Some Remarks on Predicting Multi-Label Attributes in Molecular Biosystems

ARTICLE *in* MOLECULAR BIOSYSTEMS · MARCH 2013

Impact Factor: 3.18 · DOI: 10.1039/c3mb25555g · Source: PubMed

1 AUTHOR:

Kuo-Chen Chou
Gordon Life Science Institute

509 PUBLICATIONS **30,060** CITATIONS

SEE PROFILE

Molecular BioSystems

RSC Publishing

REVIEW

Some remarks on predicting multi-label attributes in molecular biosystems

Cite this: *Mol. BioSyst.*, 2013, **9**, 1092

Kuo-Chen Chou*[ab]

Received 4th December 2012,
Accepted 20th February 2013

DOI: 10.1039/c3mb25555g

www.rsc.org/molecularbiosystems

Many molecular biosystems and biomedical systems belong to the multi-label systems in which each of their constituent molecules possesses one or more than one function or feature, and hence needs one or more than one label to indicate its attribute(s). With the avalanche of biological sequences generated in the post genomic age, it is highly desirable to develop computational methods to timely and reliably identify their various kinds of attributes. Compared with the single-label systems, the multi-label systems are much more complicated and difficult to deal with. The current mini review focuses on the recent progresses in this area from both conceptual aspects and detailed mathematical formulations.

I. Introduction

Many molecules in biosystems or biomedical systems possess multiplex features.

It has been observed that an increasing number of proteins have multiple locations in a cell,[1] meaning that they can simultaneously reside at, or move between, two or more different subcellular locations. Proteins with multiple location sites or dynamic features of this kind are particularly interesting because they may have some unique biological functions worthy of our special notice.[2] For example, of the 3106 human proteins investigated in ref. 3, 2580 occur in one location; 480 in two locations; 43 in three locations, and 3 in 4 locations. Of the 5048 animal proteins investigated in ref. 4, 2284 occur in one subcellular location, 1740 in two locations, 510 in three locations, 368 in four locations, 111 in five locations, 20 in six locations, 9 in seven locations, and 6 in eight locations.

According to the ATC (Anatomical Therapeutic Chemical) classification system recommended by the World Health Organization, drugs are generally classified into the following 14 main ATC-groups based on their therapeutic, pharmacological and chemical properties (http://www.whocc.no/atc/structure_and_principles/): (1) alimentary tract and metabolism; (2) blood and blood forming organs; (3) cardiovascular system; (4) dermatologicals; (5) genitourinary system and sex hormones; (6) systemic hormonal preparations, excluding sex hormones and insulins; (7) antiinfectives for systemic use; (8) antineoplastic and immunomodulating agents; (9) musculoskeletal system; (10) nervous system; (11) antiparasitic products, insecticides and repellents; (12) respiratory system; (13) sensory organs; (14) various. Some drugs may belong to more than one main ATC-class. For example of the 3883 drugs investigated in ref. 5, 3295 occur in one class, 370 in two classes, 110 in three classes, 37 in four classes, 27 in five classes, and 44 in six classes.

Antimicrobial peptides (AMPs), also called host defense peptides, are an evolutionarily conserved component of the innate immune response and are found among all classes of life. According to their special functions, AMPs are generally

[a] *Gordon Life Science Institute, 53 South Cottage Road, Belmont, Massachusetts 02478, USA. E-mail: kcchou@gordonlifescience.org; Tel: +1 858-380-4623*
[b] *King Abdulaziz University, Jeddah, Saudi Arabia*

Kuo-Chen Chou

Dr Kuo-Chen Chou is the chief scientist of Gordon Life Science Institute. He is also an Advisory Professor of several universities. Professor Chou has published over 450 papers in the fields of bioinformatics, computer-aided drug design, protein-structural prediction, low-frequency internal motion of protein and DNA and its biological functions, diffusion-controlled reactions of enzymes, as well as graphic rules in enzyme kinetics and other biological systems. As of March 2013, Professor Kuo-Chen Chou's publications have been cited more than 24 018 times with an h-index of 84, according to ISI Science Citation Index (Web of Science). For more information about Professor Kuo-Chen Chou, visit http://gordonlifescience.org/members/kcchou/; http://www.scirp.org/kcchou/; or http://www.researcherid.com/rid/A-8340-2009.

classified into the following ten types:[6] (1) antibacterial peptides; (2) anticancer/tumor peptides; (3) antifungal peptides; (4) anti-HIV peptides; (5) antiviral peptides; (6) antiparasital peptides; (7) anti-protist peptides; (8) AMPs with chemotactic activity; (9) insecticidal peptides; (10) spermicidal peptides. Some AMPs may belong to two or more functional types. For example, of the 878 AMPs investigated in ref. 7, 454 belong to one functional type, 296 to two different types, 85 to three types, 30 to four types, and 13 to five types.

The aforementioned examples actually belong to the so-called "multi-label system" because each of the constituent members therein may need one or more than one label to indicate its attribute(s). With the avalanche of biological sequences generated in the post genomic age, it is highly desirable to develop computational methods to timely and reliably identify their various kinds of attributes. Compared with single-label systems, multi-label systems are much more complicated and difficult to deal with. Therefore, many approaches used to deal with the former would be no longer valid for the latter. The present review focuses on what kind of special attention is needed when dealing with the multi-label systems.

II. Number of virtual samples

In developing a statistical prediction method for a given attribute, the first important thing is to construct a benchmark dataset $\mathbb{S}$ according to its possible classification; *i.e.*,

$$\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \mathbb{S}_2 \cup \mathbb{S}_3 \cup \mathbb{S}_4 \cup \mathbb{S}_5 \ldots \cup \mathbb{S}_M \quad (1)$$

where $\mathbb{S}_1$ represents the subset for category 1 of the attribute, $\mathbb{S}_2$ for category 2, and so forth; while $\cup$ represents the symbol for "union" in the set theory, and M the total number of different categories for the attribute concerned. For example, when the attribute concerned was about the subcellular localization of eukaryotic proteins as investigated in ref. 8, M would be 22 as illustrated in Fig. 1; when the attribute concerned was about the subcellular localization of animal proteins as investigated in ref. 4, M would be 20; when the attribute concerned was about the ATC groups of drugs investigated in ref. 5, M would be 14; and so forth.

