Metabolic Pathway Modeling by Using the Nearest Neighbor Algorithm

Cai, Yu-Dong and Muldoon, Mark

2007

ISSN 1749-9097

Metabolic Pathway Modeling by Using the Nearest Neighbor Algorithm

Yu-Dong Cai and Mark Muldoon*
School of Mathematics, The University of Manchester, Alan Turing Building, Manchester 13 9PL, UK

Kuo-Chen Chou
Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, USA

*Corresponding to: M.Muldoon@Manchester.ac.uk (M. Muldoon)

Keywords: Budding yeast, *Saccharomyces cerevisiae,* Biochemical regulation, Enzyme control, Gene ontology, Microarray data, Chemical functional group

Running title: Metabolic Pathway Modeling

Abstract

A new computational approach was developed for modeling the metabolic pathways. The new approach is featured by combing the knowledge of gene ontology, microarray, and chemical functional group to formulate the enzyme-substrate/product couples in a 1,660 vector space. The nearest neighbor algorithm was used to perform the prediction of the networking relationship occurring in the metabolic pathways. The average overall success rate by jackknife cross-validation tests for the 79 metabolic pathways in the budding yeast system was over 94%, suggesting that the current approach might become a useful tool for studying metabolic pathways and many other networking-related areas.

I. Introduction

Metabolism (the Greek word for “change” or “overthrow”) is the biochemical modification of chemical compounds in living organisms and cells. It comprises a series of chemical reactions that occur in a cell and enable it to keep living, growing and dividing. Without metabolism we would not be able to survive.

Metabolic processes are generally classified as (1) anabolism and (2) catabolism (Voet et al., 2002). The former includes the biosynthesis of complex organic molecules, production of new cell components, usually through processes that demand energy and reducing power obtained from nutrient catabolism; while the latter, obtaining energy and reducing power from nutrients.

Metabolism usually consists of sequences of enzymatic steps, the so-called metabolic pathways. The cell metabolism includes all chemical processes in a cell, while the total metablism includes all biochemical processes of an organism.

The number of metabolic pathways is very large, reflecting the fact that “life is extremely complicated”. The most important metabolic pathways for humans are (Voet et al., 2002): (1) glycolysis – glucose oxidation for obtaining ATP; (2) citric acid cycle (Krebs’ cycle) (Krebs & Johnson, 1937)– acetyl-CoA oxidation for obtaining GTP and valuable intermediates; (3) oxidative phosphorylation – disposal of the electrons released by glycolysis and citric acid cycle (much of the energy released in this process can be stored as ATP); (4) pentose phosphate pathways – synthesis of pentoses and release of the reducing power needed for anabolic reactions; (5) urea cycle – disposal of NH_4^+ in less toxic forms; (6) fatty acid â -oxidation – fatty acids breakdown into acetyl-CoA for being used by the Krebs' cycle; (7) gluconeogenesis – glucose synthesis from smaller precursors for being used by the brain.

Metabolic pathways interact in a complex way in order to allow an adequate regulation. This interaction includes the enzymatic control and hormone control. In this study, we are focused on the enzyme control category, where metabolic pathway is the network linking various chemical reactions of compounds (substrates or products) catalyzed by enzymes.

The present study was devoted to establish a model for predicting the network relationship of enzymes and substrates/products in a living system.

II. Materials and Method

The data studied here were taken from ftp://ftp.genome.jp/pub/kegg/pathways/ (Kanehisa et al., 2004). Here, we are considering budding yeast *Saccharomyces cerevisiae*, which has 85 pathways (Table 1). Each pathway contains many reactions. For example, for the 1st pathway in Table 1, P00010, there are 26 different reactions catalyzed by various enzymes (Table 2). From Table 2 we can construct a positive and negative training datasets (Chou, 1993) for the pathway P00010.

As shown in Table 2, a same reaction may involve several different enzymes. The positive set consists of those pairs with each formed by one compound and one enzyme associated with the same reaction.

