Hindawi Publishing Corporation
Research Letters in Signal Processing
Volume 2008, Article ID 365152, 5 pages
doi:10.1155/2008/365152

Research Letter

Is Subcellular Localization Informative for Modeling Protein-Protein Interaction Signal?

Junfeng Liu,[1,2] Hongyu Zhao,[3] Jun Tan,[4] Dajie Luo,[4] Weichuan Yu,[5] E. James Harner,[4] and Weichung Joe Shih[1,2]

[1] *Division of Biometrics, The Cancer Institute of New Jersey, 195 Little Albany Street, New Brunswick, NJ 08901, USA*
[2] *Department of Biostatistics, School of Public Health, University of Medicine and Dentistry of New Jersey, 683 Hoes Lane West, Piscataway, NJ 08854, USA*
[3] *Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College Street, New Haven, CT 06520, USA*
[4] *Department of Statistics, West Virginia University, P.O. Box 6330, Morgantown, WV 26506, USA*
[5] *Department of Electronic and Computer Engineering, The Hong Kong University of Sciences and Technology, Clear Water Bay, Kowloon, Hong Kong, China*

Correspondence should be addressed to Junfeng Liu, ljfbacc@yahoo.com

Received 21 August 2007; Accepted 2 January 2008

Recommended by Jar-Ferr Kevin Yang

Statistical methods have been intensively applied in genomic signal processing (Dougherty et al. 2005). For budding yeast *Saccharomyces cerevisiae* with around 6000 proteins, genome-wide protein-protein-interaction (PPI) (Fromont-Racine et al. 2000, Ito et al. 2001, Newman et al. 2000, and Uetz et al. 2000 among others) and protein subcellular localization (PSL) (Huh et al. 2003) data recently became available and for the latter the presence of 4152 proteins is experimentally tested in each of the 22 subcellular compartments. Recent work shows that multiple biological sources are helpful for both PSL and PPI predictions, and this paper studies statistical feasibility of modeling PPI from PSL since PSLs may play different marginal or joint roles in the complex regulatory network. However, our results indicate that PSL may be controversial for this purpose as an independent source.

1. STATISTICAL METHODS

1.1. Two-way PPI count contingency table

We extracted 2712 PPIs from MIPS [1] which were available at http://hto-b.usc.edu/~msms/AssessInteraction/MIPSMatchYPD.txt as of 2005 and used by Lin and Zhao [2] for PPI network robustness study. We use 1641 PPIs with complete PSL information in Huh et al. [3], for example, protein A has a 22-dimensional PSL vector $\widetilde{L}_A = (L_{A,1}, L_{A,2}, \ldots, L_{A,22})$, where $L_{A,i} = 1$ represents presence of protein A at PSL i and $L_{A,i} = 0$ represents absence of protein A at PSL i. For proteins A and B, we create 44-dimensional PSL vector $\widetilde{L}_{AB}$ $(\widetilde{L}_A, \widetilde{L}_B)$ along with an exchanged counterpart $\widetilde{L}_{BA}$ $(\widetilde{L}_B, \widetilde{L}_A)$ for naive balance. Since log-linear model with large number (2^{44}) of cross-classified cells may lack power where the total PPI count is relatively small (<10 000), we instead explore an alternative two-way (22^2) contingency table whose rows (compartments: $i = 1, \ldots, 22$) and columns (compartments: $j = 1, \ldots, 22$) jointly assign each PPI into cell (i, j) with one protein in compartment i and the other one in compartment j $(i, j = 1, 2, \ldots, 22)$ (Figure 1). Note that one PPI may be redundantly counted due to multiple PSL occupation. Cytoplasm and nucleus likely play crucial roles since these two compartments hold most PPI entries and other compartment pairs have much less entries. Negative binomial model avoids overdispersion and shows ER to Golgi, lipid particle and nucleus may be significant effects for this two-way contingency table.