Because a sample in a multi-label system may have two or more labels to annotate its classification, it is useful to introduce the concept of "virtual sample" as briefed below. If a sample is marked with two different labels, it will be counted as two virtual samples; if marked with three different labels, it will be counted as three virtual samples; and so forth. Thus, the number of total virtual samples, $N(\mathrm{vir})$, can be expressed as

$$N(\mathrm{vir}) = \sum_{k=1}^{N} n^{\mathrm{L}}(k) \quad (2)$$

where $n^{\mathrm{L}}(k)$ is the number of different labels "attached" on the kth sample in $\mathbb{S}$. On the other hand, the number of total virtual samples can also be expressed as[9]

$$N(\mathrm{vir}) = N + \sum_{m=1}^{M} (m-1)n(m) \quad (3)$$

where N represents the number of total different samples, $n(1)$ the number of samples with one single label, $n(2)$ the number of samples with two labels, and $n(m)$ the number of samples with $m(\leq M)$ labels; while M is the number of total different classes investigated.

As we can see from eqn (3), the number of the total virtual samples is always greater than that of the total number of different samples. When, and only when, all the samples have a single label, *i.e.* $n(m) = 0$ when $m \geq 2$, can the two be the same.

III. Multiplicity degree

To quantitatively reflect the extent of multiplicity for a benchmark dataset $\mathbb{S}$, let us introduce the multiplicity degree $\mathrm{MD}(\mathbb{S})$, which is actually the quotient of the number of total virtual samples $N(\mathrm{vir})$ divided by the number of total different samples N. Thus, according to eqn (2) and (3), it can be calculated by

$$\mathrm{MD}(\mathbb{S}) = \frac{N(\mathrm{vir})}{N} = 1 + \frac{\sum_{m=1}^{M} (m-1)n(m)}{N} = \frac{\sum_{k=1}^{N} n^{\mathrm{L}}(k)}{N} \quad (4)$$

where all the symbols have exactly the same meanings as in eqn (2) and (3). As we can see from eqn (4), when all the samples in $\mathbb{S}$ have only one single label, we have $\mathrm{MD}(\mathbb{S}) = 1$; when all the samples in $\mathbb{S}$ have two labels, $\mathrm{MD}(\mathbb{S}) = 2$; and so forth. Therefore, the closer to 1 the multiplicity degree is, the fewer the number of samples in $\mathbb{S}$ that have multi-labels.

For example, the multiplicity degree for the system investigated in ref. 3 was 3681/3106 = 1.1851; that for the system investigated in ref. 5 was 4912/3883 = 1.2650; that for the system investigated in ref. 7 was 1486/878 = 1.6925; and that for the system investigated in ref. 4 was 9522/5048 = 1.8922.

IV. Prediction of multi-label attributes

Various classifiers were developed for predicting the multi-label attributes in different molecular biosystems. For example, for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins, the *iLoc-Euk* classifier[8] was proposed based on the "multi-label K-nearest neighbor" algorithm. For predicting the subcellular locations of animal proteins with both single and multiple sites, the *iLoc-Animal* classifier[4] was proposed based on the "accumulation-label K-nearest neighbor" algorithm. For identifying the functional types of antimicrobial peptides (AMPs), the *iAMP-2L* classifier[7] was proposed based on the "fuzzy K-nearest neighbor" algorithm.[7] Here, let us give a brief introduction about the *iLoc-Animal* classifier[4] through which we can see how a multi-label classifier works.

To develop a classifier for a biological system, one of the keys is to formulate the biological samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the attribute to be identified. Such a mathematical expression is usually a feature vector. If the biological samples are proteins or peptides, their feature vector can be

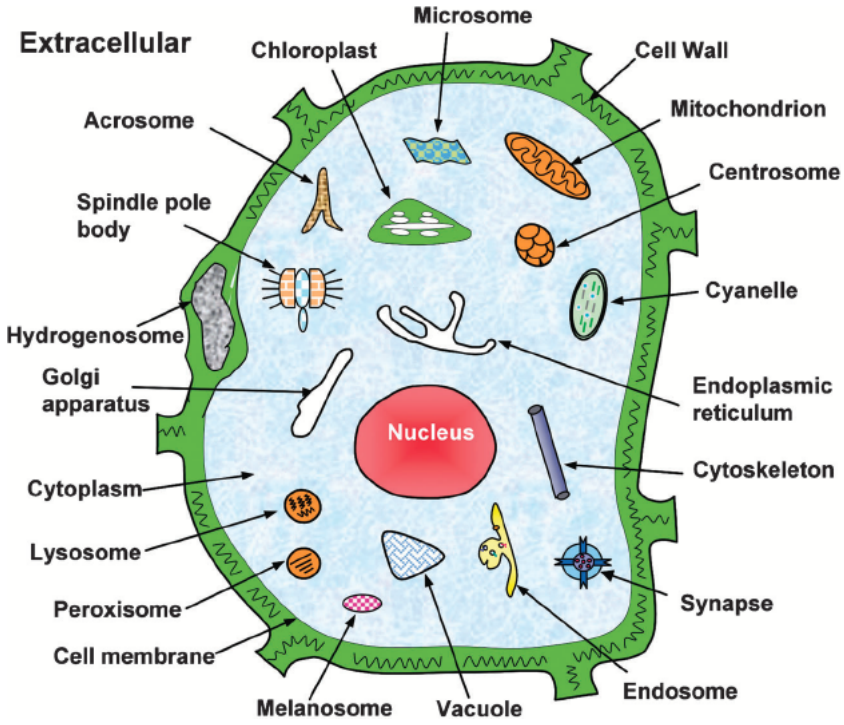

Fig. 1 Schematic illustration to show the 22 subcellular locations of eukaryotic proteins: (1) acrosome, (2) cell membrane, (3) cell wall, (4) centriole, (5) chloroplast, (6) cyanelle, (7) cytoplasm, (8) cytoskeleton, (9) endoplasmic reticulum, (10) endosome, (11) extracellular, (12) Golgi apparatus, (13) hydrogenosome, (14) lysosome, (15) melanosome, (16) microsome (17) mitochondria, (18) nucleus, (19) peroxisome, (20) spindle pole body, (21) synapse, and (22) vacuole. Adapted from ref. 39 with permission.

expressed by a general form of pseudo amino acid composition (PseAAC),[10] as given by[11]

$$\mathbf{P} = [\psi_1\ \psi_2 \cdots \psi_u \cdots \psi_\Omega]^{\mathbf{T}} \quad (5)$$

where $\mathbf{P}$ represents a protein or peptide sample, $\mathbf{T}$ is the transposing operator, the subscript Ω is an integer, and its value as well as the components ψ_u (u = 1,2,$\cdots$,Ω) will depend on how to extract the desired information from the amino acid sequence of $\mathbf{P}$. For different objects or targets, the components ψ_u may have different implications (see, *e.g.*, ref. 12–29). Actually, the concept of PseAAC and eqn (5) were not only limited to protein and peptide sequences. Recently, they were also extended to represent the feature vectors of DNA and nucleotides,[30,31] as well as other biological samples (see, *e.g.*, ref. 32 and 33).

