Correlation of Metabolic Pathways with the Primary Structure in Acetylated Proteins

Zheng Rong Yang*,[1] and Kuo-Chen Chou[2]

[1]*School of Biosciences, University of Exeter, UK*

[2]*Gordon Life Science Institute, San Diego, California, USA*

Abstract: Signaling pathways are the major component in cellular networks, but most studies done recently on signaling pathways were either aimed to enhance various molecular predictions using pathways as contexts or focused to predict pathways indirectly. The former assumed that the pathways for the biomolecules (genes or proteins) used in the modelling were known. Although the latter was well suiting the biosciences researches at the systems level, the indirect predictions would more or less rely on the prediction accuracy of other systems. So far no work whatsoever has been done for studying the direct correlation between signaling pathways and protein primary structures although acetylated proteins are one of the main players in metabolic signaling pathways. In order to investigate their correlation, the sequences of 76 experimentally verified acetylated proteins were downloaded from NCBI. They cover three major metabolic pathways, i.e., biosynthesis, degradation, and metabolism. Without any a prior knowledge about how these three metabolic pathways are correlated with the primary structures of acetylated proteins, we proposed some classification models between the pathways. It has been found through computer simulations that the signaling pathways are indeed correlated with the primary structures in acetylated proteins, further demonstrating the well-known biological law that sequence determines structure and structure determines function.

INTRODUCTION

Most cellular functions are based on the communications among signaling molecules [1]. Signals are travelling among cells delivering instructions for cell development, immunity development, and normal tissue homeostasis. Abnormal signals are the main course for altered cellular functions leading to disease development [2-4]. The most fundamental issue related with systems level biosciences or life sciences researches [5] is that how we can effectively and accurately identify pathways in different organisms [6-8]. Because the identification made in wet-laboratories is time-consuming and labour-intensive, computational approaches are urgently required. Actually, many lines of evidences have indicated that computational approaches, such as structural bioinformatics [9-13], molecular docking [14-18], pharmacophore modeling [19, 20], QSAR [21-26], protein cleavage site prediction [27-30], protein subcellular location prediction [31-35], identification of membrane proteins and their types [36], identification of enzymes and their functional classes [37], identification of G-protein coupled receptors (GPCR) and their types [38, 39], identification of proteases and their types [40], and signal peptide prediction [41, 42] can provide very useful information for drug design in a timely manner. It is equally important to provide proper functional annotation of genome sequences and to accurately integrate proteins into the signaling pathways [43]. In other words, the correlation pattern between signaling pathways and protein primary structures, if being revealed, will help fast, effective, and accurate identification of the pathways concerned.

*Address correspondence to this author at the School of Biosciences, University of Exeter, UK; E-mail: Z.R.Yang@exeter.ac.uk

Bioinformatics studies involving signaling pathway are mainly using signaling pathways for genome annotation and signaling network analysis. For the former, the signaling pathway information was used for classifying gene expression data [44]. In that study, the classification of gene expression data with two or three classes related to diseases were enhanced by the pathway information. In the corresponding data set, each gene is associated with a pathway with such classification that a pathway is ranked higher if the prediction error associated with the pathway is lower. The same research group [45] later published another related work on clustering pathways.

Some web-based tools were developed for the visualisation of pathways. For instance, PathExpress was developed for gene expression data with functional context extracted from KEGG ligand database [46]. ArrayXPath II was developed using the scalable vector graphics technique for gene expression data based on integrated biological pathway resources [47]. Path-A was developed for metabolic pathway prediction, where machine learning algorithms and homology alignment algorithms were used to build basic reaction classifiers. Based on the reaction predictions, pathways were then integrated [48]. In that work, the prediction if a query protein is a catalyst of a particular reaction was implemented using the support vector machine, hidden Markov models, and BLAST.

Because it is difficult to predict pathways in a completely automatic way, some work focused on identifying pathway fragments based on text mining [6]. In addition, database technology was employed to curate pathway data and for visualisation [49]. Because of incompleteness of pathway information in many databanks, identifying missing enzymes of pathways has also been an important research direction in

1875-0362/08

pathway analysis [50-52]. In these works, pathway context was used to guide the identification of missing enzymes in some specific pathways. Having built some pathway networks, it is very often for biologists to pin-point where a new sequenced protein is in the network, sequence similarity was therefore used in QPath (Querying pathway) for the prediction of pathways in a pathway network [53].

Although it is known that many essential pathways remain unknown or incomplete for newly sequenced proteins [6], very few work has been conducted to predict pathways for query sequences. To the best of our knowledge, the only one in the literature was predicting pathways implicitly [48], where catalytic reactions of various pathways were precompiled. The prediction was conducted on individual reactions. If a query protein turns out to exactly match the reactions belonging to a pathway, the query protein was predicted to belong to the pathway.