1.2. PSL correlation pattern

For 44-dimensional joined PSL vectors we calculate all $C_{44}^2 + C_{44}^1$ $(= 990)$ pairwise Pearson correlation coefficients

$$\mathrm{Corr}(\widetilde{L}_{AB}, \widetilde{L}_{AB}) = \begin{bmatrix} \mathrm{Corr}(\widetilde{L}_A, \widetilde{L}_A), \mathrm{Corr}(\widetilde{L}_A, \widetilde{L}_B) \\ \mathrm{Corr}(\widetilde{L}_B, \widetilde{L}_A), \mathrm{Corr}(\widetilde{L}_B, \widetilde{L}_B) \end{bmatrix}. \quad (1)$$

This calculation was carried over to the following four disjoint sets of protein pairs: (1) interacting protein pairs from PPIs (set [1]), (2) non-PPI protein pairs from those proteins with PPI (set [2]), (3) non-PPI protein pairs from those proteins without PPI (set [3]), and (4) non-PPI protein pairs from combining protein without PPI and protein with PPI (set [4]). As in Section 1.1, by selecting those PPIs (MIPS) with PSL information (Huh et al. [3]), we obtained 2883 proteins without PPI, 3282 exchanging PSL vectors for (set [1]), 1 591 338 exchanging PSL vectors for (set [2]), 8 308 806 exchanging PSL vectors for (set [3]) and 7 317 054 exchanging PSL vectors for (set [4]). 30% of non-PPI protein pairs and 69% of PPI protein pairs have colocalization, thus 31% of PPIs may be transient. The pseudoimages and contour plots for PSL correlations (1) are given in Figure 2, where the upper-left quadrat in panel 1 shows significant between-protein colocalization pattern for (set [1]) and no clear colocalization pattern occurs for between-protein PSLs for (sets [2,3,4]). These observations motivate us to study if between-protein PSL pattern could potentially discriminate between protein pairs with PPI and those without PPI.

1.3. Retrospective logistic regression

We propose a realistic model for quantifying PPI tendency from fused PSLs of proteins A and B (with exchanging). The PSL and PPI information is expressed as

$$\begin{aligned}(L_{A,1}, L_{A,2}, \ldots, L_{A,22}, L_{B,1}, L_{B,2}, \ldots, L_{B,22}) &\sim I_{AB},\\ (L_{B,1}, L_{B,2}, \ldots, L_{B,22}, L_{A,1}, L_{A,2}, \ldots, L_{A,22}) &\sim I_{BA},\end{aligned} \tag{2}$$

where $I_{AB}(= I_{BA})$ is the binary PPI indicator (response) and the logistic regression model is proposed to be logit

$$\begin{aligned}\Pr(\text{PPI} \mid A, B) = \beta_0 &+ \sum_{i=1}^{22} \beta_i (L_{A,i} + L_{B,i})\\ &+ \sum_{i<j} \beta_{ij} (L_{A,i}L_{A,j} + L_{B,i}L_{B,j})\\ &+ \sum_{i\le j} \beta'_{ij} (L_{A,i}L_{B,j} + L_{B,i}L_{A,j}),\end{aligned} \tag{3}$$

where β_0 and β_i imply default PPI probability and PPI tendency of single protein with PSL i, β_{ij}, and β'_{ij} represent PPI tendency of single protein with PSLs i and j and two proteins with PSLs i and j, respectively, where $i = j$ describes PPI tendency of two proteins with common PSL i. The number of model parameters is $1 + 2C_{22}^1 + 2C_{22}^2 = 507$. For efficiency we consider a reduced model $\beta_0 + \sum_{i\le j}\beta'_{ij}(L_{A,i}L_{B,j} + L_{B,i}L_{A,j})$ which incorporates second-order PSL effects between two proteins. The yeast interactome and proteome are inherent libraries and not subject to arbitrary experimental design, which indicates a retrospective (case-control) study. On the other hand, we have $\sim 18 \times 10^6$ total protein pairs and only $\sim 2 \times 10^3$ PPIs in our data. In order to overcome computer memory limitation and achieve reasonable sample sizes for both case (PPI) and control (non-PPI) groups, we need to select out a sample subset under statistical justification. For logistic model with responses y_is and predictors x_is, we let Z_i indicate whether subject i is selected and assume $\rho_1 = \Pr(Z_i = 1 \mid y_i = 1)$ and $\rho_0 = \Pr(Z_i = 1 \mid y_i = 0)$, both of them are free of x_i. If the logistic model based on all subjects has logit $(\Pr(y_i = 1 \mid x_i)) = \alpha + \beta x_i$, then the retrospective logistic regression (RLR) after selection probability adjustment would be logit $(\Pr(y_i = 1 \mid x_i, z_i = 1)) = \alpha + \log(\rho_1/\rho_0) + \beta x_i$ (Chapter 4.3.3, McCullagh and Nelder [4]). We apply case selection probability 1 and control selection probability 2×10^{-3} (3282 PPIs, 38 338 entries and 254 parameters) and identify around 60 significant effects. The resultant prospective PPI probabilities are to be adjusted based on foregoing theory.

