中央研究院統計科學研究所
學　術　演　講

講　題：Predicting protein-protein interactions in unbalanced data using the primary structure of proteins

演講人：Prof. Tien-Hao Chang (張天豪教授)
（國立成功大學電機工程學系）

時　間：2011年4月15日（星期五）上午10:30-12:00

地　點：中央研究院統計科學研究所二樓交誼廳

※茶　會：上午10：10統計所二樓交誼廳

Abstract

Background

Elucidating protein-protein interactions (PPIs) is essential to constructing protein interaction networks and facilitating our understanding of the general principles of biological systems. Previous studies have revealed that interacting protein pairs can be predicted by their primary structure. Most of these approaches have achieved satisfactory performance on datasets comprising equal number of interacting and non-interacting protein pairs. However, this ratio is highly unbalanced in nature, and these techniques have not been comprehensively evaluated with respect to the effect of the large number of non-interacting pairs in realistic datasets. Moreover, since highly unbalanced distributions usually lead to large datasets, more efficient predictors are desired when handling such challenging tasks.

Results

This study presents a method for PPI prediction based only on sequence information, which contributes in three aspects. First, we propose a probability-based mechanism for transforming protein sequences into feature vectors. Second, the proposed predictor is designed with an efficient classification algorithm, where the efficiency is essential for handling highly unbalanced datasets. Third, the proposed PPI predictor is

assessed with several unbalanced datasets with different positive-to-negative ratios (from 1:1 to 1:15). This analysis provides solid evidence that the degree of dataset imbalance is important to PPI predictors.

Conclusions

Dealing with data imbalance is a key issue in PPI prediction since there are far fewer interacting protein pairs than non-interacting ones. This article provides a comprehensive study on this issue and develops a practical tool that achieves both good prediction performance and efficiency using only protein sequence information.

中 央 研 究 院
統計科學研究所

Tien-Hao Chang (Darby Chang)
張天豪

National Cheng Kung University
No.1, University Road,
Tainan 70101, Taiwan

+886-6-2757575 ext. 62421
darby@ee.ncku.edu.tw
http://mbi.ee.ncku.edu.tw/

EDUCATION

National Taiwan University, Taipei, Taiwan
- Ph.D. in Computer Science and Information Engineering *June 2006*
  Dissertation: "A Study on Prediction of Protein Binding Sites"
- M.S. in Computer Science and Information Engineering *June 2004*
  Thesis: "A Study on Expediting Analysis of Protein Substructures"
- B.S. in Computer Science and Information Engineering *June 2002*

PROFESSIONAL EXPERIENCE

Associated Professor *August 2010 – present*
Department of Electrical Engineering
National Cheng Kung University

Assistant Professor *August 2006 – July 2010*
Department of Electrical Engineering
National Cheng Kung University

RESEARCH INTERESTS

- Data Mining and Machine Learning: Improving density estimation and function approximation techniques; developing advanced learning and optimization algorithms
- Bioinformatics and Systems Biology: Applying data mining and machine learning techniques on solving Bioinformatics and Systems Biology problems, which include sequence, structure and energy analyses

JOURNAL PAPERS

Yi-Zhong Weng, **Darby Tien-Hao Chang***, Yu-Feng Huang and Chih-Wei Lin, "A study on the flexibility of enzyme active sites," *BMC Bioinformatics*, in press, 2010. (SCI)

Darby Tien-Hao Chang, Cheng-Yi Huang, Chi-Yeh Wu and Wei-Sheng Wu*, "YPA: an integrated repository of promoter features in Saccharomyces cerevisiae," *Nucleic Acids Research*, **vol. 39**, Database Issue, D647-D652, 2011. (SCI) **Featured Article**, representing the **top 5%** of papers in terms of originality, significance and scientific excellence.

Darby Tien-Hao Chang, Jung-Hsin Lin, Chih-Hung Hsieh and Yen-Jen Oyang*, "On the design of optimization algorithms for prediction of molecular interactions," *International Journal on Artificial Intelligence Tools*, **19**, 3, 267-280, 2010. (SCI)

Chi-Yuan Yu, Lih-Ching Chou and **Darby Tien-Hao Chang***, "Predicting protein-protein interactions in unbalanced data using the primary structure of proteins," *BMC Bioinformatics*, **11**:167, 2010. (SCI) **Highly Accessed Article**, identifying those articles that have been especially highly accessed, relative to their age, and the journal in which they were published.