There are two opposite chemical activities in cells [54]. The catabolic pathways are used to break down large molecules to smaller molecules. Small molecules are then used as building blocks to form new molecules by the anabolic pathways. The anabolic pathways are referred to as biosynthesis processes. The combination of the two is the metabolism. Metabolism is a vital process of life maintenance through chemical reactions in living organisms [7, 55]. Degradation is a process of breaking down molecules in living organisms [54].

Acetylation is a reaction which introduces an acetyl functional group into an organic compound and is also a major metabolic pathway. In some disease developments, it has been found that acetylation plays an important role. For instance, the interplay of acetylation and methylation in gene transcription regulation is one of contributing factors in cancer development and has been investigated if patterns can be used for disease diagnosis [56]. Histone deacetylation inhibitors, if properly designed, can be used for tumour radiosensitization [57].

The relation between biosynthesis and acetylation can be seen as a cycle in which ATP generates S-adenosylmethionine for biosynthesis while acetyl-CoA is consumed in the acetylation [58]. GDP-N-acetyl-d-perosamine was found to be a precursor of the LPS-O-antigen biosynthesis in E. coli [59]. The acetylated polyamines induced by spermidine/ spermine N(1)-acetyltransferase increased in biosynthesis [60]. The naturally occurring proteins with acetylated NH2-termninal will normally be degraded except for the involvement of a conjugating enzyme (possibly a ubiquitin-protein ligase) [61]. It was shown that acetylated GATA-1 can be targeted for degradation. This is completed in the uniquitin/proteasom pathway. It was suggested that acetylation may signal unquitination. From this, GATA-1 is led to degradation [62]. In studying the inter-individual variability in 5-fluoroiracil metabolism, it was found that N-acetylation played a more important role than hydrolysis [63]. In studying oxidative hair dyes, it was found that N-acetylation is a predominant metabolism pathway [64]. In studying severe sulphonamide hypersensitivity reaction, it was found that N-acetylation of parent compound is the important metabolic pathway [65] and patients with slow acetylators showed diverse reactions [66].

The present study was initiated in an attempt to develop a computational approach by which one can predict the metabolic pathways in which a query protein involved according to its primary sequence structure.

MATERIALS AND METHODS

Benchmark Dataset

Searching NCBI database for acetylated proteins led to 1251 hits. The following rules were used to select proper data for the current study. First, a sequence without pathway annotation (denoted by [PATHWAY]) was discarded. Second, a sequence without any experimentally verified acetylation residue was dropped. From this, the sequences with the annotation of [PATHWAY] contained five types of acetylations, i.e., lysine, serine, methionine, threonine, and valine. An experimentally verified acetylated residue was denoted by "/experiment=…" in the NCBI GenPept file. A computer program written in C language for these rules was used to scan the NCBI GenPept files downloaded from NCBI automatically. This led to 87 sequences being kept and the remaining 1164 sequences discarded. To reduce the redundancy and homology bias, the sequence identity was checked using the CD-HIT algorithm [67-69] to remove those sequences which have $\geq 90\%$ pairwise sequence identity to any other in a same subset. This further removed 11 sequences and finally we obtained 76 sequences in the benchmark dataset for the current study. The accession numbers and sequences of the 76 proteins are given in the Online Supporting Information A.

Table 1. Breakdown of the 76 Proteins in the Benchmark Dataset According to their Five Involvement Modes with the Three Major Pathways

Pathway		Number of Sequences
Name	Abbreviated Name	
Biosynthesis	BIO	12
Degradation	DEG	26
Metabolism	MET	21
Metabolism + Biosynthesis	MET+BIO	12
Metabolism + Degradation	MET+DEG	5
Total		76

Table 2. The Sub-Pathways for the Three Major Pathways

Pathway	Sub-pathways			
Biosynthesis	AMP biosynthesis *via* salvage pathway	Carnitine biosynthesis	Cholesterol biosynthesis	Dopamine biosynthesis
	Ergosterol biosynthesis	Estrogen biosynthesis	Glycogen biosynthesis	Homocysteine biosynthesis
	Lanosterol biosynthesis	L-cysteine biosynthesis	Malonyl-CoA biosynthesis	Mevalonic acid biosynthesis
	Prostaglandin biosynthesis	Protoporphyrin-IX biosynthesis	S-adenosyl-L-methionine biosynthesis	
	Steroid biosynthesis	Amine and polyamine biosynthesis	Amino-acid biosynthesis	Carbohydrate biosynthesis
	Catecholamine biosynthesis	Glycan biosynthesis	Metabolic intermediate biosynthesis	Protein biosynthesis
Degradation	Ethanol degradation	HMG-CoA degradation	L-kynurenine degradation	L-leucine degradation
	L-phenylalanine degradation	Pectin degradation	Sarcosine degradation	Uric acid degradation
	Amine and polyamine degradation	Amino-acid degradation	Carbohydrate degradation	
Metabolism	Glyoxylate and dicarboxylate metabolism	Propanoate metabolism	Retinol metabolism	Alcohol metabolism
	Carbohydrate metabolism	Cofactor metabolism	Glycan metabolism	Lipid metabolism
	Metabolic intermediate metabolism	Nitrogen metabolism	One-carbon metabolism	Porphyrin metabolism