1.4. PPI prediction from PSL

After fitting the preceding model, we apply certain threshold τ to the simple classification rule

$$\begin{aligned}&\Pr\left(\text{PPI} \mid \widetilde{L}_{AB} \text{ or } \widetilde{L}_{BA}\right) > \tau \Longrightarrow \text{PPI},\\ &\Pr\left(\text{PPI} \mid \widetilde{L}_{AB} \text{ or } \widetilde{L}_{BA}\right) \le \tau \Longrightarrow \text{non-PPI}.\end{aligned} \tag{4}$$

We randomly divide the whole dataset for retrospective study into 10 disjoint portions. Each portion (includes PPIs and non-PPIs in proportion) acts as one testing set and the other nine portions are combined into one training set for 10-fold cross validation. We classify each protein-protein pair in the testing set into PPI or non-PPI by comparing the calculated probabilities (from trained model parameters) with some threshold τ. We find that PPI probability median of the non-PPI subset in the training set is always equal to that of the PPI subset in the training set and the PPI probability median (1.88×10^{-4}) for PPI subset also equals that of non-PPI subset for the whole dataset in retrospective study. For retrospective study with PPI probability median threshold, we have specificity around 98% and sensitivity around 15%, RandomForest Breiman [5] in R reaches specificity around 99% and sensitivity around 20% and support vector machine (SVM) in R reaches specificity around 50% and sensitivity around 90%. The PPI probabilities from retrospective study dataset and 10-fold cross-validation are plotted in Figure 3. The logistic model-based classification results are found to be sensitive to threshold. If we use "$\Pr(\text{PPI} \mid \widetilde{L}_{AB} \text{ or } \widetilde{L}_{BA}) \ge \tau \Rightarrow$ PPI" and "$\Pr(\text{PPI} \mid \widetilde{L}_{AB} \text{ or } \widetilde{L}_{BA}) < \tau \Rightarrow$ non-PPI", where τ equals PPI probability median, then we obtained very different classification results. After prospective PPI probability adjustment, the threshold-based classification (4) is applied to the complete PPI and PSL data ([Sets 1,2,3,4], Section 1.2) and the resultant ROC curve is given in Figure 4 with area under curve (AUC) less than 0.5. Since we may simply invert this classifier to make AUC greater than 0.5, Figure 4 indicates that the proposed logistic regression model ((3) in Section 1.3) may not be highly sufficient even if this model is carefully chosen. We also observe the following facts: selection procedure in retrospective study may involve some bias, the joined PSL patterns (from two proteins) are finite with uncertain overlap between PPI set and non-PPI set, false positives and false negatives may exist in both PPI and PSL data and others. From statistical point of view, interprotein PSL pattern may not independently determine PPI tendency,