For example, for Reaction 1, the following 6 pairs (C05125, YBR221C), (C00068, YBR221C), (C00022, YBR221C), (C05125, YER178W), (C00068, YER178W), (C00022, YER178W) belong to the positive set.

For Reaction 2, the following 8 pairs (C00002, YAL038W), (C00022, YAL038W), (C00008, YAL038W), (C00074, YAL038W), (C00002, YOR347C), (C00022, YOR347C), (C00008, YOR347C), (C00074, YOR347C) belong to the positive set.

And so forth.

The negative set consists of those pairs in which the compound and enzyme are associated with different reactions. For example, (C05125, YAL038W) belongs to the negative set because C05125 is associated with Reaction 1 while YAL038W associated with Reaction 2. Similarly, (C05125, YOR347C), (C05125, YBR221C), (C05125, YER178W), and so forth, belong to the negative set as well.

Pairs in the positive set are termed networking pairs, and those in the negative set non-networking pair. Both the networking and non-networking pairs can be generally represented thru the following feature selections.

Each pair contains an enzyme and a compound. For the enzyme part, the GO (gene ontology) (Ashburner et al., 2000) and microarray data (http://bioinfo.mbb.yale.edu/expression/) were used to represent the sample of an enzyme. The details of how to use GO to represent a protein or enzyme were elaborated in many previous publications (Cai & Chou, 2004a; Cai & Chou, 2004b; Chou & Cai, 2004; Chou & Cai, 2005a; Chou & Cai, 2005b), and there is no need to repeat here. The only difference is that the number of GO-compress entries now was reduced to 1540 from 1930. This is because all the enzymes studied here were from yeast genes rather than entire gene universe. Here, in addition to GO, the microarray knowledge is used to represent the enzyme sample as well. According to the microarray data, each enzyme corresponds to 80 components which can be obtained from http://rana.lbl.gov/data/yeast/yeastall_public.txt.gz. For reader's convenience, these data are provided in Online Supplementary Materials A. By combining the GO and microarray data, an enzyme can be expressed as

$$\mathbf{E} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_{1540} \\ \grave{\imath}_1 \\ \grave{\imath}_2 \\ \vdots \\ \grave{\imath}_{80} \end{bmatrix} = \begin{bmatrix} g_1 & g_2 & \cdots & g_{1540} & \grave{\imath}_1 & \grave{\imath}_2 & \cdots & \grave{\imath}_{80} \end{bmatrix}^{\mathbf{T}} \qquad (1)$$

where g_i =1 if there is a hit corresponding to the ith ($i = 1,\ 2,\ \cdots,\ 1540$) GO number when using the program IPRSCAN (Apweiler et al., 2001) to search the InterPro functional domain database (release 6.1) for the enzyme; otherwise, g_i =0. **T** is the transpose operator to a matrix.

For the compound part, the 40 functional groups (Marchand-Geneste et al., 2002) (Table 3) were used to represent the sample of a compound (substrate or product); i.e.,

$$\mathbf{C} = \begin{bmatrix} c_1 & c_2 & \cdots & c_{40} \end{bmatrix}^{\mathbf{T}} \qquad (2)$$

where c_i is the occurrence number of the ith functional group of Table 3 in the compound concerned. Thus, the sample of an enzyme-compound pair can be expressed as a vector with 1540+80+40=1660 dimensions; i.e.,

$$\mathbf{EC} = \begin{bmatrix} g_1 & g_2 & \cdots & g_{1540} & \grave{\imath}_1 & \grave{\imath}_2 & \cdots & \grave{\imath}_{80} & c_1 & c_2 & \cdots & c_{40} \end{bmatrix}^{\mathbf{T}} \qquad (3)$$

With the above representation for the enzyme-compound pairs in both positive and negative sets for each of the pathways, we can use the nearest neighbor algorithm (Cai & Chou, 2003; Cover & Hart, 1967; Shen & Chou, 2005b) to perform the prediction.

