Journal of Quality Measurement and Analysis
Jurnal Pengukuran Kualiti dan Analisis

JQMA 9(2) 2013, 11-26

CHARACTERISATION OF ESSENTIAL PROTEINS IN PROTEINS INTERACTION NETWORKS

(Pencirian Protein-protein Utama dalam Rangkaian Interaksi Protein)

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

ABSTRACT

The identification of essential proteins is theoretically and practically important as it is essential to understand the minimal surviving requirements for cellular lives, and it is fundamental of drug development. As conducting experimental studies to identify essential proteins are both time and resource consuming, here we present a computational approach in predicting them based on network topology properties from protein-protein interaction networks of *Saccharomyces cerevisiae*, *Escherichia coli* and *Drosophila melanogaster*. The proposed method, namely EP^{3}NN (Essential Proteins Prediction using Probabilistic Neural Network), employed a machine learning algorithm called Probabilistic Neural Network as a classifier to identify essential proteins of the organism of interest. EP^{3}NN uses degree centrality, closeness centrality, local assortativity and local clustering coefficient of each protein in the network for such predictions. Results show that EP^{3}NN managed to successfully predict essential proteins with an average accuracy of 95% for our studied organisms. Results also show that most of the essential proteins are close to other proteins, have assortativity behaviour and form clusters/ sub-graph in the network.

Keywords: essential protein; machine learning algorithms; neural networks; protein interaction networks

ABSTRAK

Pengenalpastian protein utama adalah penting secara teori dan praktikal kerana ia penting untuk memahami keperluan minimum bagi kehidupan sel, dan ia menyediakan asas untuk pembangunan dadah/ubat. Memandangkan kajian untuk mengenal pasti protein utama yang dijalankan secara eksperimen memerlukan tempoh atau masa yang lama serta sumber yang terhad, di sini dibentangkan pendekatan pengkomputeran dalam meramal protein-protein utama berdasarkan sifat-sifat topologi rangkaian daripada rangkaian interaksi protein bagi *Saccharomyces cerevisiae*, *Escherichia coli* dan *Drosophila melanogaster*. Kaedah yang dicadangkan, iaitu EP^{3}NN (Ramalan Protein Utama menggunakan Rangkaian Neural Berkebarangkalian), menggunakan al-Khwarizmi pembelajaran mesin, iaitu rangkaian neural berkebarangkalian sebagai pengelas untuk mengenal pasti protein utama organisma-organisma tersebut. EP^{3}NN menggunakan kepusatan darjah, kepusatan keakraban, assortativiti tempatan dan pekali kelompok tempatan bagi setiap protein dalam rangkaian tersebut. Hasil kajian menunjukkan bahawa EP^{3}NN berjaya meramal protein utama dengan purata ketepatan 95 % bagi organisma-organisma yang dikaji. Keputusan juga menunjukkan bahawa sebahagian besar protein utama berkedudukan dekat dengan protein lain, mempunyai tingkah laku assortativiti dan membentuk kelompok/subgraf dalam rangkaian.

Kata kunci: protein utama; al-Khwarizmi pembelajaran mesin; rangkaian neural; rangkaian interaksi protein

References

Adams M. D., Celniker S. E., Holt R. A., *et al.* 2000. The genome sequence of drosophila melanogaster. *Science* **287**: 2185-2195.

Acencio M. L. & Lemke N. 2009. Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. *BMC Bioinformatics* **10**: 1-18.

Bhan A., Galas D. J. & Dewey T. G. 2002. A duplication growth model of gene expression networks. *Bioinformatics* **18**: 1486-1493.

Celniker S. E. & Rubin G. M. 2003. The Drosophila melanogaster genome. *Annu. Rev. Genomics Hum. Genet.* **4**: 89–117.

Cole S. T. 2002. Comparative mycobacterial genomics as a tool for drug target and antigen discovery. *European Respiratory Journal* **20**: 78-86

Colizza V., Flammini A., Maritan A. & Vespignanim A. 2005. Characterization and modeling of protein-protein interaction networks. *Physica A* **352**: 1-27.

Coulomb S., Bauer M., Bernard D. & Marsolier-Kergoat M. C. 2005. Gene essentiality and the topology of protein interaction networks. *Proc. R. Soc. B* **272:** 1721-1725.

Cullen L. M. & Arndt G. M. 2005. Genome-wide screening for gene function using rnai in mammalian cells. *Immunology and Cell Biology* **83**: 217-223.