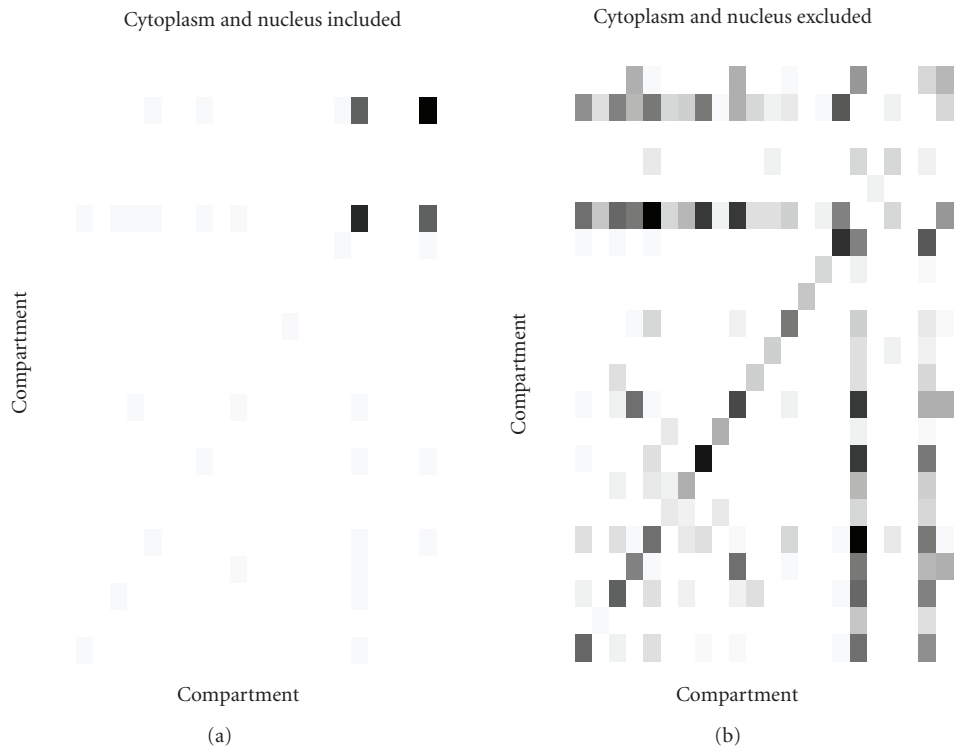

FIGURE 1: Pseudoimages for two-way PPI count contingency table by combining pairwise PSLs (Section 1.1). The left panel includes cytoplasm and nucleus and the right panel excludes cytoplasm and nucleus. From left (bottom) to right (top) on the $x(y)$-axis: compartments $[1:22]$.