III. Results and Discussion

In statistical prediction the independent dataset test, sub-sampling test, and jackknife test are the three cross-validation methods often used in literatures for examining the power of a predictor. Among these three, the jackknife test is deemed the most rigorous and objective, as indicated by a comprehensive discussion Chou, 1995 #27} and many follow-up papers (Feng et al., 2005; Feng, 2001; Liu et al., 2005; Pan et al., 2003; Shen & Chou, 2005a; Shen & Chou, 2005b; Shen et al., 2005; Wang et al., 2004; Wang et al., 2005; Xiao et al., 2005; Zhou, 1998; Zhou & Assa-Munt, 2001; Zhou & Doctor, 2003). Therefore, the jackknife cross validation was also used here to test the prediction quality.

Similar to the signal peptide prediction (Chou, 2001a; Chou, 2001b), the success rates for the positive set and negative set in the k th pathway of the budding yeast system are given by

$$\begin{cases} \Lambda_k^+ = \dfrac{N_k^+ - m_k^+}{N_k^+}, & \text{for positive set} \\ \\ \Lambda_k^- = \dfrac{N_k^- - m_k^-}{N_k^-}, & \text{for negative set} \end{cases} \tag{4}$$

where N_k^+ represents the total number of enzyme-compound networking (positive) pairs in the k th pathway, and m_k^+ is the number of positive pairs missed in prediction; N_k^- is the corresponding total number of negative pairs, and m_k^- is the number of negative pairs incorrectly predicted as positive pairs. The overall rate of correct prediction for the k th pathway is given by

$$\Lambda_k = \frac{\Lambda_k^+ N_k^+ + \Lambda_k^- N_k^-}{N_k^+ + N_k^-} = 1 - \frac{m_k^+ + m_k^-}{N_k^+ + N_k^-} \tag{5}$$

And the overall success rate for the entire budding yeast system is given by

$$\Lambda = \frac{\sum_{k=1}^{□}\left(\Lambda_k^+ N_k^+ + \Lambda_k^- N_k^-\right)}{\sum_{k=1}^{□}\left(N_k^+ + N_k^-\right)} = 1 - \frac{\sum_{k=1}^{□}\left(m_k^+ + m_k^-\right)}{\sum_{k=1}^{□}\left(N_k^+ + N_k^-\right)} \tag{6}$$

where □ is the total number of the metabolic pathways concerned in the budding yeast system. Of the 104 metabolic pathways for the budding yeast (Table 1), the data with statistical significance were obtained only for 79 pathways. Therefore, for the current study, $□ = 79$.

The predicted results by jackknife tests for each of the 79 pathways are given in Table 4, from which we can derive that the overall success rate for the entire 79 pathways is Λ=45135/47671=94.7%. The high overall success rate indicates that the current approach, which is featured by combing the knowledge of GO, microarray and chemical functional group to represent the enzyme-compound (substrate/product) pair samples, is very promising for predicting the reactions in the metabolic pathways.

Table 1. Codes of the 104 budding yeast metabolic pathways from KEGG

P00010	P00020	P00030	P00031	P00040	P00051	P00052	P00053
P00061	P00062	P00071	P00072	P00100	P00120	P00130	P00140
P00150	P00190	P00193	P00220	P00230	P00240	P00251	P00252
P00253	P00260	P00271	P00272	P00280	P00290	P00300	P00310
P00330	P00340	P00350	P00351	P00360	P00361	P00362	P00380
P00400	P00410	P00430	P00440	P00450	P00460	P00472	P00480
P00500	P00510	P00511	P00512	P00520	P00521	P00522	P00530
P00531	P00533	P00540	P00550	P00561	P00562	P00563	P00570
P00580	P00590	P00600	P00601	P00602	P00603	P00604	P00620
P00623	P00625	P00626	P00627	P00630	P00631	P00632	P00640
P00642	P00643	P00650	P00660	P00670	P00680	P00710	P00720
P00730	P00740	P00750	P00760	P00770	P00780	P00790	P00791
P00860	P00900	P00910	P00920	P00930	P00940	P00950	P00960