For the *iLoc-Animal* classifier, the general form of PseAAC (eqn (5)) was defined by a combination of a gene ontology approach and a sequential evolution approach.[4] Once the feature vector has been defined to represent the protein samples, the next thing we need to consider is how to introduce an algorithm to operate the statistical prediction, as described below.

Suppose the mth subset $\mathbb{S}_m$ of $\mathbb{S}$ (eqn (1)) contains N_m proteins, and $\mathbf{P}(m,j)$ is the jth protein in that subset, and its feature vector is defined through the same procedure as that of $\mathbf{P}$ and hence also has the form of eqn (5). Thus, the similarity between $\mathbf{P}$ and $\mathbf{P}(m,j)$ can be defined by

$$D\{\mathbf{P},\mathbf{P}(m,j)\} = ||\mathbf{P} - \mathbf{P}(m,j)|| \quad (6)$$

where $||\mathbf{P} - \mathbf{P}(m,j)||$ represents the module of the vector difference between $\mathbf{P}$ and $\mathbf{P}(m,j)$ in the Euclidean space. According to eqn (6), when $\mathbf{P} \equiv \mathbf{P}(m,j)$ we have $D\{\mathbf{P},\mathbf{P}(m,j)\} = 0$, indicating that the distance between these two protein sequences is zero and hence they have perfect or 100% similarity.

Suppose $\mathbf{P}_1^\#, \mathbf{P}_2^\#, \cdots, \mathbf{P}_K^\#$ are the K nearest neighbor proteins to the protein $\mathbf{P}$ and they form a set denoted by $\mathbb{S}_K^{\mathbf{P}}$, which is a subset of $\mathbb{S}$ (*cf.* eqn (1)); *i.e.*,$\mathbb{S}_K^{\mathbf{P}} \subseteq \mathbb{S}$.

Based on the K nearest neighbor proteins in $\mathbb{S}_K^{\mathbf{P}}$, the "accumulation-label scale" is defined as given by

$$\mathbb{Q}(\mathbf{P},K) = \{\rho_1^K \rho_2^K \cdots \rho_m^K \cdots \rho_M^K\} \quad (7)$$

where

$$\rho_m = \frac{\sum_{i=1}^K \delta(\mathbf{P}_i^\#, m)}{\mathbb{N}_K^\#} \quad (m = 1,2,\cdots,M) \quad (8)$$

where

$$\delta(\mathbf{P}_i^\#, m) = \begin{cases} 1, & \text{if } \mathbf{P}_i^\# \text{ is labeled with the } m\text{th location} \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

and

$$\mathbb{N}_K^\# = \sum_{m=1}^M \sum_{i=1}^K \delta(\mathbf{P}_i^\#, m) \quad (10)$$

Note that $\mathbb{N}_K^\# \geq K$ because a protein may have more than one subcellular location label in the current system.

Now, for a query protein $\mathbf{P}$, its subcellular location(s) will be predicted according to the following steps.

Step 1. The number of different subcellular locations it belongs to will be determined by its nearest neighbor protein in $\mathbb{S}$ (*cf.* eqn (1)). For example, suppose $\mathbf{P}^{\#}$ is the nearest protein to $\mathbf{P}$ in $\mathbb{S}$. If $\mathbf{P}^{\#}$ has only one location label, *i.e.*, belongs to only one subcellular location, then $\mathbf{P}$ will be predicted belonging to only one location; if $\mathbf{P}^{\#}$ has two subcellular locations, then $\mathbf{P}$ will also have two locations; and so forth. Therefore, in general we have

$$\mathfrak{M}(\mathbf{P}) = \mathfrak{M}(\mathbf{P}^{\#}) \tag{11}$$

where $\mathfrak{M}(\mathbf{P}^{\#})$ is an integer ($\leq M$) representing the number of different subcellular locations to which $\mathbf{P}^{\#}$ belongs, and $\mathfrak{M}(\mathbf{P})$ represents the number of different subcellular locations to which $\mathbf{P}$ belongs.

Step 2. The actual location site(s) where $\mathbf{P}$ resides will not be determined by the location site(s) of $\mathbf{P}^{\#}$, but by the element(s) of the accumulation-label scale of eqn (7) that has (have) the highest score(s), as can be expressed by $\{\ell\}$, the subscript(s) of eqn (1). For example, if $\mathbf{P}$ is found belonging to only one location in Step 1, *i.e.*, $\mathfrak{M}(\mathbf{P}) = 1$, and the highest score in eqn (7) is ρ_3^K, then $\mathbf{P}$ will be predicted as $\{\ell\} = \{3\}$ meaning that it belongs to $\mathbb{S}_3$ or resides at the ''Centriole'' (*cf.* Table 1). If $\mathbf{P}$ is found belonging to four locations, *i.e.*, $\mathfrak{M}(\mathbf{P}) = 4$, and the first four highest scores in eqn (7) are ρ_1^K, ρ_8^K, ρ_{15}^K and ρ_{20}^K, then $\mathbf{P}$ will be predicted as $\{\ell\} = \{1, 8, 15, 20\}$ meaning that it belongs to $\mathbb{S}_1$, $\mathbb{S}_8$, $\mathbb{S}_{15}$ and $\mathbb{S}_{20}$ or resides simultaneously at the ''Acrosome'', ''Endoplasmic reticulum'', ''Microsome'', and ''Synapse'' (*cf.* Table 1), and so forth. In other words, the actual predicted subcellular location(s) for $\mathbf{P}$ can be formulated as

$$\{\ell\} = \text{Max} \rhd_{\textbf{Sub}}^{\mathfrak{M}(\mathbf{P})} \{\rho_1^K \rho_2^K \cdots \rho_m^K \cdots \rho_M^K\} (\mathfrak{M}(\mathbf{P}) \leq M) \tag{12}$$

Table 1 The multi-label system $\mathbb{S}$ contains 5048 animal proteins classified into 20 subcellular location sites. Of the 5048 proteins, 2284 occur in one subcellular location, 1740 in two locations, 510 in three locations, 368 in four locations, 111 in five locations, 20 in six locations, 9 in seven locations, 6 in eight locations, and none in nine or more locations. Reproduced from ref. 4 with permission

