EXCLI Journal 2006;5:55-65 – ISSN 1611-2156
Received: 1. May 2006, accepted: 9. May 2006, published 19. May 2006

Original article:

Predicting networking couples for metabolic pathways of Arabidopsis

Kuo-Chen Chou[1,*], Yu-Dong Cai[1,2,3], Wei-Zhu Zhong[1]

[1]Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, USA (*corresponding author e-mail: kchou@san.rr.com)

[2]CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Biological Sciences, Chinese Academy of Sciences,320 Yue Yang Road, Shanghai , China

[3]Department of Mathematics, University of Manchester, P.O. Box 88, Sackville Street, Manchester M60 1QD, UK

ABSTRACT

Given an enzyme-compound couple, how can we identify whether it belongs to a networking couple or non-networking couple? This is very important for investigating the metabolic pathways. To address this problem, a novel approach was developed that is featured by using the knowledge of gene ontology (GO), chemical functional group (FunG), and pseudo amino acid composition (PseAA) to represent the samples of enzyme-compound couples. Two basic identifiers were formulated: one is called "GO-FunG", and the other, "PseAA-FunG". The prediction was operated by fusing these two basic identifiers into one. As a showcase, the metabolic pathways were investigated for Arabidopsis thaliana, a small flowering plant widely used as a model organism for studies of the cellular and molecular biology of flowering plants. The average overall success rate via the jackknife cross-validation tests for the 72 metabolic pathways in the Arabidopsis system was over 95%, suggesting that the current approach might become a very useful tool for studying metabolic pathways and many other problems in the cellular networking related areas.

Keywords: Arabidopsis thaliana; Enzyme control regulation; Gene ontology; Chemical functional group; Pseudo amino acid composition; Cellular networking; Metabolic pathway; System biology

INTRODUCTION

A living organism must not be a closed, equilibrium system but an open, steady-state one. To maintain its order, and hence life, in a universe bent on maximizing disorder, a continuous influx of free energy is indispensable. Metabolism, the Greek word for "change" or "overthrow", is the overall process thru which living systems acquire and utilize the free energy they need for performing various functions to keep their life. Metabolism comprises a set of sophistigated metabolic pathways, which are series of consecutive enzymatic reactions that produce specific products, and thru which the steady state in a living system is maintained. The cell metabolism covers all chemical processes in a cell, while the total metablism, all biochemical processes of an organism. Because a living system utilizes many metabolites (i.e., reactants, intermediates, and products), it has many metabolic pathways.

Metabolic pathways are generally classified into two categories: (a) anabolism (biosynthesis) and (b) catabolism (degradation) (Voet et al., 2002). The former includes the process of biosynthesizing complex organic molecules and producing new cell components; while the latter, the process of obtaining energy and reducing power from nutrients.

One of the important characteristics of metabolic pathways is that they are highly exergonic, i.e., having large negative free energy changes, which provides them with distinct direction to complete their reactions. Accordingly, if two metabolites are metabolically interconvertible, the pathway from the first to the second must differ from the pathway from the second back to the first. Also, in order to exert control on the flux of metabolites thru a metabolic pathway, it is necessary to use enzymatic control to realize various regulations, such as regulating glycolysis, gluconeogenesis, citric acid cycle (Krebs' cycle) (Krebs & Johnson, 1937), urea cycle, glycogen metabolism, fatty acids metabolism, and pentose phosphate pathway (Voet et al., 2002).

Table 1: Codes of the 102 metabolic pathways of Arabidopsis thaliana

P00010	P00020	P00030	P00031	P00040	P00051	P00052	P00053
P00061	P00071	P00072	P00100	P00120	P00130	P00150	P00190
P00193	P00195	P00220	P00230	P00240	P00251	P00252	P00260
P00271	P00272	P00280	P00290	P00300	P00310	P00330	P00340
P00350	P00351	P00360	P00361	P00362	P00380	P00400	P00401
P00410	P00430	P00440	P00450	P00460	P00480	P00500	P00510
P00511	P00512	P00513	P00520	P00521	P00522	P00530	P00531
P00540	P00550	P00561	P00562	P00564	P00590	P00600	P00601
P00602	P00603	P00604	P00620	P00624	P00626	P00628	P00630
P00632	P00640	P00642	P00643	P00650	P00670	P00680	P00710
P00720	P00730	P00740	P00750	P00760	P00770	P00780	P00790
P00860	P00900	P00901	P00902	P00903	P00904	P00910	P00920
P00930	P00940	P00941	P00950	P00960	P00970		

Knowledge of metabolic pathways is indispensable for understanding a living system at the level of molecular networks. However, owing to the extreme complexity of the problem, it is both time-consuming and costly to determine the metabolic pathways and the network interactions therein purely by means of biochemical experiments even for a very simple living system. Besides, for those whose metabolic pathways are known, the knowledge might be still not complete, meaning that some network interactions between enzymes and substrates/products might be missing. In view of this, it would be highly desired to develop an automated method, or a complementary tool, for fast predicting the network relationship of enzymes and substrates/products in a living system. The present study was initiated in an attempt to explore this problem.