Table 2. Listing of 26 different reactions catalyzed by various enzymes for pathway P00010

Reaction	Compound A □ Compund B	Enzyme
1	C05125 <=> C00068+C00022	YBR221C
	C05125 <=> C00068+C00022	YER178W
2	C00002+C00022 <=> C00008+C00074	YAL038W
	C00002+C00022 <=> C00008+C00074	YOR347C
3	C00022 <=> C00024	YBR221C
	C00022 <=> C00024	YER178W
	C00022 <=> C00024	YFL018C
	C00022 <=> C00024	YNL071W
	C00022 <=> C00024	YPL017C
4	C00033 <=> C00024	YAL054C
	C00033 <=> C00024	YLR153C
5	C00631 <=> C00074	YGR254W
	C00631 <=> C00074	YHR174W
	C00631 <=> C00074	YMR323W
	C00631 <=> C00074	YOR393W
	C00631 <=> C00074	YPL281C
6	C00084 <=> C00033	YER073W
	C00084 <=> C00033	YMR169C
	C00084 <=> C00033	YMR170C
	C00084 <=> C00033	YOR374W
	C00084 <=> C00033	YPL061W
	C00084 <=> C00033	YER073W

	C00084 <=> C00033	YMR169C
	C00084 <=> C00033	YMR170C
	C00084 <=> C00033	YOR374W
	C00084 <=> C00033	YPL061W
7	C00469 <=> C00084	YBR145W
	C00469 <=> C00084	YDL168W
	C00469 <=> C00084	YGL256W
	C00469 <=> C00084	YMR083
	C00469 <=> C00084	YMR303C
	C00469 <=> C00084	YOL086C
8	C00084 <=> C05125	YDL080C
	C00084 <=> C05125	YGR087C
	C00084 <=> C05125	YLR044C
	C00084 <=> C05125	YLR134W
9	C00103 <=> C00668	YKL127W
	C00103 <=> C00668	YMR105C
10	C00118 <=> C00111	YDR050C
11	C00118 <=> C00236	YGR192C
	C00118 <=> C00236	YJL052W
	C00118 <=> C00236	YJR009C
12	C05378 <=> C00111+C00118	YKL060C
13	C00197 <=> C00236	YCR012W
14	C00631 <=> C00197	YDL021W
	C00631 <=> C00197	YKL152C

	C00631 <=> C00197	YOL056W
15	C00221 <=> C01172	YCL040W
	C00221 <=> C01172	YDR516C
	C00221 <=> C01172	YFR053C
	C00221 <=> C01172	YGL253W
16	C00267 <=> C00221	YBR019C
17	C00236 <=> C01159	YDL021W
	C00236 <=> C01159	YKL152C
	C00236 <=> C01159	YOL056W
18	C00579 <=> C00248	YFL018C
	C00579 <=> C00248	YPL017C
19	C00267 <=> C00668	YCL040W
	C00267 <=> C00668	YDR516C
	C00267 <=> C00668	YFR053C
	C00267 <=> C00668	YGL253W
20	C00024+C00579 <=> C01136	YNL071W
21	C00668 <=> C01172	YBR196C
22	C00668 <=> C05345	YBR196C
23	C05125+C00248 <=> C01136+C00068	YBR221C
	C05125+C00248 <=> C01136+C00068	YER178W
24	C01172 <=> C05345	YBR196C
25	C05345 <=> C05378	YGR240C
26	C05378 <=> C05345	YLR377C

Table 3. List of 40 chemical groups used for representing the samples of compounds