Chih-Hung Hsieh, **Darby Tien-Hao Chang***, Cheng-Hao, Hsueh, Chi-Yeh Wu and Yen-Jen Oyang, "Predicting microRNA precursors with a generalized Gaussian components based density estimation algorithm," *BMC Bioinformatics*, **11**(Suppl 1):S52, 2010. (SCI)

Darby Tien-Hao Chang*, Yu-Tang Syu and Po-Chang Lin, "Predicting the protein-protein interactions using primary structures with predicted protein surface," *BMC Bioinformatics*, **11**(Suppl 1):S3, 2010. (SCI)

Darby Tien-Hao Chang, Ting-Ying Chien and Chien-Yu Chen*, "seeMotif: exploring and visualizing sequence motifs in 3D structures," *Nucleic Acids Research*, **vol. 37**, Web Server Issue, W552-W558, 2009. (SCI)

Darby Tien-Hao Chang*, Hsuan-Yu Huang, Yu-Tang Syu and Chih-Peng Wu, "Real value prediction of protein solvent accessibility using enhanced PSSM features," *BMC Bioinformatics*, **9**(Suppl 12):S12, 2008. (SCI)

Darby Tien-Hao Chang*, Chih-Ching Wang and Jian-Wei Chen, "Using a kernel density estimation based classifier to predict species-specific microRNA precursors," *BMC Bioinformatics*, **9**(Suppl 12):S2, 2008. (SCI)

Ting-Ying Chien, **Darby Tien-Hao Chang***, Chien-Yu Chen, Yi-Zhong Weng and Chen-Ming Hsu, "E1DS: catalytic site prediction based on 1D signatures of concurrent conservation," *Nucleic Acids Research*, **vol. 36**, Web Server Issue, W291-W296, 2008. (SCI)

Darby Tien-Hao Chang, Yu-Yen Ou, Hao-Geng Hung, Meng-Han Yang, Chien-Yu Chen and Yen-Jen Oyang*, "Prediction of protein secondary structures with a novel kernel density estimation based classifier," *BMC Research Notes*, **1**:51, 2008.

Darby Tien-Hao Chang, Yi-Zhong Weng, Jung-Hsin Lin, Ming-Jing Hwang, and Yen-Jen Oyang, "Protemot: prediction of protein binding sites with automatically extracted geometrical templates", *Nucleic Acids Research*, **Vol. 34**, Web Server Issue, W303-W309, 2006. (SCI)

Darby Tien-Hau Chang, Yen-Jen Oyang, and Jung-Hsin Lin, "MEDock: a Web Server for Efficient Prediction of Ligand Binding Sites Based on a Novel Optimization Algorithm", *Nucleic Acids Research*, **Vol. 33**, Web Server Issue, W233-W238, 2005. (SCI)

Darby Tien-Hau Chang, Chien-Yu Chen, Wen-Chin Chung, Yen-Jen Oyang, Hsueh-Fen Juan, and Hsuan-Cheng Huang, "ProteMiner-SSM: A Web Server for Identifying Possible Protein-Ligand Interactions Based on Analysis of Protein Tertiary Substructures", *Nucleic Acids Research*, **Vol. 32**, Web Server Issue, W76-W82, 2004. (SCI)

SELECTED CONFERENCE PAPERS

Chen-Hao Hsueh, **Darby Tien-Hao Chang** and Chih-Yun Chien, "Predicting DNA-binding Proteins using Disorder Information," in poster proceedings of the 9th International Conference of Bioinformatics, Tokyo, Japan, Sepetmber 26-28, 2010.

Ting-Ying Chien, Chien-Yu Chen, Chih-Kang Lin, Chih-Wei Lin, Cheng-Yi Huang and **Darby Tien-Hao Chang**, "Prediction of DNA-binding profiles by protein-DNA complexes," in poster proceedings of the 9th International Conference of Bioinformatics, Tokyo, Japan, Sepetmber 26-28, 2010.

Jian-Wei Wu, **Darby Tien-Hao Chang** and Chien-Ju Li, "Predicting Protein-protein Interactions using a Hybrid Approach," in poster proceedings of the 9th International Conference of Bioinformatics, Tokyo, Japan, Sepetmber 26-28, 2010.

Chih-Hung Hsieh, **Darby Tien-Hao Chang** and Yen-Jen Oyang, "Data Classification with a Generalized Gaussian Components based Density Estimation Algorithm," in poster Proceedings of the 2009 International Joint Conference on Neural Networks, Atlanta, Georgia, 2009.

Yi-Zhong Weng, Chien-Kang Huang, Yu-Feng Huang, Chi-Yuan Yu and **Darby Tien-Hao Chang**, "Introducing sequence-order constraint into prediction of protein binding sites with automatically extracted templates," in proceedings of the 6th International Conference on Bioinformatics and Bioengineering, Tokyo, Japan, 2009.

INVITED TALKS

Predicting Protein-protein Interactions using a Hybrid Approach *December 2010*
The 2010 International Statistical Conference,
National Central University, Jhongli, Taiwan

Machine Learning in Bioinformatics - *December 2007*
Applications, concepts and our approaches
Department of Engineering Science and Ocean Engineering,
National Taiwan University, Taipei, Taiwan

Machine Learning in Bioinformatics - *December 2007*
Applications, concepts and our approaches
Institute of Biomedical Informatics,
National Yang Ming University, Taipei, Taiwan

From a Basic Statistical Concept to Advanced Bioinformatics *December 2007*
The 2nd Taiwan-Japan Bilateral Symposium on Bioinformatics,
Tainan, Taiwan

From a Basic Statistical Concept to Advanced Bioinformatics *November 2006*
Institute of Information Science,
Academic Sinica, Taipei, Taiwan

From a Basic Statistical Concept to Advanced Bioinformatics *November 2006*
The 6th Association of Asian Societies for Bioinformatics Symposium,
Singapore