MATERIALS AND METHOD

Here, let us consider Arabidopsis thaliana, a small flowering plant belonging to a member of the mustard (Brassicaceae) family, which includes cultivated species such as cabbage and radish.

Arabidopsis is not of major agronomic significance, but it offers important advantages for basic research in genetics and molecular biology, and hence is widely used as a model organism in plant biology. Its metabolic pathways were taken from ftp://ftp.genome.jp/pub/kegg/pathways/. There are 102 pathways (Table 1). Each pathway contains many reactions. The enzymes and compounds (ligands) involved in these reactions were taken from http://mips.gsf.de/proj/thal/db/index.html and ftp://ftp.genome.jp/pub/kegg/ligand/, respectively. For example, for the 1st pathway in Table 1, P00010, there are 18 different reactions catalyzed by various enzymes listed in Appendix A, from which we can construct a positive and negative training datasets (Chou, 1993; Elhammer et al., 1993; Poorman et al., 1991) for the pathway P00010.

As shown in Appendix A, a same reaction may involve several different enzymes. The positive training set S^{+} consists of those couples with each formed by one compound and one enzyme associated with the same reaction. For example, for Reaction 1, the following 21 couples (C05125, AT1G01090), (C05125, AT1G24180), (C05125, AT1G30120), (C05125, AT1G59900), (C05125, AT2G34590), (C05125, AT3G48560), (C05125, AT5G50850), (C00068, AT1G01090), (C00068, AT1G24180), (C00068, AT1G30120), (C00068, AT1G59900), (C00068, AT2G34590), (C00068, AT3G48560), (C00068, AT5G50850), (C00022, AT1G01090), (C00022, AT1G24180), (C00022, AT1G30120), (C00022, AT1G59900), (C00022, AT2G34590), (C00022, AT3G48560), and (C00022, AT5G50850) belong to the positive set S^{+}. For Reaction 2, there are 40 couples, such as (C00002, AT3G04050), (C00002, AT3G25690), and (C00074, AT5G63680), belonging to the positive set. And so forth.

The negative training set S^{-} consists of those pairs in which the compound and enzyme are associated with different reactions. For example, (C05125, AT3G04050) belongs to the negative training set because C05125 is associated with Reaction 1 while AT3G04050 associated with Reaction 2. Similarly, (C05125, AT3G25960), (C05125, AT3G52990), (C05125, AT3G55650), and so forth, belong to the negative set S^{-} as well.

Couples in the positive set S^{+} are termed “networking couples”, and those in the negative set S^{-} “non-networking couples”. Both the networking and non-networking couples can be generally represented thru the following feature selections.

Each couple contains an enzyme and a compound. For the enzyme part, the GO (gene ontology) (Ashburner et al., 2000) and the pseudo amino acid composition (PseAA) were used to represent the sample of an enzyme. The GO database is very useful in representing the samples of proteins by grasping their core features (Camon et al., 2004; Harris et al., 2004; Lee et al., 2005), while the PseAA allows us to incorporate a considerable amount of sequence-order effects into a discrete model (Chou, 2001). The details of how to use GO-PseAA to represent the sample of protein or enzyme were elaborated in previous publications (Chou & Cai, 2004). The only difference is that the GO