The 76 sequences involve many sub-pathways of three major metabolic pathways, most of which occur only once in one protein sequence. This made the task of model construction very difficult. To simplify the problem, let us focus the following three major pathways: biosynthesis (BIO), degradation (DEG), and metabolism (MET). Although some of the 76 proteins may involve two of the three major pathways, none involves both BIO and DEG, or more than two of the three pathways. Therefore, according to the modes of their involvement with the three pathways, we can classify the 76 proteins into the following five categories: (1) BIO, (2) DEG, (3) MET, (4) BIO+MET, and (5) DEG+MET. Listed in Table **1** are numbers of the 76 proteins among the five categories. Moreover, the 23 sub-pathways of BIO, 11 sun-pathways of DEG, and 15 sub-pathways of MET are listed in Table **2**.

Experimental Design

The computer simulation was designed in several steps each of which was for a specific target. First, sequence data needed to be coded to fit the machine learning algorithms for classification model construction. The coded data were then fed to the machine learning algorithms for training the predictor. In order to have an automatic model construction process, the evolutionary algorithm was used to tune the predictor for optimizing its parameters.

Each protein sequence was coded as a 20-D (dimensional) vector [70, 71] with each component representing the occurrence frequency of one of the 20 native amino acids. For instance, if there are 20 alanine in a sequence while the sequence length is 100 (with 100 residues), the frequency or the first component in this 20-D vector is 0.2.

The SVM (support vector machine) algorithm [72, 73] was used as the identification engine for the current study. SVM has been successfully used to predict protein subcellular locations, membrane protein types, and other protein attributes (see, e.g., [74-76]). For a brief introduction of using SVM to classify protein attributes based on a discrete model, refer to [75]. The SVM^{light} package (http://svmlight.joachims.org/) [72] was employed for model construction and evaluation. The package provides four kernel functions, i.e. the linear kernel, radial-basis function kernel, sigmoid function kernel and polynomial kernel. The radial basis function was found the best to fit this application and was used. The package needs the user to tune three hyper-parameters (C, J and γ). The C parameter was designed for trading-off between training and testing error. The J parameter was introduced for dealing with heavily imbalanced data. The γ parameter is associated with the radial basis function to determine the sensitivity of the function. The combination of these three parameters sits in a very large space and finding the best one is actually a non-trivial problem. The evolutionary algorithm was therefore used for this optimization problem. Through many generations of multiple-solution competition, the final solution is believed to be most close to the best solution.

RESULTS

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for

Table 3. The Jackknife Success Rates in Identifying the Five Classes of Pathways for the 76 Proteins in the Benchmark Dataset (see the Online Supporting Information A)

Pathway Class	Number of Proteins	Number of Correctly Predicted Events	Success Rate
BIO	12	7	58.3%
DEG	26	18	69.2%
MET	21	15	71.4%
METBIO	12	6	50%
METDEG	5	4	80%
Ocerall	76	50	65.8%

its effectiveness in practical application: independent dataset test, subsampling test, and jackknife test [77]. However, as elucidated in [34] and demonstrated by Eq.50 of [78], among the three cross-validation methods, the jackknife test is deemed the most objective that can always yield a unique result for a given benchmark dataset, and hence has been increasingly used by investigators to examine the accuracy of various predictors [1, 17, 79-102]. Accordingly, the jackknife test was also used to examine the success rates in identifying the pathway a query protein involves with. The optimal C values are 4977, 865, 5782, 8205, 1039 for the BIO, DEG, MET, BIO+MET, and DEG+MET classifications, respectively. The optimal J values are 2685, 6748, 1530, 1358, and 20021039 for the BIO, DEG, MET, BIO+MET, and DEG+MET classifications, respectively. The γ values are 1.73, 1.11, 0.26, 1.87, and 1.10 for the BIO, DEG, MET, BIO+MET, and DEG+MET classifications, respectively.

Listed in Table **3** are the jackknife success rates by the current approach in identifying the five classes of pathways for the 76 proteins. As we can see from the table, the overall success rate is about 66%. Let us imagine: if the involvements of the 76 proteins are completely randomly distributed among the 5 possible pathways, the overall success rate by random assignments would generally be $1/5 = 20\%$; if the random assignments are weighted according to the number of proteins in each pathway class (see column 2 of Table **3**), then the overall success rate would be [103]

$$\frac{1}{76^2}\left(12^2 + 26^2 + 21^2 + 12^2 + 5^2\right) = 24.8\%$$

which is about 41% lower than the overall success rate by the current approach, indicating that the metabolic pathways are really correlated with the primary structure in acetylated proteins.