General feature	Key group				
Two dimensional structure	halogen	alcohol	aldehyde	amide	amine
	hydroxamic_acid	phosphorus	phosphorus_opo3	carboxylate	carboxylic_acid
	ester	ether	imine	ketone	methyl
	nitro	ar_alcohol	thiol	sulfonic_aci	sulfide
	sulfone	sulfonamide	sulfoxide	sulfo	halogen
	hacc	hdonor	neg_charge	pos_charge	hydrophobic
Cycle two dimensional structure	ar_5c_ring	ar_6c_ring	non_ar_5c_ring	non_ar_6c_ring	hetero_ar_5_ring
	hetero_ar_6_ring	hetero_non_ar_5_ring	hetero_non_ar_6_ring	five_ring	six_ring

Table 4. The successful rates for the 79 pathways (the numerators in columns 2, 3, and 4 represent the numbers of correct predictions for the positive, negative, and overall pairs for each of the pathways, respectively; while the denominators represent those of the corresponding total pairs concerned)

Index *k*	Pathway code	Positive (Λ_k^+)	Negative (Λ_k^-)	Overall (Λ_k)
1	P00010	91/111=0.819820	1039/1065=0.975587	1130/1176=0.960884
2	P00020	50/66=0.757576	392/398=0.984925	442/464=0.952586
3	P00030	53/65=0.815385	431/435=0.990805	484/500=0.968000
4	P00040	7/10=0.700000	30/30=1.000000	37/40=0.925000
5	P00051	98/189=0.518519	2109/2115=0.997163	2207/2304=0.957899
6	P00052	64/93=0.688172	642/651=0.986175	706/744=0.948925
7	P00053	9/14=0.642857	13/19=0.684211	22/33=0.666667
8	P00061	10/12=0.833333	2/4=0.500000	12/16=0.750000
9	P00062	16/18=0.888889	37/38=0.973684	53/56=0.946429
10	P00071	29/33=0.878788	249/252=0.988095	278/285=0.975439
11	P00100	30/36=0.833333	178/185=0.962162	208/221=0.941176

P00120	30/35=0.857143	192/196=0.979592	222/231=0.961039
P00130	110/125=0.8800	595/603=0.986733	705/728=0.968407
P00150	15/15=1.000000	76/76=1.000000	91/91=1.000000
P00190	42/42=1.000000	198/198=1.000000	240/240=1.000000
P00220	22/42=0.523810	346/357=0.969188	368/399=0.922306
P00230	231/325=0.710769	6092/6174=0.986718	6323/6499=0.972919
P00240	197/226=0.871681	2454/2472=0.992718	2651/2698=0.982580
P00251	33/71=0.464789	639/657=0.972603	672/728=0.923077
P00252	41/75=0.546667	612/627=0.976077	653/702=0.930199
P00260	43/71=0.605634	1074/1076=0.998141	1117/1147=0.973845
P00271	19/28=0.678571	131/137=0.956204	150/165=0.909091
P00272	18/23=0.782609	24/32=0.750000	42/55=0.763636

25	P00280	48/55=0.872727	196/198=0.989899	244/253=0.964427
26	P00290	38/47=0.808511	264/265=0.996226	302/312=0.967949
27	P00300	25/39=0.641026	244/250=0.976000	269/289=0.930796
28	P00310	32/52=0.615385	485/488=0.993852	517/540=0.957407
29	P00330	34/76=0.447368	974/1004=0.970120	1008/1080=0.933333
30	P00340	24/54=0.444444	532/540=0.985185	556/594=0.936027
31	P00350	59/78=0.756410	771/772=0.998705	830/850=0.976471
32	P00360	8/20=0.400000	74/76=0.973684	82/96=0.854167
33	P00361	13/22=0.590909	27/33=0.818182	40/55=0.727273
34	P00362	8/9=0.888889	17/18=0.944444	25/27=0.925926
35	P00380	65/100=0.650000	1105/1116=0.990143	1170/1216=0.962171
36	P00400	36/64=0.562500	420/442=0.950226	456/506=0.901186