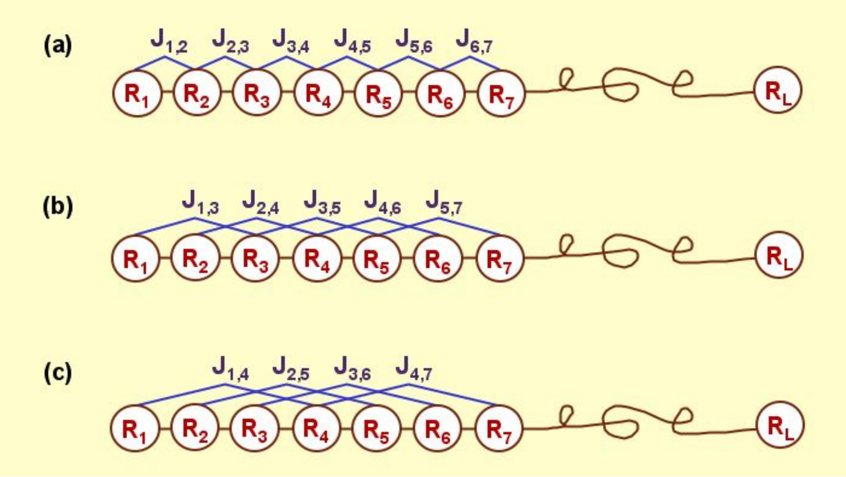

Figure 1: A schematic drawing to show **(a)** the 1st-tier, **(b)** the 2nd-tier, and **(c)** the 3rd-tier sequence-order-correlation mode along a protein sequence, where R_1 represents the amino acid residue at the sequence position 1, R_2 at position 2, and so forth, and the coupling factors $J_{i,j}$ are given by eq.3 of (Chou, 2001). Panel **(a)** reflects the correlation mode between all the most contiguous residues, panel **(b)** that between all the 2nd most contiguous residues, and panel **(c)** that between all the 3rd most contiguous residues. Adapted from (Chou, 2001) with permission.

information was now downloaded from Genemerge (version 2003) at http://genemerge.bioteam.net/download.html because all the enzymes studied here are from Arabidopsis thaliana genes rather than the entire gene universe. The number of GO_compress entries thus obtained was reduced to 663 from 1930 as in the case of (Chou & Cai, 2004). The following steps were followed to represent enzyme-compound couple.

Step 1. Each of the 663 GO numbers in GO_compress will serve as a base to define a 663D (dimensional) vector for a given enzyme $\mathbf{E}$, as formulated below

$$\mathbf{E} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_i \\ \vdots \\ g_{663} \end{bmatrix}, \qquad (1)$$

where $g_i = 1$ if there is a hit corresponding to the *i*th ($i = 1, 2, \ldots, 663$) GO number when searching the GO_compress entries for the enzyme $\mathbf{E}$; otherwise, $g_i = 0$, as treated in the case for defining the functional domain composition (Chou & Cai, 2002).

Step 2. If no hit whatsoever is found for any of the 663 GO numbers, the enzyme $\mathbf{E}$ will correspond to a naught vector. Under such a circumstance, the enzyme should be instead defined in the $(20+\lambda)$D PseAA space (Chou, 2001), as formulated below

$$\mathbf{E} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_{20} \\ p_{20+1} \\ \vdots \\ p_{20+\lambda} \end{bmatrix}, \qquad (2)$$

where $p_1, p_2, \cdots, p_{20}$ represent the 20 components of the classical amino acid composition (Chou, 1995; Nakashima et al., 1986; Zhou, 1998), while p_{20+1} is the first-tier sequence order correlation factor, p_{20+2} the second-tier sequence order correlation factor, and so forth (Fig.1). It is the additional λ components that incorporate some sequence order effects into the representation of the enzyme. For different datasets, λ usually has different optimal value (Chou, 2001). For the current study, the optimal value of λ is 37. Given a enzyme, the (20+37)=57 PseAA components in eq.2 can be easily derived by following the procedures as described in the paper (Chou, 2001) that has originally introduced the concept of PseAA. Thus, the enzyme that corresponds to a naught vector in the 663D GO space (eq.1) can always be explicitly defined in the 57D PseAA space (eq.2).

For the compound part, the 34 functional groups (FunG) were used (cf. Table 3 of Marchand-Geneste et al., 2002) to represent the sample of a compound (substrate or product); i.e.,

$$\mathbf{C} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_{34} \end{bmatrix} = \begin{bmatrix} c_1 & c_2 & \cdots & c_{34} \end{bmatrix}^{\mathbf{T}} \qquad (3)$$

where c_i is the occurrence number of the *i*th functional group in the compound concerned, and $\mathbf{T}$ is transpose operator to a matrix. Thus, the sample of an enzyme-compound pair can be expressed as a vector with 663+34=697 dimensions if the enzyme is expressed in the 663D GO system (eq.1) or 57+34=91 dimensions if the enzyme expressed in the 57D PseAA system (eq.2); i.e.,

$$£^{\mathrm{EC}} = \begin{cases} \begin{bmatrix} g_1 \; g_2 \; \cdots \cdots \; g_{663} \, c_1 \; c_2 \; \cdots \; c_{34} \end{bmatrix}^{\mathbf{T}}, \\ \qquad \text{in GO-FunG system} \\ \begin{bmatrix} p_1 \; p_2 \; \cdots \; p_{57} \, c_1 \; c_2 \; \cdots \; c_{34} \end{bmatrix}^{\mathbf{T}}, \\ \qquad \text{in PseAA-FunG system} \end{cases} \qquad (4)$$