CONCLUSION

It has been demonstrated through this study that there exists a correlation between the signaling pathways and the protein primary structures. This is a quite encouraging sign, indicating that it is possible to predict the pathway property or the involvement of a query protein, and hence its functions at the systems level can be analyzed as well. Particularly, for a protein known from some disease-related tissue, it is possible to use the current approach to explore which kinds of signaling pathways might be triggered for the disease development. It is instructive to point out that in the current approach, the simplest discrete model, i.e., the 20-D vector was adopted to express the protein samples. It is anticipated that if using more sophisticated discrete models [78], such as the pseudo amino acid (PseAA) composition approach [104] or functional domain (FunD) approach [36], or the hybridization approach by fusing FunD with the sequential evolution information [40], the success rates in predicting protein metabolic pathways will be further enhanced. The bioinformatics tool thus established will be very useful for studying biomedicine at the systems level.

REFERENCES

[1] X. Xiao, S. Shao, Y. Ding, Z. Huang, X. Chen and K.C. Chou, "An application of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation", *Journal of Theoretical Biology*, vol. 235, pp. 555-565, 2005.

[2] K.C. Chou, "Review: Prediction of protein signal sequences", *Current Protein and Peptide Science*, vol. 3, pp. 615-622, 2002.

[3] K.C. Chou and Y.D. Cai, "Predicting protein-protein interactions from sequences in a hybridization space", *Journal of Proteome Research*, vol. 5, pp. 316-322, 2006.

[4] H. Gonzalez-Díaz, S. Vilar, L. Santana and E. Uriarte, "Medicinal chemistry and bioinformatics - current trends in drugs discovery with networks topological indices", *Current Topics in Medicinal Chemistry*, vol. 10, pp. 1015-1029, 2007.

[5] P. Kelley, R. Sharan, R.M. Karp, T. Sittler, D.E. Root, B.R. Stockwell and T. Ideker, "Conserved pathways within bacteria and yeast as revealed by global protein network alignment", *Proceedings of the National Academy of Sciences of the United States*, vol. 100, pp. 11394-11399, 2003.

[6] A. Cakmak and Q. Qzsoyoglu, "Mining biological networks for unknown pathways", *Bioinformatics*, vol. 23, pp. 2775-2783, 2007.

[7] K.C. Chou, Y.D. Cai and W.Z. Zhong, "Predicting networking couples for metabolic pathways of Arabidopsis", *EXCLI Journal*, vol. 5, pp. 55-65, 2006.

[8] H. Gonzalez-Díaz, Y. Gonzalez-Díaz, L. Santana, F.M. Ubeira and E. Uriarte, "Proteomics, networks, and connectivity indices", *Proteomics*, vol. 8, pp. 750-778, 2008.

[9] K.C. Chou, "Insights from modelling the 3D structure of the extracellular domain of alpha7 nicotinic acetylcholine receptor", *Biochemical and Biophysical Research Communication*, vol. 319, pp. 433-438, 2004.

[10] K.C. Chou, "Modelling extracellular domains of GABA-A receptors: subtypes 1, 2, 3, and 5", *Biochemical and Biophysical Research Communications*, vol. 316, pp. 636-642, 2004.

[11] K.C. Chou, "Molecular therapeutic target for type-2 diabetes", *Journal of Proteome Research*, vol. 3, pp. 1284-1288, 2004.

[12] K.C. Chou, "Review: Structural bioinformatics and its impact to biomedical science", *Current Medicinal Chemistry*, vol. 11, pp. 2105-2134, 2004.

[13] K.C. Chou, "Coupling interaction between thromboxane A2 receptor and alpha-13 subunit of guanine nucleotide-binding protein", *Journal of Proteome Research*, vol. 4, pp. 1681-1686, 2005.

[14] K.C. Chou, D.Q. Wei and W.Z. Zhong, "Binding mechanism of coronavirus main proteinase with ligands and its implication to

drug design against SARS", (Erratum: ibid., 2003, vol.310, 675). *Biochemical and Biophysical Research Communication*, vol. 308, pp. 148-151, 2003.

[15] W.N. Gao, D.Q. Wei, Y. Li, H. Gao, W.R. Xu, A.X. Li and K.C. Chou, "Agaritine and its derivatives are potential inhibitors against HIV proteases", *Medicinal Chemistry*, vol. 3, pp. 221-226, 2007.

[16] Y. Li, D.Q. Wei, W.N. Gao, H. Gao, B.N. Liu, C.J. Huang, W.R. Xu, D.K. Liu, H.F. Chen and K.C. Chou, "Computational approach to drug design for oxazolidinones as antibacterial agents", *Medicinal Chemistry*, vol. 3, pp. 576-582, 2007.