Subset	Subcellular location or label	Number of proteins
$\mathbb{S}_1$	Acrosome	87
$\mathbb{S}_2$	Cell membrane	1096
$\mathbb{S}_3$	Centriole	75
$\mathbb{S}_4$	Centrosome	243
$\mathbb{S}_5$	Cell cortex	108
$\mathbb{S}_6$	Cytoplasm	2170
$\mathbb{S}_7$	Cytoskeleton	729
$\mathbb{S}_8$	Endoplasmic reticulum	541
$\mathbb{S}_9$	Endosome	185
$\mathbb{S}_{10}$	Extracellular space	105
$\mathbb{S}_{11}$	Golgi apparatus	413
$\mathbb{S}_{12}$	Lysosome	136
$\mathbb{S}_{13}$	Mitochondrion	595
$\mathbb{S}_{14}$	Melanosome	49
$\mathbb{S}_{15}$	Microsome	71
$\mathbb{S}_{16}$	Nucleus	1458
$\mathbb{S}_{17}$	Peroxisome	81
$\mathbb{S}_{18}$	Plasma membrane	1096
$\mathbb{S}_{19}$	Spindle	159
$\mathbb{S}_{20}$	Synapse	155
Total different virtual or locative proteins		9552
Total different proteins		5048

where the operator ''Max $\rhd_{\textbf{Sub}}^{\mathfrak{M}(\mathbf{P})}$'' means identifying the $\mathfrak{M}(\mathbf{P})$ highest scores for the elements in the brackets right after them, followed by taking their $\mathfrak{M}(\mathbf{P})$ subscripts. The value for the parameter K in eqn (12) will be determined by optimizing the overall jackknife[11] success rate on the benchmark dataset $\mathbb{S}$.

V. Metrics for measuring the prediction quality of a multi-label system

For a multi-label system $\mathbb{S}$ consisting of N different samples, suppose $\mathbb{L}$ is the label set that contains all the possible labels for $\mathbb{S}$. Obviously, we have

$$||\mathbb{L}|| = M \tag{13}$$

where $||\ ||$ is the operator acting on the set therein to count the number of its elements, and M has the same meaning as in eqn (1). Thus, the kth sample $\mathbf{P}_k$ and its label(s) can be expressed by

$$\{\mathbf{P}_k, \mathbb{L}_k\}\ (k = 1, 2, \cdots, N) \tag{14}$$

where $\mathbb{L}_k$ is the subset that contains all the labels for the kth sample $\mathbf{P}_k$. Obviously, we have

$$\mathbb{L}_1 \cup \mathbb{L}_2 \cup \cdots \cup \mathbb{L}_N \subseteq \mathbb{L} = \{\ell_1, \ell_2, \cdots \ell_M\} \tag{15}$$

where ℓ_1 is the 1st label in $\mathbb{S}$, ℓ_2 is the 2nd label, and so forth, while the symbol $\subseteq$ means ''being the subset of''.

Suppose $\mathbb{L}_k^*$ represents the subset that contains all the predicted labels for the kth sample $\mathbf{P}_k$. Note that the predicted labels of a protein sample are usually not the same as its real labels. Only when the prediction is perfectly correct, will the two subsets $\mathbb{L}_k$ and $\mathbb{L}_k^*$ be the same (for further explanation, see Fig. 2 and its legend). Thus, we can have the following five metrics to measure the prediction quality for the multi-label system:

$$\begin{cases} \text{Aiming} = \dfrac{1}{N}\sum_{k=1}^{N}\left(\dfrac{\|\mathbb{L}_k \cap \mathbb{L}_k^*\|}{\|\mathbb{L}_k^*\|}\right) \\ \text{Coverage} = \dfrac{1}{N}\sum_{k=1}^{N}\left(\dfrac{\|\mathbb{L}_k \cap \mathbb{L}_k^*\|}{\|\mathbb{L}_k\|}\right) \\ \text{Accuracy} = \dfrac{1}{N}\sum_{k=1}^{N}\left(\dfrac{\|\mathbb{L}_k \cap \mathbb{L}_k^*\|}{\|\mathbb{L}_k \cup \mathbb{L}_k^*\|}\right) \\ \text{Absolute-True} = \dfrac{1}{N}\sum_{k=1}^{N}\Delta(\mathbb{L}_k, \mathbb{L}_k^*) \\ \text{Absolute-False} = \dfrac{1}{N}\sum_{k=1}^{N}\left(\dfrac{\|\mathbb{L}_k \cup \mathbb{L}_k^*\| - \|\mathbb{L}_k \cap \mathbb{L}_k^*\|}{M}\right) \end{cases} \tag{16}$$

where N and M have the same meanings as in eqn (3), $\cup$ is the symbol of union in the set theory, $\cap$ is the intersection symbol, $||\ ||$ has the same meaning as in eqn (13), and

$$\Delta(\mathbb{L}_k, \mathbb{L}_k^*) = \begin{cases} 1, & \text{if all the labels in } \mathbb{L}_k \text{ are identical to those in } \mathbb{L}_k^* \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

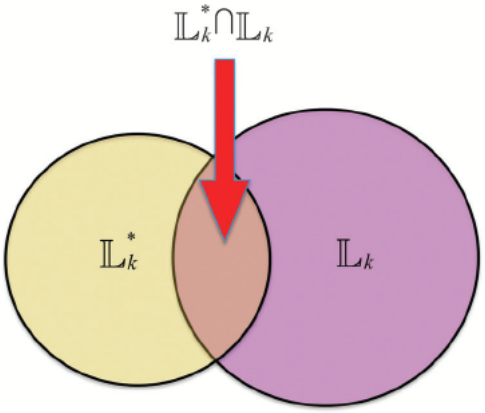

Fig. 2 Schematic drawing to show the meanings of some set theory symbols used in eqn (16): $\mathbb{L}_k$ represents the subset that contains all the label(s) for the kth sample; $\mathbb{L}_k^*$, the subset that contains all the predicted label(s) for the kth sample. Union of the subsets $\mathbb{L}_k$ and $\mathbb{L}_k^*$, denoted $\mathbb{L}_k \cup \mathbb{L}_k^*$, is the set of all labels that are a member of $\mathbb{L}_k$, or $\mathbb{L}_k^*$, or both. Intersection of the subsets $\mathbb{L}_k$ and $\mathbb{L}_k^*$, denoted $\mathbb{L}_k \cap \mathbb{L}_k^*$, is the set of all labels that are members of both $\mathbb{L}_k$ and $\mathbb{L}_k^*$. For example, if $\mathbb{L}_k = \{\ell_1, \ell_2, \ell_3\}$ contains $\|\mathbb{L}_k\| = 3$ labels and $\mathbb{L}_k^* = \{\ell_2, \ell_3, \ell_4\}$ also contains $\|\mathbb{L}_k^*\| = 3$ labels, their union $\mathbb{L}_k \cup \mathbb{L}_k^* = \{\ell_1, \ell_2, \ell_3, \ell_4\}$ will contain $\|\mathbb{L}_k \cup \mathbb{L}_k^*\| = 4$ labels; while their intersection $\mathbb{L}_k \cap \mathbb{L}_k^* = \{\ell_2, \ell_3\}$ will contain $\|\mathbb{L}_k \cap \mathbb{L}_k^*\| = 2$ labels. Therefore, when $\mathbb{L}_k$ and $\mathbb{L}_k^*$ are perfectly identical to each other, we should have $\mathbb{L}_k \cup \mathbb{L}_k^* = \mathbb{L}_k \cap \mathbb{L}_k^*$.