where $£^{\mathrm{EC}}$ represent an enzyme-compound couple. The prediction was performed with the ISort (Intimate Sorting) predictor, which can be briefed below. Suppose there are N enzyme-compound couples $\left(\pounds_1^{\mathrm{EC}},\pounds_2^{\mathrm{EC}},\mathrm{L}\ ,\pounds_N^{\mathrm{EC}}\right)$ which have been classified into categories 1, 2, …, μ. Now, for a query enzyme-compound couple $\pounds^{\mathrm{EC}}$, how can we predict which category it belongs to? To deal with this problem, let us define the following scale to measure the similarity between $\pounds^{\mathrm{EC}}$ and $\pounds_i^{\mathrm{EC}}$ (i = 1, 2, …, N)

$$\Psi(\pounds^{\mathbf{EC}},\pounds_i^{\mathbf{EC}}) = \frac{\pounds^{\mathbf{EC}}\cdot\pounds_i^{\mathbf{EC}}}{\left\|\pounds^{\mathbf{EC}}\right\|\left\|\pounds_i^{\mathbf{EC}}\right\|}, \qquad (5)$$
$$(i=1,\ 2,\ \mathrm{L}\ ,\ N)$$

where $\pounds^{\mathbf{EC}}\cdot\pounds_i^{\mathbf{EC}}$ is the dot product of vectors $\pounds^{\mathbf{EC}}$ and $\pounds_i^{\mathbf{EC}}$, and $\left\|\pounds^{\mathbf{EC}}\right\|$ and $\left\|\pounds_i^{\mathbf{EC}}\right\|$ their modulus, respectively. Obviously, when $\pounds^{\mathbf{EC}}\equiv\pounds_i^{\mathbf{EC}}$, we have $\Psi(\pounds^{\mathbf{EC}},\pounds_i^{\mathbf{EC}})=1$, meaning they have perfect or 100% similarity. Generally speaking, the similarity is within the range of 0 and 1; *i.e.,* $0\leq\Psi(\pounds^{\mathbf{EC}},\pounds_i^{\mathbf{EC}})\leq 1$. Accordingly, the ISort predictor can be formulated as follows. If the similarity between $\pounds^{\mathbf{EC}}$ and $\pounds_k^{\mathbf{EC}}$ $(k=1,\ 2,\ \mathrm{L}\ ,\text{ or } N)$ is the highest; i.e.

$$\Psi(\pounds^{\mathbf{EC}},\pounds_k^{\mathbf{EC}}) = \mathbf{Max}\left\{\Psi(\pounds^{\mathbf{EC}},\pounds_1^{\mathbf{EC}}),\ \Psi(\pounds^{\mathbf{EC}},\pounds_2^{\mathbf{EC}})\ \mathrm{L}\ ,\Psi(\pounds^{\mathbf{EC}},\pounds_N^{\mathbf{EC}})\right\} \qquad (6)$$

where the operator **Max** means taking the maximum one among those in the brackets, then the query couple $\pounds^{\mathbf{EC}}$ is predicted belonging to the same category as of $\pounds_k^{\mathbf{EC}}$. If there is a tie, the query protein may not be uniquely determined and will be randomly assigned among those with a tie, but cases like that rarely occur. The ISort classifier is particularly useful for the situation when the distributions of the samples are unknown.

To make the operation consistent, the following rule must be observed during the course of computation: the predictor's parameters should be derived based on all those enzyme-compound couples in the training set that can be meaningfully defined in the same space as of the query enzyme-compound couple. Accordingly, the current ISort predictor actually consists of two sub-predictors: **(1)** the ISort-697D predictor that operates in the 697D GO-FunG space (the 1st equation of eq.4), and **(2)** the ISort-91D predictor that operates in the 91D PseAA-FunG space (the 2nd equation of eq.4). The whole predictor is called GO-PseAA-FunG hybridization predictor, or just GO-PseAA-FunG predictor, which was operated by fusing the two sub-predictors according to the following "flowchart". If the enzyme of the query enzyme-compound couple was meaningfully defined in the 663D GO space (eq.1), then the ISort-697D GO-FunG predictor was used to predict its attribute; if the enzyme in the 663D GO space is a naught vector and hence must be redefined in the 57D PseAA space (eq.2), then the ISort-91D PseAA-FunG predictor was used to predict the attribute of the query enzyme-compound couple.