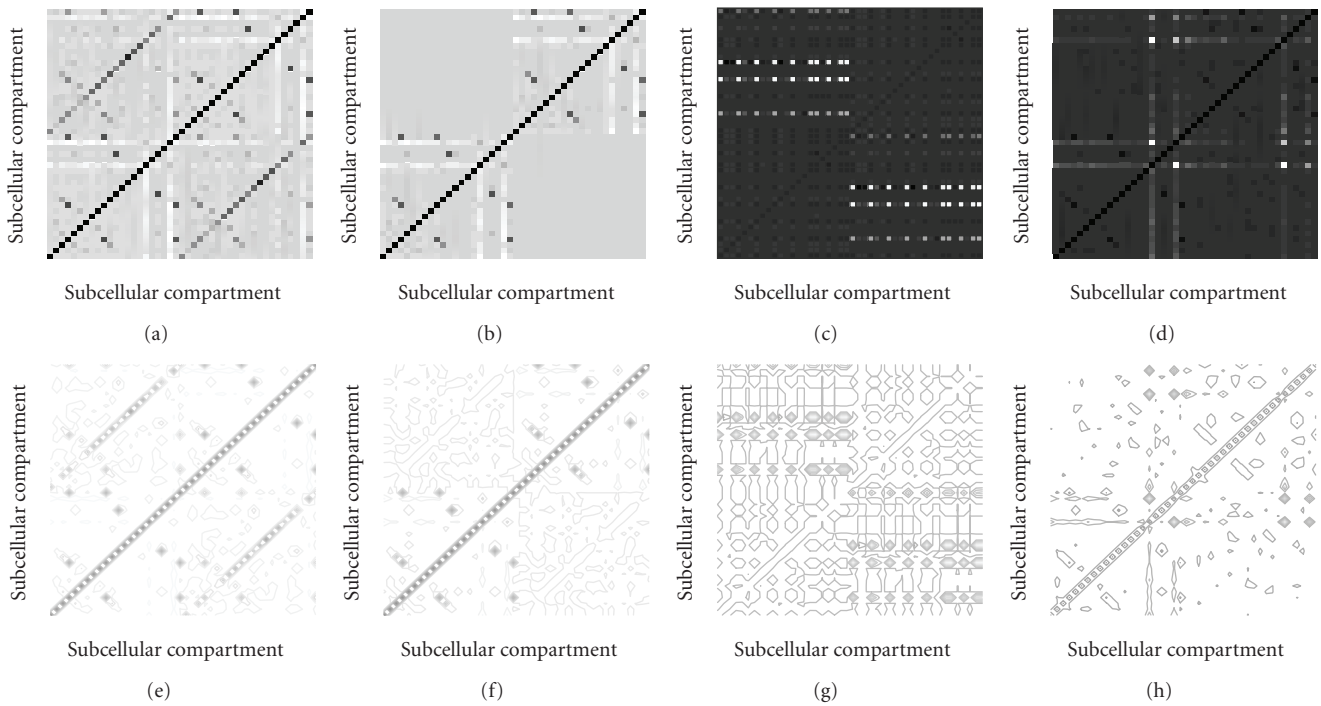

FIGURE 2: Correlation pseudoimages and contour plots for PPIs (set [1]: panel 1) and non-PPIs (sets [2,3,4], panels 2, 3, 4). From left (bottom) to right (top) on the $x(y)$-axis in each panel: compartments $[1:22]$ for protein $A(B)$.

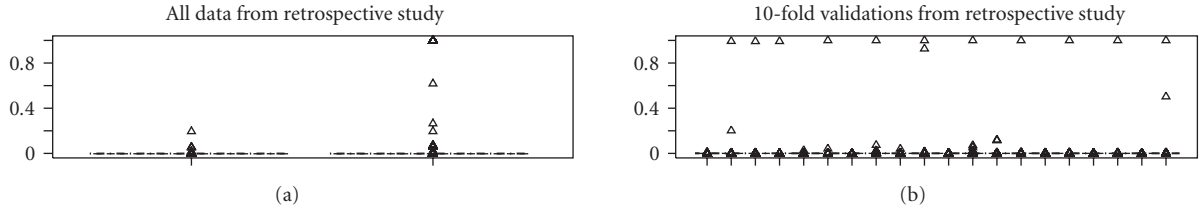

Figure 3: (Top panel) PPI probabilities from retrospective study. (Lower panel) PPI probabilities from 10-fold cross-validation (after prospective adjustment), each pair of consecutive boxplots is for individual testing set where PPI subset follows non-PPI subset.

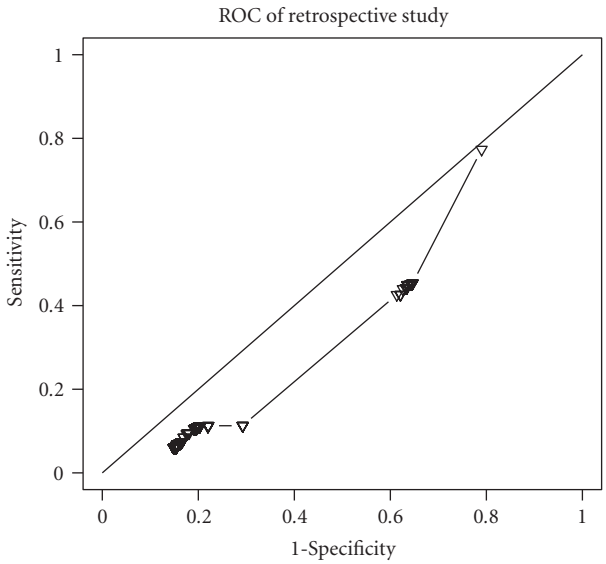

Figure 4: ROC (∇ connection) based on prospective PPI probability (adjusted from retrospective study) threshold-based classification.

and threshold-based PPI prediction rule may not discriminate PPI from non-PPI either. The former conclusion is also a major concern from biologists who consider PPI mechanism far beyond only PSL information.

2. DISCUSSION

In this article, we proposed statistical analysis of the association between PPI and PSL with the possibility of offering clues for further specific biological experiments. The aforementioned model is only one possible approach out of many helpful tries. It is likely that a totally different approach based on PSL information may lead to disparate results. As an alternative, if we could describe the distribution of 44-dimensional joined binary PSL vectors given PPI or non-PPI: $\Pr(\widetilde{L}_{AB} \mid \text{PPI})$ and $\Pr(\widetilde{L}_{AB} \mid \text{non-PPI})$, then armed with some prior PPI probability, say $\Pr(\text{PPI}) = 3 : (1.8 \times 10^4)$, we can predict PPI probability for joined PSL pattern $\widetilde{L}_{AB}$ by Bayes rule

$$\Pr\left(\text{PPI} \mid \widetilde{L}_{AB}\right) = \frac{\Pr\left(\widetilde{L}_{AB} \mid \text{PPI}\right)\Pr(\text{PPI})}{Q}, \tag{5}$$