For readers' convenience, the symbols used in eqn (16) are illustrated in Fig. 2, from which we can easily see that the rates of all the above five metrics are within the range of 0–1 (or 0–100%).

In eqn (16) the "Aiming" rate (also called "Precision"[34]) is to reflect the average ratio of the correctly predicted labels over the predicted labels; *i.e.*, to measure the percentage of the predicted labels that hit the target of the real labels.

The "Coverage" rate (also called "Recall"[34]) is to reflect the average ratio of the correctly predicted labels over the real labels; *i.e.*, to measure the percentage of the real labels that are covered by the hits of prediction.

The "Accuracy" rate is to reflect the average ratio of correctly predicted labels over the total labels including correctly and incorrectly predicted labels as well as those real labels but are missed in the prediction.

Compared with the aforementioned three rates in eqn (16) that are actually used to reflect the ratios of corrected predictions from different angles, the "Absolute-True" rate (or "Subset-Accuracy"[34]), is the most intuitive and easier-to-understand one for a multi-label system. According to its definition (see the 4th equation of eqn (16)), for a query sample, *e.g.*, the kth sample, when and only when all its labels are exactly predicted without any underprediction (the number of predicted labels is less than the number of real labels) or overprediction (the number of predicted labels is more than the number of real labels), *i.e.*, $\mathbb{L}_k \equiv \mathbb{L}_k^*$ (*cf.* eqn (17)), can the prediction event be scored with 1; otherwise, 0. For example, for a sample having, say, four labels, if three of the four are correctly predicted, or the predicted result contains a label not one of the four, the prediction score will be counted as 0.[3,8] Therefore, the absolute true rate is much more strict and harsh than the other metrics such as aiming rate, coverage rate, and accuracy rate used previously.[35,36] For the absolute true rate, if the labels of a query sample were partially correctly predicted, no score at all would be credited; but for the other metrics rates, a corresponding proportional score would be credited.

The last equation in eqn (16) is the formulation for the "Absolute-False" (also called "Hamming-Loss"[37]) that is completely opposite to the other four metrics. As can be easily seen from its formulation, when the multi-labels for all the samples are correctly predicted, *i.e.*, $\mathbb{L}_k \equiv \mathbb{L}_k^*$ or $\|\mathbb{L}_k \cup \mathbb{L}_k^*\| = \|\mathbb{L}_k \cap \mathbb{L}_k^*\|$ ($k = 1, 2, \cdots, N$), the rate of absolute-false is equal to 0. When each of the samples $\mathbf{P}_k$ ($k = 1, 2, \cdots, N$) is wrongly predicted having all the possible labels except its own true label(s), *i.e.*, $\mathbb{L}_k \cap \mathbb{L}_k^* = +$ (where + means the empty set) and $\mathbb{L}_k \cup \mathbb{L}_k^* = \mathbb{L}$, or $\|\mathbb{L}_k \cup \mathbb{L}_k^*\| = M$ and $\|\mathbb{L}_k \cap \mathbb{L}_k^*\| = 0$, the rate of absolute-false is equal to 1. Therefore, the lower the absolute-false rate is, the better the prediction quality will be. However, for the other four metrics, the meanings of their rates are just opposite; *i.e.*, the higher their rates are, the better the prediction quality will be.

For example, for the multi-label system $\mathbb{S}$ investigated in ref. 4 that contains $N = 5048$ protein samples, of which 2284 were each located in one subcellular location site and hence with one single label, 1740 were each located at two different subcellular locations and hence with two labels, 510 with three labels, 368 with four labels, 111 with five labels, 20 with six labels, 9 with seven labels, 6 with eight labels, and none with nine or more labels. Substituting these data into eqn (3), we have

$$\begin{aligned} N(\text{vir}) &= N + (1-1) \times 2284 + (2-1) \times 1740 \\ &\quad + (3-1) \times 510 + (4-1) \times 368 + (5-1) \\ &\quad \times 111 + (6-1) \times 20 + (7-1) \times 9 \\ &\quad + (8-1) \times 6 + \sum_{m=9}^{20} (m-1) \times 0 \\ &= 5048 + 1740 + 1020 + 1104 + 444 \\ &\quad + 100 + 54 + 42 = 9552 \end{aligned} \tag{18}$$

meaning that the 5048 proteins cover $M = 20$ different subcellular locations *via* $N(\text{vir}) = 9552$ virtual proteins, fully consistent with the data in Table 1.

The rates for the five metrics achieved by the predictor *iLoc-Animal* developed in ref. 4 on such a multi-label system were

$$\begin{cases} \text{Aiming} = 0.7356 \\ \text{Coverage} = 0.6949 \\ \text{Accuracy} = 0.6288 \\ \text{Absolute-True} = 0.4562 \\ \text{Absolute-False} = 0.0518 \end{cases} \tag{19}$$

from which we can see that 73.56% of the predicted labels (or subcellular locations) hit the real labels, that 69.49% of the real labels are covered by the prediction, and that 62.88% of all

Table 2 A comparison of the *iLoc-Animal* predictor with the other existing predictors that are also able to deal with a system with both single- and multiple-location proteins. Reproduced from ref. 4 with permission

Predictor	Absolute-true rate[a]	Coverage or recall rate[a]	Multiplicity degree[b]	Number of locations covered
IMMMLGP[c]	0.2740	0.5950	1.1851	14
Hum-mPLoc2.0[d]	0.2940	0.5190	1.1851	14
iLoc-Animal	0.4562	0.6949	1.8922	20

[a] See eqn (16) for the definitions of ''absolute-true'' and ''coverage'' or ''recall''. [b] See eqn (4) for the definition of ''multiplicity degree'' for a benchmark dataset. [c] The predictor proposed by He *et al.*[38] [d] The predictor proposed by Shen and Chou.[36]

the possible labels are correctly predicted. It can be also seen from eqn (19) that the overall absolute-true rate is 45.62% while the overall absolute-false rate is 5.18%, indicating that even if the prediction reaches accuracy over 62.0% with the absolute-false rate under 5.2%, its absolute-true rate is only 45.62%.