The success rates for the positive set and negative set in the k th pathway of the Arabidopsis system are given by

$$\begin{cases} \Lambda_k^+ = \dfrac{N_k^+ - m_k^+}{N_k^+}, & \text{for positive set} \\ \Lambda_k^- = \dfrac{N_k^- - m_k^-}{N_k^-}, & \text{for negative set} \end{cases} \qquad (7)$$

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

$$\Lambda_k = \frac{\Lambda_k^+ N_k^+ + \Lambda_k^- N_k^-}{N_k^+ + N_k^-} = 1 - \frac{m_k^+ + m_k^-}{N_k^+ + N_k^-} \qquad (8)$$

And the overall success rate for the entire Arabidopsis system is given by

$$\Lambda = \frac{\sum_{k=1}^{¥}\left(\Lambda_k^+ N_k^+ + \Lambda_k^- N_k^-\right)}{\sum_{k=1}^{¥}\left(N_k^+ + N_k^-\right)} = 1 - \frac{\sum_{k=1}^{¥}\left(m_k^+ + m_k^-\right)}{\sum_{k=1}^{¥}\left(N_k^+ + N_k^-\right)} \qquad (9)$$

where ¥ is the total number of the metabolic pathways concerned in the Arabidopsis system. Of the 102 metabolic pathways for the Arabidopsis system (Table 1), the data with statistical significance were obtained only for 72 pathways (Appendix B). Therefore, for the current study, $¥ = 72$.

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. See a monograph by Mardia et al. (Mardia et al., 1979) for the mathematical principle and a review (Chou & Zhang, 1995) for a comprehensive discussion about this. More and more investigators have adopted the jackknife test to examine the power of various predictors (Feng, 2001; Feng, 2002; Luo et al., 2002; Pan et al., 2003; Zhou, 1998; Zhou & Assa-Munt, 2001; Zhou & Doctor, 2003). Here, the jackknife cross validation was also used to test the prediction quality.

The computation was carried out in a Silicon Graphics IRIS Indigo workstation (Elan 4000). According to the search procedures as described in Section II, we obtained the following results. In the 72 pathways of Arabidopsis system there are 26,755 possible enzyme-compound couples, of which 3,771 belong to the positive set S^+, and 22,984 belong to the negative set S^-. Furthermore, it was found according to Steps $1-4$ of Section II that, of the 3,771 networking couples in S^+, 3,391 got hits in the GO system and hence were defined in the 697D GO-FunG space (the 1st equation of eq.4), and the remaining 380 couples were defined in the 91D PseAA-FunG space (the 2nd equation of eq.4). Also, of the 22,984 non-networking couples in S^-, 20,203 got hits in the GO system and hence were defined in the 697D GO-FunG space (the 1st equation of eq.4), and the remaining 2,781 couples were defined in the 91D PseAA-FunG space (the 2nd equation of eq.4).

The predicted results by jackknife tests for each of the 72 pathways are given in Appendix B, from which we can derive that the overall success rate for the entire 72 pathways is Λ =25607/26755=95.7%. The high overall success rate indicates that the current approach, which is featured by combing the knowledge of GO, PseAA and chemical functional group to represent the enzyme-compound (substrate/product) couple samples, is very promising for predicting the reactions in the metabolic pathways. The present work just represents the seeds of investigating a very important but extremely complicated problem in system biology by means of computational approach. Of course, substantially more work is needed and is currently under way in our lab.

CONCLUSION

Knowledge of metabolic pathways is very important for understanding a living system at the level of molecular networks. During the process of studying a metabolic pathway, a key problem is how to identify a query enzyme-compound couple belongs to a networking couple or non-networking couple. It is both expensive and time-consuming to characterize all the query couples purely by means of biochemical experiments even for a very simple living system. Therefore, it would be of great help to develop an automated method as a complementary tool. The method developed here is featured by fusing two identifiers: one is based on the gene ontology (GO) and chemical functional group (FunG); while the other, the pseudo amino acid composition (PseAA) and FunG. The results thus obtained are quite promising, implying that the fusing approach might become a useful vehicle for studying metabolic pathways and many other system biology related problems.

Appendix A: Listing of 18 different reactions catalyzed by various enzymes for pathway P00010

