dimensional structures, and consequently, very similar functions. As a result, they must interact with each other. In this study we used our approach, namely FIS-PNN, to predict the interacting proteins in yeast from the information of their secondary structures using hybrid machine learning algorithms. Two main stages of our approach are similarity score computation, and classification. The first stage is further divided into three steps: (1) Multiple-sequence alignment, (2) Secondary structure prediction, and (3) Similarity measurement. In the classification stage, several independent first order Sugeno Fuzzy Inference Systems and probabilistic neural networks are generated to model the behavior of similarity scores of all possible proteins pairs. The final results show that the multiple classifiers have significantly improved the performance of the single classifier. Our method, namely FIS-PNN, successfully predicts PPI with 96% of accuracy, a level that is significantly greater than all other sequence-based prediction methods.

Keywords: protein-protein interaction prediction; hybrid method; secondary structure; machine learning algorithm

ABSTRAK

Interaksi protein-protein merupakan suatu bidang kajian biologi yang aktif kerana ia menjadi perantara bagi hampir semua fungsi biologi di dalam organisma. Kajian ini diilhamkan daripada hakikat bahawa protein yang mempunyai struktur sekunder yang serupa akan mempunyai struktur tiga-dimensi yang hampir serupa, dan seterusnya mempunyai fungsi yang sangat serupa. Oleh yang demikian, protein-protein tersebut akan berinteraksi di antara satu sama lain. Dalam kajian ini, digunakan pendekatan FIS-PNN untuk meramal interaksi di antara protein dalam ragi menggunakan maklumat struktur sekunder dan al-Khwarizmi hibrid pembelajaran mesin. Dua peringkat utama pendekatan ini adalah pengiraan skor keserupaan dan pengelasan. Peringkat pertama pula mempunyai tiga langkah: (1) Penjajaran multi-jujukan, (2) Peramalan struktur sekunder, dan (3) Pengukuran keserupaan. Dalam peringkat pengelasan pula, beberapa sistem pentaadbiran kabur Sugeno peringkat pertama yang tak bersandar dan rangkaian neural berkeberangkalian dijana untuk memodelkan telatah skor keserupaan bagi semua pasangan protein yang mungkin. Hasil kajian mendapati pengelas berbilang telah meningkatkan ketepatan ramalan berbanding dengan pengelas tunggal. FIS-PNN yang dicadangkan telah berjaya meramal interaksi protein-protein dengan ketepatan 96%, suatu tahap ketepatan yang jauh lebih baik berbanding dengan kesemua kaedah ramalan berasaskan jujukan yang lain.

Kata kunci: ramalan interaksi protein-protein; kaedah hibrid; struktur sekunder; al-Khwarizmi pembelajaran mesin

References

Bakar S. A., Taheri J., & Zomaya A. Y. 2009. Fuzzy Systems Modeling for Protein-protein Interaction Prediction in Saccharomyces Cerevisie. *18th World IMACS / MODSIM Congress: Australia*, pp. 782-788.

Bock J. R. & Gough D. A. 2000. Predicting Protein-protein Interactions from Primary Structure. *Bioinformatics* 12(5): 455-460.

Craig R. A. & Liao L. 2007. Phylogenetic Tree Information Aids Supervised Learning for Predicting Protein-protein Interaction based on Distance Matrices. *BMC Bioinformatics* **8**(6): 1-12.

Enright A. J., Iliopoulos I., Kyrpides N. C. & Ouzounis C. A. 1999. Protein Interaction Maps for Complete Genomes based on Gene Fusion Events. *Letters of Nature* 402: 86-90.

Espadaler J., Romero-Isart O., Jackson R. M. & Oliva B. 2005. Prediction of Protein-protein Interactions using Distant Conservation of Sequence Patterns and Structure Relationships. *Bioinformatics* **21**(16): 3360-3368.

Lee M. S., Park S. S. & Kim M. K. 2005. A Protein Interaction Verification System Based on a Neural Network Algorithm. *IEEE Computational Systems Bioinformatics Conference Workshops.* IEEE Computer Society; 151-154.

Martin S., Roe D. & Faulon J. 2005. Predicting Protein-protein Interactions using Signature Products. *Bioinformatics* **21**(2): 218-226.

Pazos F. & Valencia A. 2001. Similarity of Phylogenetic Trees as Indicator of Protein-protein Interaction. *Protein Engineering* **14**(9): 609-614.

Ruepp A., Zollner A., Maier D., Albermann K., Hani J., Mokrejs M., Tetko I., Guldener U., Mannhaupt G. & Munsterkotter M. 2004. The FunCat: A Funtional Annotation Scheme for Systematic Classification of Proteins from Whole Genomes. *Nucleic Acids Research* **32**(18): 5539-5545.

Sato T., Yamanishi Y. & Horimoto K. 2003. Prediction of Protein-protein Interactions from Phylogenetic Trees using Partial Correlation Coefficient. *Genome Imformatics* **14**: 496-497.

Specht D. F. 1990. Probabilistic Neural networks and the Polynomial Adaline as Complementary Techniques for Classification. *IEEE Transactions on Neural Networks* **1**(1): 111-121.

Taheri J. & Zomaya A. Y. 2006. Fuzzy Logic. In *Handbook of Nature-Inspired and Innovative Computing.* Edited by Zomaya A. Y.: New York: Springer Science + Business Media Inc.

Taheri J. & Zomaya A. Y. 2010. RBT-L: A Location Based Approach for Solving the Multiple Sequence Alignment Problem. *International Journal of Bioinformatics Research and Applications (IJBRA)* **6**: 37-57.

Tamames J., Casari G., Ouzounis C. & Valencia A. 1997. Conserved Clusters of Functionally Related Genes in Two Bacterial Genomes. *Journal of Molecular Evolution* **44**: 66-73.

Thomas G. D. 2000. Ensemble Methods in Machine Learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems.* Springer-Verlag.

Tramontano A. 2005. The Ten Most Wanted Solutions in Protein Bioinformatics: Chapman & Hall / CRC.

Xenarios I., Rice D. W., Salwinski L., Baron M. K., Marcotte E. M. & Eisberg D. 2000. DIP: The Database of Interacting Proteins. *Nucleic Acids Research* **28**(1): 289-291.

[1]Pusat Pengajian Sains Matematik
Fakulti Sains dan Teknologi
Universiti Kebangsaan Malaysia
43600 UKM Bangi
Selangor DE, MALAYSIA
*E-mail: sakhinah@ukm.my**

[2]School of Information Technologies
Faculty of Engineering and IT
The University of Sydney
NSW 2006, AUSTRALIA.
E-mail: javid.taheri@sydney.edu.au, albert.zomaya@sydney.edu.au

** Corresponding author*