The difficulty in obtaining a high absolute-true rate using a predictor for a multi-label system can also been seen from Table 2. It lists not only the absolute-true rate predicted by *iLoc-Animal* but also the absolute rates reported by *IMMMLGP*[38] and *Hum-mPLoc2.0*,[36] two powerful predictors developed recently that are also able to deal with a system with both single- and multiple-location proteins. As we can see from the table, the absolute-true success rates predicted by *IMMMLGP*[38] and *Hum-mPLoc2.0*[36] were, respectively, 27.40% and 29.40%, and both are even lower than the aforementioned absolute-true success rate achieved by *iLoc-Animal.*

Furthermore, it can also be seen from Table 2 that, compared with the dataset used to test *IMMMLGP* and *Hum-mPLoc2.0*, the benchmark dataset used to test *iLoc-Animal* is much more harsh due to the following facts. (1) It covers 20 subcellular location sites, whereas the dataset used to test *IMMMLGP* and *Hum-mPLoc2.0* only covered 14 location sites. As is well known, the more classes a benchmark dataset covers, the more difficult it is to get a high success rate when using it to test a classifier.[11] (2) The multiplicity degree of the benchmark dataset used to test *iLoc-Animal* was 1.8922 (see Table 2), which is much higher than 1.1851, the multiplicity degree of the benchmark dataset used to test *IMMMLGP* and *Hum-mPLoc2.0*. It is easy to imagine that the higher the multiplicity degree of a benchmark dataset is, the more number of multiple subcellular locations it contains, and hence the more difficult it is to achieve a high absolute-true success rate when using it to test a predictor, as will be further discussed later. However, the overall absolute-true success rate achieved by *iLoc-Animal* is still significantly higher than that achieved by *IMMMLGP* and *Hum-mPLoc2.0*, indicating that *iLoc-Animal* holds a high potential to become a useful high throughput tool in this area.

It is instructive to point out that since a sample may have two or more labels in a multi-label system, it is meaningless to define the absolute-true success rate for the samples in each of the individual groups, such as the proteins in each of the subcellular locations,[4] the drugs in each of the ATC classes,[5] and the antimicrobial peptides in each of the AMP types.[7] This is because the concept of the absolute-true rate is in conflict or inconsistent with the success rate for each of the individual groups in a multi-label system. For instance, suppose a protein, say $\mathbf{P}_k$, can simultaneously occur in the following three subcellular locations: ''cytoplasm'', ''nucleus'', and ''plasma membrane''. However, in its predicted result, only ''cytoplasm'' is shown but the other two locations are missing. Thus, according to the absolute-true rate, the score for predicting $\mathbf{P}_k$ is zero although its subcellular location is correctly predicted as far as the ''cytoplasm'' sub-set is concerned. That is why in all the papers focused on the multi-label systems (see, *e.g.*, ref. 4, 36, and 38) no absolute-true success rate value was provided for each of the individual labels.

Instead, we should consider the absolute true success rates for the samples with different numbers of labels. For example, reported in Table 3 are the absolute-true success rates achieved by using the *iLoc-Animal* predictor[4] for identifying the animal proteins with different numbers of subcellular location sites. For facilitating in-depth analysis, listed in this table are also the corresponding rates by the completely random guess and weighted random guess, as defined below.

The completely random guess (CRG) rates can be derived as follows. First of all, the query protein may have one label, two labels, or up to M labels; *i.e.*, one of the M possibilities regarding the number of labels it may bear. In other words,

Table 3 The absolute true success rates obtained by *iLoc-Animal* for proteins with different numbers of subcellular location sites. Adapted from ref. 4 with permission

Number of location sites or labels	Number of proteins	Absolute-true rate		
		iLoc-Animal	Completely random guess[a] (%)	Weighted random guess[b] (%)
1	2284	$\frac{1240}{2284} = 54\%$	2.5×10^{-1}	2.2
2	1740	$\frac{928}{1740} = 53\%$	2.6×10^{-2}	1.8×10^{-1}
3	510	$\frac{77}{510} = 15\%$	4.4×10^{-3}	8.9×10^{-3}
4	368	$\frac{53}{368} = 14\%$	1.1×10^{-3}	1.5×10^{-4}
5	111	$\frac{5}{111} = 4.5\%$	3.2×10^{-4}	1.4×10^{-4}

[a] The completely random guess was calculated according to eqn (21). [b] The weighted random guess was calculated according to eqn (22).

for a completely random guess, the probability for a query protein to have $m(=1, 2, \cdots, \text{or } M)$ symbol(s) is $1/M$. Secondly, of the M labels, we have

$$C(M,m) = \frac{M!}{(M-m)!m!} \quad (20)$$

distinct ways to pick the m labels. In the above equation, the symbol $C(M,m)$ represents the number of combinations of M distinct things (or locations) taken m at a time. Accordingly, the completely random guess (CRG) rate for a sample to have $m(\leq M)$ labels should be calculated by

$$P(\text{CRG}) = \frac{1}{M} \cdot \frac{1}{C(M,m)} \quad (m \leq M) \quad (21)$$

where M is the total number of all the possible subcellular locations that is equal to 20 for the multi-label system investigated in ref. 4.

The weighted random guess (WRG) rates can be derived *via* the following consideration. For a weighted random guess, the probability for a query protein to have $m(\leq M)$ symbol(s) is $n(m)/N$, where $n(m)$ has the same meaning as in eqn (3), *i.e.*, the number of samples with m different labels, and N is the number of the total samples concerned. Therefore, the weighted random guess (WRG) rates for a sample to have $m(=1, 2, \cdots, \text{or } M)$ label(s) should be calculated by

$$P(\text{WRG}) = \frac{n(m)}{N} \cdot \frac{1}{C(M,m)} \quad (m \leq M) \quad (22)$$

From Table 3, we can see that for those proteins with more number of labels (different location sites), the overall absolute true success rates achieved by the *iLoc-Animal*[4] are generally lower. This is fully consistent with results derived from the random guess approaches. As shown in Table 3, the absolute-true success rates by the completely random guess and weighted random guess for the proteins with one label (location site) are 0.25% and 2.2%, those for the proteins with two labels are 0.026% and 0.18%, those for the proteins with three labels are 0.0044% and 0.0089%, those for the proteins with four labels are 0.0011% and 0.00015%, and those for the proteins with five labels are 0.00032% and 0.00014%, respectively. It can be also seen from the table that the absolute-true rates achieved by *iLoc-Animal* are about 200–14 000 times higher than those achieved by the completely random guess, and about 25–93 000 times higher than those achieved by the weighted random guess.