Reaction	Compound A Compund B	Enzyme
1	C05125 <=> C00068 + C00022	AT1G01090
	C05125 <=> C00068 + C00022	AT1G24180
	C05125 <=> C00068 + C00022	AT1G30120
	C05125 <=> C00068 + C00022	AT1G59900
	C05125 <=> C00068 + C00022	AT2G34590
	C05125 <=> C00068 + C00022	AT3G48560
	C05125 <=> C00068 + C00022	AT5G50850
2	C00002 + C00022 <=> C00008 + C00074	AT3G04050
	C00002 + C00022 <=> C00008 + C00074	AT3G25960
	C00002 + C00022 <=> C00008 + C00074	AT3G52990
	C00002 + C00022 <=> C00008 + C00074	AT3G55650
	C00002 + C00022 <=> C00008 + C00074	AT3G55810
	C00002 + C00022 <=> C00008 + C00074	AT4G26390
	C00002 + C00022 <=> C00008 + C00074	AT5G08570
	C00002 + C00022 <=> C00008 + C00074	AT5G52920
	C00002 + C00022 <=> C00008 + C00074	AT5G56350
	C00002 + C00022 <=> C00008 + C00074	AT5G63680
3	C00022 <=> C00024	AT1G01090
	C00022 <=> C00024	AT1G24180
	C00022 <=> C00024	AT1G30120
	C00022 <=> C00024	AT1G34430
	C00022 <=> C00024	AT1G48030
	C00022 <=> C00024	AT1G54220
	C00022 <=> C00024	AT1G59900
	C00022 <=> C00024	AT2G34590
	C00022 <=> C00024	AT3G13930
	C00022 <=> C00024	AT3G16950
	C00022 <=> C00024	AT3G17240
	C00022 <=> C00024	AT3G25860
	C00022 <=> C00024	AT3G52200
	C00022 <=> C00024	AT5G50850
4	C00631 <=> C00074	AT1G74030
	C00631 <=> C00074	AT2G36530
5	C00084 <=> C05125	AT4G33070
	C00084 <=> C05125	AT5G01320
	C00084 <=> C05125	AT5G01330
	C00084 <=> C05125	AT5G54960
6	C00103 <=> C00668	AT1G23190
	C00103 <=> C00668	AT1G70730
7	C00103 <=> C00668	AT5G51820
	C00118 <=> C00111	AT2G21170
	C00118 <=> C00111	AT3G55440
	C00118 <=> C00236	AT1G12900
	C00118 <=> C00236	AT1G13440
	C00118 <=> C00236	AT1G16300
	C00118 <=> C00236	AT1G42970
	C00118 <=> C00236	AT1G79530
	C00118 <=> C00236	AT3G04120
	C00118 <=> C00236	AT3G26650

8	C05378 <=> C00111 + C00118	AT2G01140
	C05378 <=> C00111 + C00118	AT2G21330
	C05378 <=> C00111 + C00118	AT2G36460
	C05378 <=> C00111 + C00118	AT3G52930
	C05378 <=> C00111 + C00118	AT4G26520
	C05378 <=> C00111 + C00118	AT4G26530
	C05378 <=> C00111 + C00118	AT4G38970
	C05378 <=> C00111 + C00118	AT5G03690
9	C00197 <=> C00236	AT1G56190
	C00197 <=> C00236	AT1G79550
	C00197 <=> C00236	AT3G12780
10	C00221 <=> C01172	AT1G47840
	C00221 <=> C01172	AT2G19860
	C00221 <=> C01172	AT3G20040
	C00221 <=> C01172	AT4G37840
11	C00267 <=> C00221	AT3G17940
	C00267 <=> C00221	AT3G47800
	C00267 <=> C00221	AT5G15140
12	C00579 <=> C00248	AT1G48030
	C00579 <=> C00248	AT3G16950
	C00579 <=> C00248	AT3G17240
13	C00267 <=> C00668	AT1G47840
	C00267 <=> C00668	AT2G19860
	C00267 <=> C00668	AT3G20040
	C00267 <=> C00668	AT4G37840
14	C00024 + C00579 <=> C01136	AT1G34430
	C00024 + C00579 <=> C01136	AT1G54220
	C00024 + C00579 <=> C01136	AT3G13930
	C00024 + C00579 <=> C01136	AT3G25860
	C00024 + C00579 <=> C01136	AT3G52200
15	C00668 <=> C01172	AT4G24620
	C00668 <=> C01172	AT5G42740
	C00668 <=> C05345	AT4G24620
	C00668 <=> C05345	AT5G42740
16	C05125 + C00248 <=> C01136 + C00068	AT1G01090
	C05125 + C00248 <=> C01136 + C00068	AT1G24180
	C05125 + C00248 <=> C01136 + C00068	AT1G30120
	C05125 + C00248 <=> C01136 + C00068	AT1G59900
	C05125 + C00248 <=> C01136 + C00068	AT2G34590
	C05125 + C00248 <=> C01136 + C00068	AT5G50850
17	C01172 <=> C05345	AT4G24620
	C01172 <=> C05345	AT5G42740
18	C05378 <=> C05345	AT1G43670
	C05378 <=> C05345	AT3G54050