[17] J.F. Wang, D.Q. Wei, C. Chen, Y. Li and K.C. Chou, "Molecular modeling of two CYP2C19 SNPs and its implications for personalized drug design", *Protein and Peptide Letters*, vol. 15, pp. 27-32, 2008.

[18] R. Zhang, D.Q. Wei, Q.S. Du and K.C. Chou, "Molecular modeling studies of peptide drug candidates against SARS", *Medicinal Chemistry*, vol. 2, pp. 309-314, 2006.

[19] K.C. Chou, D.Q. Wei, Q.S. Du, S. Sirois and W.Z. Zhong, "Review: Progress in computational approach to drug development against SARS", *Current Medicinal Chemistry*, vol. 13, pp. 3263-3270, 2006.

[20] S. Sirois, D.Q. Wei, Q.S. Du and K.C. Chou, "Virtual screening for SARS-CoV protease based on KZ7088 pharmacophore points", *Journal of Chemical Information and Computer Science*, vol. 44, pp. 1111-1122, 2004.

[21] M.A. Dea-Ayuela, Y. Perez-Castillo, A. Meneses-Marcel, F.M. Ubeira, F. Bolas-Fernandez, K.C. Chou and H. Gonzalez-Diaz, "HP-Lattice QSAR for dynein proteins: Experimental proteomics (2D-electrophoresis, mass spectrometry) and theoretic study of a Leishmania infantum sequence", *Bioorganic and Medicinal Chemistry*, vol. 16, pp. 7770-7776, 2008.

[22] Q.S. Du, Huang, R. B., Wei, Y. T., Du, L. Q. and K.C. Chou "Multiple field three dimensional quantitative structure-activity relationship (MF-3D-QSAR)", *Journal of Computational Chemistry*, vol. 29, pp. 211-219, 2008.

[23] Q.S. Du, Mezey, P. G. and K.C. Chou "Heuristic Molecular Lipophilicity Potential (HMLP): A 2D-QSAR Study to LADH of Molecular Family Pyrazole and Derivatives", *Journal of Computational Chemistry*, vol. 26, pp. 461-470, 2005.

[24] H. Gonzalez-Díaz, A. Perez-Bello, E. Uriarte and Gonzalez-Diaz, "QSAR study for mycobacterial promoters with low sequence homology", *Bioorganic and Medicinal Chemistry Letter*, vol. 16, pp. 547-53, 2006.

[25] H. Gonzalez-Díaz, A. Sanchez-Gonzalez and Y. Gonzalez-Diaz, "3D-QSAR study for DNA cleavage proteins with a potential antitumor ATCUN-like motif", *Journal of Inorganic Biochemistry*, vol. 100, pp. 1290-7, 2006.

[26] F.J. Prado-Prado, H. Gonzalez-Diaz, Q.M. de la Vega, F.M. Ubeira and K.C. Chou, "Unified QSAR approach to antimicrobials. Part 3: First multi-tasking QSAR model for Input-Coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds", *Bioorganic and Medicinal Chemistry*, vol. 16, pp. 5871-5880, 2008.

[27] K.C. Chou, "A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins", *Journal of Biological Chemistry*, vol. 268, pp. 16938-16948, 1993.

[28] K.C. Chou, Review: "Prediction of HIV protease cleavage sites in proteins", *Analytical Biochemistry*, vol. 233, pp. 1-14, 1996.

[29] Q.S. Du, S.Q. Wang, Z.Q. Jiang, W.N. Gao, Y.D. Li, D.Q. Wei and K.C. Chou, "Application of bioinformatics in search for cleavable peptides of SARS-CoV Mpro and chemical modification of octapeptides", *Medicinal Chemistry*, vol. 1, pp. 209-213, 2005.

[30] H.B. Shen and K.C. Chou, "HIVcleave: a web-server for predicting HIV protease cleavage sites in proteins", *Analytical Biochemistry*, vol. 375, pp. 388-390, 2008.

[31] K.C. Chou and H.B. Shen, Hum-PLoc: "A novel ensemble classifier for predicting human protein subcellular localization", *Biochemical and Biophysical Research Communication*, vol. 347, pp. 150-157, 2006.

[32] K.C. Chou and H.B. Shen, "Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers", *Journal of Proteome Research*, vol. 5, pp. 1888-1897, 2006.

[33] K.C. Chou and H.B. Shen, "Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites", *Journal of Proteome Research*, vol. 6, pp. 1728-1734, 2007.

[34] K.C. Chou and H.B. Shen, "Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms", *Nature Protocols*, vol. 3, pp. 153-162, 2008.

[35] H.B. Shen and K.C. Chou, "Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites", *Biochemical and Biophysical Research Communication*, vol. 355, pp. 1006-1011, 2007.