Table 4 The rates for the other four metrics obtained by *iLoc-Animal* for proteins with different numbers of subcellular location sites

Number of location sites or labels	Number of proteins	Aiming	Coverage	Accuracy	Absolute-false
1	2284	0.6484	0.7631	0.6484	0.0366
2	1740	0.7857	0.7078	0.6781	0.0427
3	510	0.7898	0.5405	0.5113	0.0841
4	368	0.8899	0.5238	0.5127	0.1029
5	111	0.9580	0.4450	0.4375	0.1446

The reason why in Table 3 only the absolute-true rates are shown is because the comparison of the results by the *iLoc-Animal* predictor with the results by the random guess is much easier and intuitive if using the absolute-true metrics. Of course, the corresponding values for the other four metrics can also be calculated as given in Table 4.

VI. Some remarks on the GO approach

In studying multi-label molecular biosystems, particularly in identifying the subcellular localization of multiplex proteins, the gene ontology (GO) approach has been increasingly used to develop various powerful predictors (see, *e.g.*, ref. 3, 8, 9, 16, 18, 36, and 38–49).

GO is a controlled vocabulary used to describe the biology of a gene product in any organism.[50,51] GO database[52] was established according to the molecular function, biological process, and cellular component. Thus, the following questions might be asked regarding the GO approach. If a protein already has GO annotation, why does one need to predict its subcellular location? Is it merely a procedure of converting the annotation from one format into another?

To address these questions, let us consider the following facts. In the literature almost all the existing benchmark datasets constructed by many investigators for predicting protein subcellular localization were taken from the Swiss-Prot database, in which all the proteins were explicitly annotated with their subcellular location information determined from experiments. Can we hence say that the outputs from these predictors are not prediction? Of course, we cannot. This is because all these predictors such as those proposed in ref. 53–61, once established, would yield the desired subcellular locations of query proteins by using the input only containing the sequence information alone without needing any Swiss-Prot annotation information at all. This is exactly the same for those predictors developed by using the GO approach, such as *Euk-mPloc2.0*,[9] *iLoc-Euk*,[8] *iLoc-Hum*,[3] as well as the predictors proposed in ref. 38, 48, and 49. For these GO-approach predictors, once established, the only input for them to perform prediction is the sequences of query proteins without needing any of the GO annotation information whatsoever. Accordingly, as far as the requirement for the input is concerned, there is no difference at all between the non-GO-approach predictors and GO-approach predictors.

Furthermore, it is instructive to note that, of the 5048 protein samples in the benchmark dataset $\mathbb{S}$ used to train and test *iLoc-Animal*, a subcellular location predictor for animal proteins developed recently based on the GO approach,[4] 3977 had no GO annotation terms at all, and their feature vectors in the GO-space were derived through their homologous proteins *via* the following steps. (1) Without losing generality, let us use $\mathbb{P}$ to represent one of the 3977 proteins without any GO terms. (2) Use the BLAST[62] program to search all the proteins in the Swiss-Prot database for those having homologous sequences to $\mathbb{P}$. (3) The homologous proteins thus found were collected into a set, $\mathbb{S}_{\mathbb{P}}^{\text{homo}}$, called the "homology set" of $\mathbb{P}$. All the

elements in $\mathbb{S}_{\mathbb{P}}^{\text{homo}}$ can be deemed as the ''representative proteins'' of $\mathbb{P}$, sharing some similar attributes such as structural conformations and biological functions.[63–65] (4) Use the GO terms of these representative proteins to define the feature vector $\mathbb{P}$ *via* the following equations

$$\psi_u = \frac{\sum_{k=1}^{\mathbb{N}_{\mathbb{P}}^{\text{homo}}} \delta(u,k)}{\mathbb{N}_{\mathbb{P}}^{\text{homo}}} \quad (u = 1, 2, \cdots, \Omega) \tag{23}$$

where $\mathbb{N}_{\mathbb{P}}^{\text{homo}}$ is the number of representative proteins in $\mathbb{S}_{\mathbb{P}}^{\text{homo}}$, $\Omega = 3043$ is the number of total GO terms used in ref. 4, and

$$\delta(u,k) = \begin{cases} 1, & \text{if the } k\text{th representative protein hits} \\ & \quad \text{the } u\text{th GO term considered} \\ 0, & \text{otherwise} \end{cases} \tag{24}$$

Substituting eqn (23) and (24) into eqn (5), we immediately obtain the feature vector in the GO space for $\mathbb{P}$ although it did not have any GO term. Among the 3977 proteins, 1744 were perfectly correctly predicted for their subcellular locations without any over- or under-prediction. In other words, the absolute true success rate achieved by *iLoc-Animal* for the 3977 proteins without any GO annotation terms was 43.85%. Obviously, this is certainly not what could be done by simply converting the format of protein subcellular location annotations from one format to the other, as thought by those who hold a skeptical point of view about the GO approach.

Actually, the essence of why using the GO approach can significantly improve the prediction quality is due to the fact that proteins mapped into the GO-space (instead of the Euclidean space or any other simple geometric space) would be clustered in a way much better reflecting their subcellular locations, as elaborated in ref. 2 and 66.

VII. Conclusion and perspectives

Compared with the single-label systems, it is much more difficult to develop prediction methods for the multi-label systems, for which it is also much more complicated to properly define the prediction quality. A set of five metrics was introduced to reflect the prediction quality from five different angles. Of the five metrics, the ''Absolute-True'' rate is the most intuitive one. It is also the most strict and harsh one. How to enhance its success rate is a big challenge for the future work in this area.

During the last two years or so, some concepts and mathematical formulations developed for the multi-label systems have been utilized to investigate the subcellular localization of proteins in various organisms, classification of anatomical therapeutic chemicals (ATC) for drugs, and identification of the functional type for antimicrobial peptides (AMPs). Although it is still very difficult to gather very accurate information for multi-label biosystems and the current collection of the benchmark dataset might not be complete and accurate, it is anticipated that, with the rapid accumulation of experimental data in the post-genomic age, the existing benchmark datasets will become more complete and accurate, and the method formulated in this review can be straightforwardly used on the improved benchmark datasets to enhance the quality of prediction. Meanwhile, we may also face more and more multi-label molecular biosystems. The concept and approach summarized here will become a useful vehicle to study these new multi-label biosystems.

Since the powerful GO approach has been increasingly used to develop various methods for predicting subcellular localization of proteins with both single and multiple sites (a typical multi-label system), an analysis for justifying the GO approach and the essence of why it is so powerful have been briefly elaborated.

Acknowledgements

The author wishes to thank the two anonymous reviewers for their constructive comments, which were very helpful for strengthening the presentation of this paper.