Appendix B: The successful rates for the 72 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	195/205=0.951220	1216/1225=0.992653	1411/1430=0.986713
2	P00020	59/77=0.766234	430/435=0.988506	489/512=0.955078
3	P00030	80/92=0.869565	479/484=0.989669	559/576=0.970486
4	P00040	5/12=0.416667	12/18=0.666667	17/30=0.566667
5	P00051	74/84=0.880952	264/276=0.956522	338/360=0.938889
6	P00052	74/92=0.804348	444/454=0.977974	518/546=0.948718
7	P00053	15/16=0.937500	4/8=0.500000	19/24=0.791667
8	P00061	11/12=0.916667	20/21=0.952381	31/33=0.939394
9	P00071	30/32=0.937500	44/45=0.977778	74/77=0.961039
10	P00100	73/87=0.839080	566/578=0.979239	639/665=0.960902
11	P00130	14/19=0.736842	47/51=0.921569	61/70=0.871429
12	P00190	34/36=0.944444	96/96=1.000000	130/132=0.984848
13	P00220	34/51=0.666667	352/363=0.969697	386/414=0.932367
14	P00230	270/345=0.782609	4123/4191=0.983775	4393/4536=0.968474
15	P00240	168/193=0.870466	1627/1643=0.990262	1795/1836=0.977669
16	P00251	34/68=0.500000	553/570=0.970175	587/638=0.920063
17	P00252	43/63=0.682540	460/466=0.987124	503/529=0.950851
18	P00260	68/87=0.781609	950/957=0.992685	1018/1044=0.975096
19	P00271	27/43=0.627907	183/191=0.958115	210/234=0.897436
20	P00272	46/58=0.793103	94/102=0.921569	140/160=0.875000
21	P00280	106/114=0.929825	506/510=0.992157	612/624=0.980769
22	P00290	105/112=0.937500	667/668=0.998503	772/780=0.989744
23	P00300	24/30=0.800000	102/102=1.000000	126/132=0.954545
24	P00310	19/26=0.730769	61/65=0.938462	80/91=0.879121
25	P00330	51/66=0.772727	692/702=0.985755	743/768=0.967448
26	P00340	19/23=0.826087	96/97=0.989691	115/120=0.958333
27	P00350	26/29=0.896552	134/136=0.985294	160/165=0.969697
28	P00360	18/20=0.900000	49/50=0.980000	67/70=0.957143
29	P00361	2/4=0.500000	2/4=0.500000	4/8=0.500000
30	P00380	39/44=0.886364	296/298=0.993289	335/342=0.979532
31	P00400	51/80=0.637500	674/695=0.969784	725/775=0.935484
32	P00410	23/26=0.884615	152/154=0.987013	175/180=0.972222
33	P00450	42/46=0.913043	118/122=0.967213	160/168=0.952381
34	P00460	43/45=0.955556	154/155=0.993548	197/200=0.985000
35	P00480	52/63=0.825397	263/278=0.946043	315/341=0.923754
36	P00500	113/139=0.812950	903/917=0.984733	1016/1056=0.962121
37	P00510	5/16=0.312500	82/94=0.872340	87/110=0.790909
38	P00520	4/8=0.500000	14/16=0.875000	18/24=0.750000
39	P00521	20/26=0.769231	76/78=0.974359	96/104=0.923077
40	P00522	17/20=0.850000	48/50=0.960000	65/70=0.928571
41	P00530	15/21=0.714286	74/79=0.936709	89/100=0.890000
42	P00540	2/2=1.000000	2/3=0.666667	4/5=0.800000
43	P00550	24/24=1.000000	20/20=1.000000	44/44=1.000000
44	P00561	31/42=0.738095	326/332=0.981928	357/374=0.954545
45	P00562	9/14=0.642857	36/40=0.900000	45/54=0.833333
46	P00600	23/24=0.958333	53/57=0.929825	76/81=0.938272

47	P00603	3/4=0.750000	2/2=1.000000	5/6=0.833333
48	P00620	88/115=0.765217	393/413=0.951574	481/528=0.910985
49	P00630	32/38=0.842105	155/157=0.987261	187/195=0.958974
50	P00632	11/11=1.000000	29/31=0.935484	40/42=0.952381
51	P00640	23/32=0.718750	139/144=0.965278	162/176=0.920455
52	P00643	3/3=1.000000	0/2=0.000000	3/5=0.600000
53	P00650	37/50=0.740000	240/244=0.983607	277/294=0.942177
54	P00670	32/64=0.500000	190/208=0.913462	222/272=0.816176
55	P00710	147/164=0.896341	957/970=0.986598	1104/1134=0.973545
56	P00720	19/22=0.863636	32/33=0.969697	51/55=0.927273
57	P00730	7/8=0.875000	13/16=0.812500	20/24=0.833333
58	P00740	17/20=0.850000	26/29=0.896552	43/49=0.877551
59	P00750	12/14=0.857143	27/31=0.870968	39/45=0.866667
60	P00760	2/4=0.500000	2/4=0.500000	4/8=0.500000
61	P00770	30/30=1.000000	126/126=1.000000	156/156=1.000000
62	P00780	4/4=1.000000	4/4=1.000000	8/8=1.000000
63	P00790	16/24=0.666667	29/36=0.805556	45/60=0.750000
64	P00860	25/41=0.609756	348/358=0.972067	373/399=0.934837
65	P00900	68/70=0.971429	203/205=0.990244	271/275=0.985455
66	P00901	11/11=1.000000	3/5=0.600000	14/16=0.875000
67	P00904	13/20=0.650000	71/79=0.898734	84/99=0.848485
68	P00910	82/104=0.788462	870/880=0.988636	952/984=0.967480
69	P00920	25/34=0.735294	107/110=0.972727	132/144=0.916667
70	P00940	123/132=0.931818	985/990=0.994949	1108/1122=0.987522
71	P00950	5/6=0.833333	2/3=0.666667	7/9=0.777778
72	P00960	10/10=1.000000	8/8=1.000000	18/18=1.000000