where $Q = \Pr(\widetilde{L}_{AB} \mid \text{PPI})\Pr(\text{PPI}) + \Pr(\widetilde{L}_{AB} \mid \text{non-PPI})\Pr(\text{non-PPI})$. Section 1.2 is essentially an attempt to work on either the PPI or non-PPI set to study PSL pattern without considering the non-PPI or PPI counterpart, which may be only a matter of exploring $\Pr(\widetilde{L}_{AB} \mid \text{PPI})$ or $\Pr(\widetilde{L}_{AB} \mid \text{non-PPI})$ separately. However, the explicit probability of high-dimensional binary vector is difficult to be constructed. Empirical approaches (Sections 1.1 and 1.2, Huh et al. [3]) offer informative results from different perspectives. On the other hand, Liu et al. [6] modeled PPI based on domain-domain interaction information and computational PSL prediction from other sources which are also feasible, the readers are referred to Lu et al. [7], Szafron et al. [8], Höglund et al. [9], Horton et al. [10], Guda [11], Yu et al. [12], and Zhang et al. [13, 14] among many others.

ACKNOWLEDGMENTS

The authors are very grateful to the Associate Editor and two referees for the constructive comments which led to great improvement of their presentation. This research was partially supported by NIH/COBRE Grant NOT-RR-06-001 and NIH Grant GM59507 (*J. Liu*) and NIH/NCI Grant CA-072720-11 (*J. Liu and W. J. Shih*).

REFERENCES

[1] H. W. Mewes, D. Frishman, C. Gruber, et al., "MIPS: a database for genomes and protein sequences," *Nucleic Acids Research*, vol. 28, no. 1, pp. 37–40, 2000.

[2] N. Lin and H. Zhao, "Are scale-free networks robust to measurement errors?" *BMC Bioinformatics*, vol. 6, no. 1, p. 119, 2005.

[3] W.-K Huh, J. V. Falvo, L. C. Gerke, et al., "Global analysis of protein localization in budding yeast," *Nature*, vol. 425, pp. 686–691, 2003.

[4] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, Chapman & Hall, London, UK, 2nd edition, 1989.

[5] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[6] Y. Liu, N. Liu, and H. Zhao, "Inferring protein-protein interactions through high-throughput interaction data from diverse organisms," *Bioinformatics*, vol. 21, no. 15, pp. 3279–3285, 2005.

[7] Z. Lu, D. Szafron, R. Greiner, et al., "Predicting subcellular localization of proteins using machine-learned classifiers," *Bioinformatics*, vol. 20, no. 4, pp. 547–556, 2004.

[8] D. Szafron, P. Lu, R. Greiner, et al., "Proteome analyst: custom predictions with explanations in a web-based tool for

high-throughput proteome annotations," *Nucleic Acids Research*, vol. 32, Web Server issue, pp. W365–W371, 2004.

[9] A. Höglund, P. Dönnes, T. Blum, H.-W. Adolph, and O. Kohlbacher, "MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition," *Bioinformatics*, vol. 22, no. 10, pp. 1158–1165, 2006.

[10] P. Horton, K.-J Park, T. Obayashi, and K. Nakai, "Protein subcellular localization prediction with WoLF PSORT," in *Proceedings of the 4th Asia-Pacific Bioinformatics Conference (APBC '06)*, pp. 39–48, Taipei, Taiwan, February 2006.

[11] C. Guda, "pTARGET: a web server for predicting protein subcellular localization," *Nucleic Acids Research*, vol. 34, Web Server issue, pp. W210–W213, 2006.

[12] C.-S. Yu, Y.-C. Chen, C.-H. Lu, and J.-K. Hwang, "Prediction of protein subcellular localization," *Proteins*, vol. 64, no. 3, pp. 643–651, 2006.

[13] T. Zhang, Y. Ding, and S. Shao, "Protein subcellular location prediction based on pseudo amino acid composition and immune genetic algorithm," in *Proceedings of the International Conference on Intelligent Computing (ICIC '06)*, vol. 4115, part 3, pp. 534–542, Kunming, China, August 2006.

[14] T. Zhang, Y. Ding, and K.-C. Chou, "Prediction of protein subcellular location using hydrophobic patterns of amino acid sequence," *Computational Biology and Chemistry*, vol. 30, no. 5, pp. 367–371, 2006.