P00410	18/19=0.947368	81/81=1.000000	99/100=0.990000
P00430	4/6=0.666667	2/4=0.500000	6/10=0.600000
P00440	6/20=0.300000	71/84=0.845238	77/104=0.740385
P00450	10/17=0.588235	84/85=0.988235	94/102=0.921569
P00460	18/29=0.620690	160/163=0.981595	178/192=0.927083
P00472	14/14=1.000000	2/7=0.285714	16/21=0.761905
P00480	15/27=0.555556	112/123=0.910569	127/150=0.846667
P00500	131/310=0.422581	2460/2502=0.983213	2591/2812=0.921408
P00510	99/144=0.687500	662/720=0.919444	761/864=0.880787
P00520	36/42=0.857143	59/62=0.951613	95/104=0.913462
P00521	11/14=0.785714	21/21=1.000000	32/35=0.914286
P00522	10/12=0.833333	4/6=0.666667	14/18=0.777778

P00530	23/35=0.657143	215/220=0.977273	238/255=0.933333
P00561	115/148=0.777027	1748/1772=0.986456	1863/1920=0.970313
P00562	213/225=0.946667	818/895=0.913966	1031/1120=0.920536
P00580	18/19=0.947368	15/16=0.937500	33/35=0.942857
P00590	4/4=1.000000	3/4=0.750000	7/8=0.875000
P00600	119/169=0.704142	626/678=0.923304	745/847=0.879575
P00603	33/49=0.673469	52/63=0.825397	85/112=0.758929
P00620	41/65=0.630769	471/479=0.983299	512/544=0.941176
P00626	2/2=1.000000	12/12=1.000000	14/14=1.000000
P00630	20/30=0.666667	93/102=0.911765	113/132=0.856061
P00632	178/251=0.709163	1701/1765=0.963739	1879/2016=0.932044
P00640	18/23=0.782609	131/133=0.984962	149/156=0.955128

63	P00643	4/9=0.444444	17/21=0.809524	21/30=0.700000
64	P00650	26/40=0.650000	484/487=0.993840	510/527=0.967742
65	P00670	28/52=0.538462	200/214=0.934579	228/266=0.857143
66	P00680	8/11=0.727273	37/37=1.000000	45/48=0.937500
67	P00710	56/69=0.811594	450/459=0.980392	506/528=0.958333
68	P00720	28/34=0.823529	124/126=0.984127	152/160=0.950000
69	P00730	5/13=0.384615	25/32=0.781250	30/45=0.666667
70	P00740	14/32=0.437500	115/122=0.942623	129/154=0.837662
71	P00750	36/36=1.000000	69/72=0.958333	105/108=0.972222
72	P00760	197/222=0.887387	873/878=0.994305	1070/1100=0.972727
73	P00770	18/21=0.857143	78/78=1.000000	96/99=0.969697
74	P00780	4/13=0.307692	34/41=0.829268	38/54=0.703704

P00790	42/68=0.617647	245/274=0.894161	287/342=0.839181
P00860	201/294=0.683673	3478/3567=0.975049	3679/3861=0.952862
P00900	16/18=0.888889	15/18=0.833333	31/36=0.861111
P00910	39/68=0.573529	792/802=0.987531	831/870=0.955172
P00920	3/10=0.300000	27/32=0.843750	30/42=0.714286
P00940	12/12=1.000000	12/12=1.000000	24/24=1.000000
P00950	9/12=0.750000	19/23=0.826087	28/35=0.800000
P00970	32/112=0.285714	2192/2219=0.987832	2224/2331=0.954097

References

Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D. R., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, L., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N. J., Oinn, T. M., Pagni, M., Servant, F., Sigrist, C. J. A. & Zdobnov, E. M. (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research*, **29**, 37-40.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25-29.

Cai, Y. D. & Chou, K. C. (2003). Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. *Biochem Biophys Res Comm*, **305**, 407-411.

Cai, Y. D. & Chou, K. C. (2004a). Predicting 22 protein localizations in budding yeast. *Biochem. Biophys. Res. Comm.*, **323**, 425-428.

Cai, Y. D. & Chou, K. C. (2004b). Predicting subcellular localization of proteins in a hybridization space. *Bioinformatics*, **20**, 1151-1156.