Estrada E. 2006. Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomic* **6**: 35-40.

Gerdes S. Y., Scholle M. D., Campbell J. W., *et al.* 2003. Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. *Journal of Bacteriology* **185**: 5673–5684.

Giaever G., Chu A. M., Ni L., *et al.* 2002. Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **419**: 387-391.

Gofieau A., Barrell B. G., Bussey H., Davis, R. W., Dujon, B. *et al.* 1996. Life with 6000 genes. *Science* **274**(5287): 546, 563-567.

Gustafson A. M., Snitkin E. S., Parker S. C., DeLisi C. & Kasif S. 2006. Towards the identification of essential genes using targeted genome sequencing and comparative analysis. *BMC Genomics* **7**: 1-16.

Hakes L., Pinney J. W., Robertson D. L. & Lovell S. C. 2008. Protein-protein interaction networks and biology - what's the connection? *Nature Biotechnology* **26**: 69-72.

Hudault S., Guignot J. & Servin A. L. 2001. Escherichia coli strains colonising the gastrointestinal tract protect germfree mice against Salmonella typhimurium infection. *Gut* **49**: 47–55.

Junker B. H. & Schreiber F. 2008. Analysis of biological networks. Hoboken, NJ: John Wiley & Sons, Inc.

Lee P. H., Tsai J. J. P., Fand J. F. & Ng K. L. 2005. Study of protein-protein interaction networks via random graph approach. Fourth IEEE Conference of Cognitive Informatics pp. 110-119.

Masters T. 1995. Advanced algorithms for neural networks: A C++ sourcebook. New York: John Wiley & Sons, Inc.

Newman M. E. J. 2002. Assortative mixing in networks. *Physical Review Letters* **89**: 208701-1-208701-4.

Ng K. L. & Huang C.-H. 2004. A cross-species study of the protein-protein interaction networks via the random graph approach. Fourth IEEE Symposium on Bioinformatics and Bioengineering, pp. 561-567.

Piraveenan M., Prokopenko M. & Zomaya A. Y. 2009. Local assortativity and growth of internet. *The European Physical Journal B* **70**: 275-285.

Piraveenan M., Prokopenko M. & Zomaya A. Y. 2012. Assortative mixing in directed biological networks. *IEEE Transactions on Computational Biology and Bioinformatics* **9**: 66–78.

Roemer T., Jiang B., Davison J., Ketela, T., Veillette, K. *et al*. 2003. Large-scale essential gene identification in candida albicans and applications to antifungal drug discovery. *Molecular Microbiology* **50**: 167–181.

Seringhaus M., Paccanaro A., Borneman A., Snyder M. & Gerstein M. 2006. Predicting essential genes in fungal genomes. *Genome Research* **12**: 1126-1135.

Silva J. P. M. D., Acencio M. L., Mombach J. C. M., *et al.* 2008. In silico network topology-based prediction of gene essentiality. *Physica A* **387**: 1049-1055.

Specht D. F. 1988. Probability neural networks for classification, mapping or associative memory. Proc. IEEE International Conference on Neural Networks, pp. 525 - 532.

Specht D. F. 1990. Probabilistic neural networks and the polynomial adaline as complementary techniques for classification. *IEEE Transactions on Neural Networks* **1**: 111-121.

Taheri J. & Zomaya A. Y. 2006a. An overview of neural network models. *Handbook of Bioinspired Algorithms and Applications.* Boca Raton, Florida: Chapman & Hall/CRC Press.

Taheri J. & Zomaya A. Y. 2006b. Artificial neural networks. *Handbook of Nature-Inspired and Innovative Computing*. New York: Springer Science + Business Media Inc.

Vogt R. L. & Dippold L. 2005. Escherichia coli O157:H7 outbreak associated with consumption of ground beef, June–July 2002. *Public Health Reports* **120**: 174-178.

Wagner A. 2003. How the global structure of protein interaction networks evolves. *Proc. R. Soc. Lond. B Biol. Sci.* **270**: 457–466.
Xenarios I., Rice D. W., Salwinski L., *et al.* 2000. DIP: the database of interacting proteins. *Nucleic Acids Research* **28**: 289-291.
Zhang R., Ou H.-Y. & Zhang C.T. 2004. DEG: A database of essential genes. *Nucleic Acids Research* **32**: D271-D272.

[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