[36] K.C. Chou and H.B. Shen, "MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM", *Biochemical and Biophysical Research Communication*, vol. 360, pp. 339-345, 2007.

[37] H.B. Shen and K.C. Chou, EzyPred: "A top-down approach for predicting enzyme functional classes and subclasses", *Biochemical and Biophysical Research Communication*, vol. 364, pp. 53-59, 2007.

[38] K.C. Chou, "Prediction of G-protein-coupled receptor classes", *Journal of Proteome Research*, vol. 4, pp. 1413-1418, 2005.

[39] K.C. Chou and D.W. Elrod, "Bioinformatical analysis of G-protein-coupled receptors", *Journal of Proteome Research*, vol. 1, pp. 429-433, 2002.

[40] K.C. Chou and H.B. Shen, "ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information", *Biochemical and Biophysical Research Communication*, doi:10.1016/j.bbrc.2008.08.125, 2008.

[41] K.C. Chou and H.B. Shen, "Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides", *Biochemical and Biophysical Research Communication*, vol. 357, pp. 633-640, 2007.

[42] H.B. Shen and K.C. Chou, "Signal-3L: a 3-layer approach for predicting signal peptide", *Biochemical and Biophysical Research Communication*, vol. 363, pp. 297-303, 2007.

[43] M.B. Yaffe, G.G. Leparc, J. Lai, T. Obata, S. Volinia and L.C. Cantley, "A motif-based profile scanning approach for genome-wide prediction of signaling pathways", *Nature Biotechnology*, vol. 19, pp. 348-53, 2001.

[44] H. Pang, A. Lin, M. Holford, B.E. Enerson, B. Lu, M.P. Lawton, E. Floyd and H. Zhao, "Pathway analysis using random forests classification and regression", *Bioinformatics*, vol. 22, pp. 2028-2036, 2006.

[45] H. Pang, H. Zhao, "Building pathway clusters from random forests classification using class votes", *BMC Bioinformatics*, vol. 9, pp. 87, 2008.

[46] N. Goffard and G. Weiller, "PathExpress: a web-based tool to identify relevant pathways in gene expression data", *Nucleic Acid Research*, vol. 35, pp. w176-81, 2007.

[47] H.J. Chung, C.H. Park, M.R. Han, S. Lee, J.H. Ohn, J. Kim, J. Kim and J.H. "ArrarXPath II: mapping and visualizing microarray gene-expression data with biomedical ontologies and integrated biological pathway resources using scalable vector graphics", *Nucleic Acids Research*, vol. 33, pp. w621-6, 2005.

[48] L. Pireddu, D. Szafron, P. Lu and R. Greiner, "The Path-A metabolic pathway prediction web server", *Nucleic Acids Reasearch*, vol. 34, pp. W714-9, 2006.

[49] E. Urbanczyk-Wochniak and L.W. Summer, "MedicCyc: a biochemical pathway database for Medicago Truncatula", *Bioinformatics*, vol. 23, pp. 1418-1423, 2007.

[50] M.L. Green, P.D. Karp, "Using genome-context data to identify specific types of functional associations in pathway/genome databases", *Bioinformatics*, vol. 23, pp. i205-11, 2007.

[51] A. Osterman and R. Overbeek, "Missing genes in metabolic pathways: a comparative genomics approach", *Current Opinion*, vol. 7, pp. 238-251, 2003.

[52] Y. Yamanishi, H. Mihara, M. Osaki, H. Muramatsu, N. Esaki, T. Sato, Y. Hizukuri, S. Goto and M. Kanehisa, "Prediction of missing enzyme genes in a bacterial metabolic network", *FEBS Journals*, vol. 274, pp. 2262-73, 2007.

[53] T. Shlomi, D. Segal, E. Ruppin and R. Sharan, "QPath: a method for querying pathways in a protein-protein interaction network", *BMC Bioinformatics*, vol. 7, pp. 199, 2006.

[54] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, *Molecular Biology of the Cell*. 4th ed, Garland Science, New York, 2002.

[55] E. Smith and H.J. Morowitz, "Universality in intermediary metabolism", *Proceedings of the National Academy of Sciences of the United States*, vol. 101, pp. 13168-73, 2004.

[56] V. Shukla, T. Vaissiere and Z. Herceg, "Histone acetylation and chromatin signature in stem cell identity and cancer", *Mutation Research*, vol. 637, pp. 1-15, 2007.

[57] K. Camphausen and P.J. Tofilon, "Inhibition of histone deacetylation: a strategy for tumor radiosensitization", *Journal Clinical Oncology*, vol. 25, pp. 4051-6, 2007.

[58] A.E. Pegg, "Spermidine/spermine-N(1)-acetyltransferase: a key metabolic regulator", *Ammerican Journal of Physiology Endocrinology and Metabolism*, vol. 294, pp. E995-1010, 2008.