Chou, K. C. (1993). A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. *Journal of Biological Chemistry*, **268**, 16938-16948.

Chou, K. C. (2001a). Prediction of protein signal sequences and their cleavage sites. *PROTEINS: Structure, Function, and Genetics*, **42**, 136-139.

Chou, K. C. (2001b). Using subsite coupling to predict signal peptides. *Protein Engineering*, **14**, 75-79.

Chou, K. C. & Cai, Y. D. (2004). Prediction of protein subcellular locations by GO-FunD-PseAA predicor. *Biochemical and Biophysical Research Communications*, **320**, 1236-1239.

Chou, K. C. & Cai, Y. D. (2005a). Predicting protein localization in budding yeast. *Bioinformatics*, **21**, 944-950.

Chou, K. C. & Cai, Y. D. (2005b). Using GO-PseAA predictor to identify membrane proteins and their types. *Biochem. Biophys. Res. Comm.*, **327**, 845-847.

Cover, T. M. & Hart, P. E. (1967). Nearest neighbour pattern classification. *IEEE Transaction on Information Theory*, **IT-13**, 21-27.

Feng, K. Y., Cai, Y. D. & Chou, K. C. (2005). Boosting classifier for predicting protein domain structural class. *Biochemical & Biophysical Research Communications*, **334**, 213-217.

Feng, Z. P. (2001). Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. *Biopolymers*, **58**, 491-499.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. (2004). The KEGG resources for deciphering the genome. *Nucleic Acids Res.*, **32**, D277-D280.

Krebs, H. A. & Johnson, W. A. (1937). The role of citric acid in intermediate metabolism in animal tissues. *Enzymologia*, **4**, 148-156.

Liu, H., Wang, M. & Chou, K. C. (2005). Low-frequency Fourier spectrum for predicting membrane protein types. *Biochem Biophys Res Commun*, **336**, 737-739.

Marchand-Geneste, N., Watson, K. A., Alsberg, B. K. & King, R. D. (2002). New approach to pharmacophore mapping and QSAR analysis using inductive logic programming. Application to thermolysin inhibitors and glycogen phosphorylase B inhibitors. *J Med Chem*, **45**, 399-409.

Pan, Y. X., Zhang, Z. Z., Guo, Z. M., Feng, G. Y., Huang, Z. D. & He, L. (2003). Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. *Journal of Protein Chemistry*, **22**, 395-402.

Shen, H. & Chou, K. C. (2005a). Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. *Biochemical & Biophysical Research Communications*, **334**, 288-292.

Shen, H. B. & Chou, K. C. (2005b). Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. *Biochem. Biophys. Res. Comm.*, **337**, 752-756.

Shen, H. P., Yang, J., Liu, X. J. & Chou, K. C. (2005). Using supervised fuzzy clustering to predict protein structural classes. *Biochem. Biophys. Res. Commun.*, **334**, 577-581.

Voet, D., Voet, J. G. & Pratt, C. W. (2002). *Fundamentals of Biochemistry, Chap.13*, John Wiley & Sons, New York.

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

Wang, M., Yang, J., Xu, Z. J. & Chou, K. C. (2005). SLLE for predicting membrane protein types. *Journal of Theoretical Biology*, **232**, 7-15.

Xiao, X., Shao, S., Ding, Y., Huang, Z., Huang, Y. & Chou, K. C. (2005). Using complexity measure factor to predict protein subcellular location. *Amino Acids*, **28**, 57-61.

Zhou, G. P. (1998). An intriguing controversy over protein structural class prediction. *Journal of Protein Chemistry*, **17**, 729-738.

Zhou, G. P. & Assa-Munt, N. (2001). Some insights into protein structural class prediction. *PROTEINS: Structure, Function, and Genetics*, **44**, 57-59.

Zhou, G. P. & Doctor, K. (2003). Subcellular location prediction of apoptosis proteins. *PROTEINS: Structure, Function, and Genetics*, **50**, 44-48.