REFERENCES

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, and Sherlock G. Gene ontology: tool for the unification of biology. *Nature Genetics* 2000; 25, 25-29.

Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, and Apweiler R. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res* 2004; 32, D262-6.

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

Chou KC. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. *Proteins: Structure, Function & Genetics* 1995; 21, 319-344.

Chou KC. Prediction of protein cellular attributes using pseudo amino acid composition. *PROTEINS: Structure, Function, and Genetics (Erratum: ibid., 2001, Vol.44, 60)* 2001; 43, 246-255.

Chou KC, and Cai YD. Using functional domain composition and support vector machines for prediction of protein subcellular location. *Journal of Biological Chemistry* 2002; 277, 45765-45769.

Chou KC, and Cai YD. Predicting enzyme family class in a hybridization space. *Protein Science* 2004; 13, 2857-2863.

Chou KC, and Zhang CT. Review: Prediction of protein structural classes. *Critical Reviews in Biochemistry and Molecular Biology* 1995; 30, 275-349.

Elhammer AP, Poorman RA, Brown E, Maggiora LL, Hoogerheide JG, and Kezdy FJ. The specificity of UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferase as inferred from a database of in vivo substrates and from the in vitro glycosylation of proteins and peptides. *Journal of Biological Chemistry* 1993; 268, 10029-10038.

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

Feng ZP. An overview on predicting the subcellular location of a protein. *In Silico Biol* 2002; 2, 291-303.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, and White R. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004; 32, D258-61.

Krebs HA, and Johnson WA. The role of citric acid in intermediate metabolism in animal tissues. *Enzymologia* 1937; 4, 148-156.

Lee V, Camon E, Dimmer E, Barrell D, and Apweiler R. Who tangos with GOA?-Use of Gene Ontology Annotation (GOA) for biological interpretation of '-omics' data and for validation of automatic annotation tools. *In Silico Biol* 2005; 5, 5-8.

Luo RY, Feng ZP, and Liu JK. Prediction of protein strctural class by amino acid and polypeptide composition. *Eur. J. Biochem.* 2002; 269, 4219-4225.

Marchand-Geneste N, Watson KA, Alsberg BK, and King RD. New approach to pharmacophore mapping and QSAR analysis using inductive logic programming. Application to thermolysin inhibitors and glycogen phosphorylase B inhibitors. *J Med Chem* 2002; 45, 399-409.

Mardia KV, Kent JT, and Bibby JM. (1979). *Multivariate Analysis: Chapter 11 Discriminant Analysis; Chapter 12 Multivariate analysis of variance; Chapter 13 cluster analysis (pp. 322-381),* Academic Press, London.

Nakashima H, Nishikawa K, and Ooi T. The folding type of a protein is relevant to the amino acid composition. *J. Biochem* 1986; 99, 152-162.

Pan YX, Zhang ZZ, Guo ZM, Feng GY, Huang ZD, and He L. Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. *Journal of Protein Chemistry* 2003; 22, 395-402.

Poorman RA, Tomasselli AG, Heinrikson RL, and Kezdy FJ. A cumulative specificity model for proteases from human immunodeficiency virus types 1 and 2, inferred from statistical analysis of an extended substrate data base. *Journal of Biological Chemistry* 1991; 266, 14554-14561.

Voet D, Voet JG, and Pratt CW. (2002). *Fundamentals of Biochemistry, Chap.13,* John Wiley & Sons, New York.

Zhou GP. An intriguing controversy over protein structural class prediction. *Journal of Protein Chemistry* 1998; 17, 729-738.

Zhou GP, and Assa-Munt N. Some insights into protein structural class prediction. *PROTEINS: Structure, Function, and Genetics* 2001; 44, 57-59.

Zhou GP, and Doctor K. Subcellular location prediction of apoptosis proteins. *PROTEINS: Structure, Function, and Genetics* 2003; 50, 44-48.