[59] C. Albermann and H. Beuttler, H. "Identification of the GDP-N-acetyl-d-perosamine producing enzymes from Escherichia coli O157:H7", *FEBS Letters*, vol. 582, pp. 479-84, 2008.

[60] D.L. Kramer, P. Diegelman, J. Jell, S. Vujcic, S. Merali and C.W. Porter, "Polyamine acetylation modulates polyamine metabolic flux, a prelude to broader metabolic consequences", *Journal of Biological Chemistry*, vol. 283, pp. 4241-51, 2008.

[61] A. Mayer, N.R. Siegel, A.L. Schwartz and A. Ciechanover, "Degradation of proteins with acetylated amino termini by the ubiquitin system", *Science*, vol. 244, pp. 1480-1483, 1989.

[62] Hernandez-Hernandez, P. Ray, G. Litos, M. Ciro, S. Ottolenghi, H. Beug and J. Boyes, "Acetylation and MAPK phosphorylation cooperate to regulate the degradation of active GATA-1", *The EMBO Journal*, vol. 25, pp. 3264-3274, 2006.

[63] T. Niwa, T. Shiraga, Y. Ohno and A. Kagayama, "Interindividual variability in 5-fluorouracil metabolism and procainamide N-acetylation in human liver cytosol", *Biological and Pharmaceutical Bulletin*, vol. 28, pp. 1071, 2005.

[64] J. Skare, G. Nohynek, R. Powrie, C. Goebel, S. Pfuhler, D. Duche, A. Zeller, M. Aardema, T. Hu and W. Meuling, "Metabolism of oxidative hair dyes: An overview of *in vitro*, *in vivo*, and clinical studies", *Toxicology Letters*, vol. 172, pp. S31, 2007.

[65] Anonymous, "Hypersensitivity to sulphonamides - a clue?" *The Lancet*, vol. 328, pp. 958-959, 1986.

[66] M.J. Rieder N.H. Shear, A. Kanee, B.K. Tang and S.P. Spielberg, "Prominence of slow acetylator phenotype among patients with Sulfonamide hypersensitivity reactions", *Clinical Pharmacology and Therapeutics*, vol. 49, pp. 13-17, 1991.

[67] W. Li, L. Jaroszewski and A. Godzik, "Clustering of highly homologous sequences to reduce the size of large protein databases", *Bioinformatics*, vol. 17, pp. 282-283, 2001.

[68] W. Li, L. Jaroszewski and A. Godzik, "Tolerating some redundancy significantly speeds up clustering of large protein databases", *Bioinformatics*, vol. 18, pp. 77-82, 2002.

[69] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences", *Bioinformatics*, vol. 22, pp. 1658-1659, 2006.

[70] K.C. Chou, "A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space", *Proteins: Structure, Function and Genetics*, vol. 21, pp. 319-344, 1995.

[71] K.C. Chou and C.T. Zhang, "Predicting protein folding types by distance functions that make allowances for amino acid interactions", *Journal of Biological Chemistry*, vol. 269, pp. 22014-22020, 1994.

[72] T. Joachims, "Text categorization with support vector machines: learning with many relevant features", *Proceedings of the European Conference on Machine Learning*, Springer, 1998.

[73] V. Vapnik, The *Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[74] Y.D. Cai, G.P. Zhou and K.C. Chou, "Support vector machines for predicting membrane protein types by using functional domain composition", *Biophysical Journal*, vol. 84, pp. 3257-3263, 2003.

[75] K.C. Chou and Y.D. Cai, "Using functional domain composition and support vector machines for prediction of protein subcellular location", *Journal of Biological Chemistry*, vol. 277, pp. 45765-45769, 2002.

[76] M. Wang, J. Yang, G.P. Liu, Z.J. Xu and K.C. Chou, "Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition", *Protein Engineering, Design, and Selection*, vol. 17, pp. 509-516, 2004.

[77] K.C. Chou and C.T. Zhang, "Review: Prediction of protein structural classes", *Critical Reviews in Biochemistry and Molecular Biology*, vol. 30, pp. 275-349, 1995.

[78] K.C. Chou and H.B. Shen, "Review: Recent progresses in protein subcellular location prediction", *Analytical Biochemistry*, vol. 370, pp. 1-16, 2007.

[79] Y.D. Cai and K.C. Chou, "Predicting membrane protein type by functional domain composition and pseudo amino acid composition", *Journal of Theoretical Biology*, vol. 238, pp. 395-400, 2006.

[80] Y.D. Cai, K.Y. Feng, W.C. Lu and K.C. Chou, "Using LogitBoost classifier to predict protein structural classes", *Journal of Theoretical Biology*, vol. 238, pp. 172-176, 2006.

[81] Y.D. Cai, G.P. Zhou and K.C. Chou, "Predicting enzyme family classes by hybridizing gene product composition and pseudo-amino acid composition", *Journal of Theoretical Biology*, vol. 234, pp. 145-149, 2005.

[82] Chen, L.X. Chen, X.Y. Zou and P.X. Cai, "Predicting protein structural class based on multi-features fusion", *Journal of Theoretical Biology*, vol. 253, pp. 388-392, 2008.

[83] Y.L. Chen and Q.Z. Li, "Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition", *Journal of Theoretical Biology*, vol. 248, pp. 377-381, 2007.

[84] Y.L. Chen and Q.Z. Li, "Prediction of the subcellular location of apoptosis proteins", *Journal of Theoretical Biology*, vol. 245, pp. 775-783, 2007.

[85] Y.S. Ding, T.L. Zhang and K.C. Chou, "Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network", *Protein and Peptide Letters*, vol. 14, pp. 811-815, 2007.

[86] P. Du and Y. Li, "Prediction of C-to-U RNA editing sites in plant mitochondria using both biochemical and evolutionary information", *Journal of Theoretical Biology*, vol. 253, pp. 579-589, 2008.

[87] X. Jiang, R. Wei, T.L. Zhang and Q. Gu, "Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy", *Protein and Peptide Letters*, vol. 15, pp. 392-396, 2008.

[88] Y. Jin, B. Niu, K.Y. Feng, W.C. Lu, Y.D. Cai and G.Z. Li, "Predicting subcellular localization with AdaBoost learne", *Protein and Peptide Letters*, vol. 15, pp. 286-289, 2008.

[89] F.M. Li and Q.Z. Li, "Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach", *Protein and Peptide Letters*, vol. 15, pp. 612-616, 2008.

[90] H. Lin, "The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition", *Journal of Theoretical Biology*, vol. 252, pp. 350-356, 2008.

[91] H. Lin, H. Ding, F.B. Guo, A.Y. Zhang and J. Huang, "Predicting subcellular localization of mycobacterial proteins by using chou's pseudo amino acid composition", *Protein and Peptide Letters*, vol. 15, pp. 739-744, 2008.

[92] Niu,, Y.D. Cai, W.C. Lu, G.Y. Zheng and K.C. Chou, "Predicting protein structural class with AdaBoost learner", *Protein and Peptide Letters*, vol. 13, pp. 489-492, 2006.

[93] B. Niu,, Y.H. Jin, K.Y. Feng, L. Liu, W.C. Lu, Y.D. Cai and G.Z. Li, "Predicting membrane protein types with bagging learner", *Protein and Peptide Letters*, vol. 15, pp. 590-594, 2008.

[94] H.B. Shen, J. Yang and K.C. Chou, "Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition", *Journal of Theoretical Biology*, vol. 240, pp. 9-13, 2006.

[95] M.G. Shi, D. Huang and X.L. Li, "A protein interaction network analysis for yeast integral membrane protein", *Protein and Peptide Letters*, vol. 15, pp. 692-699, 2008.

[96] S.Q. Wang, J. Yang, J. and K.C. Chou, "Using stacked generalization to predict membrane protein types based on pseudo amino acid composition", *Journal of Theoretical Biology*, vol. 242, pp. 941-946, 2006.

[97] G. Wu and S. Yan, "Prediction of mutations in H3N2 hemagglutinins of influenza A virus from north america based on different datasets", *Protein and Peptide Letters*, vol. 15, pp. 144-152, 2008.

[98] X. Xiao, P. Wang and K.C. Chou, "Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image", *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2008.06.016, 2008.

[99] G.Y. Zhang and B.S. Fang, "Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition", *Journal of Theoretical Biology*, vol. 253, pp. 310-315, 2008.

[100] G.Y. Zhang, H. Li and B.S. Fang, "Predicting lipase types by improved Chou's pseudo-amino acid composition", *Protein and Peptide Letters*, vol. 15, pp. 1132-1137, 2008.
[101] T.L. Zhang, Y.S. Ding and K.C. Chou, "Prediction protein structural classes with pseudo amino acid composition: approximate entropy and hydrophobicity pattern", *Journal of Theoretical Biology*, vol. 250, pp. 186-193, 2008.
[102] X.B. Zhou, C. Chen, Z.C. Li and X.Y. Zou, "Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes", *Journal of Theoretical Biology*, vol. 248, pp. 546-551, 2007.
[103] K.C. Chou and D.W. Elrod, "Protein subcellular location prediction", *Protein Engineering*, vol. 12, pp. 107-118, 1999.
[104] K.C. Chou, "Prediction of protein cellular attributes using pseudo amino acid composition", *PROTEINS: Structure, Function, and Genetics* (Erratum: ibid., 2002, vol. 44, 60), vol. 43, pp. 246-255, 2001.

Received: October 08, 2008 Revised: October 28, 2008 Accepted: November 03, 2008