References

1 E. Glory and R. F. Murphy, *Dev. Cell*, 2007, **12**, 7–16.
2 K. C. Chou and H. B. Shen, *Nat. Sci.*, 2010, **2**, 1090–1103 (openly accessible at http://www.scirp.org/journal/NS/; DOI: 10.4236/ns.2010.210136).
3 K. C. Chou, Z. C. Wu and X. Xiao, *Mol. BioSyst.*, 2012, **8**, 629–641.
4 W. Z. Lin, J. A. Fang, X. Xiao and K. C. Chou, *Mol. BioSyst*, 2013, **9**, 634–644.
5 L. Chen, W. M. Zeng, Y. D. Cai, K. Y. Feng and K. C. Chou, *PLoS One*, 2012, **7**, e35254.
6 G. Wang, X. Li and Z. Wang, *Nucleic Acids Res.*, 2009, **37**, D933–D937.
7 X. Xiao, P. Wang, W. Z. Lin, J. H. Jia and K. C. Chou, *Anal. Biochem.*, 2013, **436**, 168–177.
8 K. C. Chou, Z. C. Wu and X. Xiao, *PLoS One*, 2011, **6**, e18258.
9 K. C. Chou and H. B. Shen, *PLoS One*, 2010, **5**, e9931.
10 K. C. Chou, *Proteins: Struct., Funct., Genet.*, 2001, **43**, 246–255; K. C. Chou, *Proteins: Struct., Funct., Genet.*, 2001, **44**, 60.
11 K. C. Chou, *J. Theor. Biol.*, 2011, **273**, 236–247.
12 C. Chen, Z. B. Shen and X. Y. Zou, *Protein Pept. Lett.*, 2012, **19**, 422–429.
13 H. Mohabatkar, M. M. Beigi, K. Abdolahi and S. Mohsenzadeh, *Med. Chem.*, 2013, **9**, 133–137.
14 G. L. Fan and Q. Z. Li, *J. Theor. Biol.*, 2012, **304**, 88–95.
15 M. Hayat and A. Khan, *Protein Pept. Lett.*, 2012, **19**, 411–421.
16 G. L. Fan and Q. Z. Li, *Amino Acids*, 2012, **43**, 545–555.
17 M. M. Beigi, M. Behjati and H. Mohabatkar, *J. Struct. Funct. Genomics*, 2011, **12**, 191–197.
18 L. Q. Li, Y. Zhang, L. Y. Zou, Y. Zhou and X. Q. Zheng, *Protein Pept. Lett.*, 2012, **19**, 375–387.
19 B. Liao, Q. Xiang and D. Li, *Protein Pept. Lett.*, 2012, **19**, 1133–1138.
20 H. Mohabatkar, M. Mohammad Beigi and A. Esmaeili, *J. Theor. Biol.*, 2011, **281**, 18–23.
21 S. W. Zhang, Y. L. Zhang, H. F. Yang, C. H. Zhao and Q. Pan, *Amino Acids*, 2008, **34**, 565–572.

22 L. Nanni, S. Brahnam and A. Lumini, *Amino Acids*, 2012, **43**, 657–665.
23 L. Liu, X. Z. Hu, X. X. Liu, Y. Wang and S. B. Li, *Protein Pept. Lett.*, 2012, **19**, 439–449.
24 X. H. Niu, X. H. Hu, F. Shi and J. B. Xia, *Protein Pept. Lett.*, 2012, **19**, 940–948.
25 Y. F. Qin, C. H. Wang, X. Q. Yu, J. Zhu, T. G. Liu and X. Q. Zheng, *Protein Pept. Lett.*, 2012, **19**, 388–397.
26 X. Y. Sun, S. P. Shi, J. D. Qiu, S. B. Suo, S. Y. Huang and R. P. Liang, *Mol. BioSyst.*, 2012, **8**, 3178–3184.
27 X. W. Zhao, Z. Q. Ma and M. H. Yin, *Protein Pept. Lett.*, 2012, **19**, 492–500.
28 Y. Xu, J. Ding, L. Y. Wu and K. C. Chou, *PLoS One*, 2013, **8**, e55844.
29 Y. K. Chen and K. B. Li, *J. Theor. Biol.*, 2013, **318**, 1–12.
30 W. Chen, P. M. Feng, H. Lin and K. C. Chou, *Nucleic Acids Res.*, 2013, DOI: 10.1093/nar/gks1450.
31 W. Chen, H. Lin, P. M. Feng, C. Ding, Y. C. Zuo and K. C. Chou, *PLoS One*, 2012, **7**, e47843.
32 B. Q. Li, T. Huang, L. Liu, Y. D. Cai and K. C. Chou, *PLoS One*, 2012, **7**, e33393.
33 T. Huang, J. Wang, Y. D. Cai, H. Yu and K. C. Chou, *PLoS One*, 2012, **7**, e34460.
34 G. Tsoumakas, I. Katakis and I. Vlahavas, in *Data Mining and Knowledge Discovery Handbook*, ed. O. Maimon and L. Rokach, Springer, Heidelberg, 2nd edn, 2010, pp. 1–19.
35 K. C. Chou and H. B. Shen, *Anal. Biochem.*, 2007, **370**, 1–16.
36 H. B. Shen and K. C. Chou, *Anal. Biochem.*, 2009, **394**, 269–274.
37 G. Tsoumakas and l. katakis, *Int. J. Data Warehousing Mining*, 2007, **3**, 13.
38 J. He, H. Gu and W. Liu, *PLoS One*, 2012, **7**, e37155.
39 K. C. Chou and H. B. Shen, *J. Proteome Res.*, 2007, **6**, 1728–1734.
40 H. B. Shen and K. C. Chou, *Biochem. Biophys. Res. Commun.*, 2007, **355**, 1006–1011.
41 H. B. Shen and K. C. Chou, *J. Theor. Biol.*, 2010, **264**, 326–333.
42 K. C. Chou and H. B. Shen, *PLoS One*, 2010, **5**, e11335.
43 X. Xiao, Z. C. Wu and K. C. Chou, *J. Theor. Biol.*, 2011, **284**, 42–51.
44 X. Xiao, Z. C. Wu and K. C. Chou, *PLoS One*, 2011, **6**, e20592.
45 Z. C. Wu, X. Xiao and K. C. Chou, *Mol. BioSyst.*, 2011, **7**, 3287–3297.
46 Z. C. Wu, X. Xiao and K. C. Chou, *Protein Pept. Lett.*, 2012, **19**, 4–14.
47 X. Wang and G. Z. Li, *PLoS One*, 2012, **7**, e36317.
48 S. Mei, *J. Theor. Biol.*, 2012, **293**, 121–130.
49 S. Mei, *J. Theor. Biol.*, 2012, **310**, 80–87.
50 E. Camon, M. Magrane, D. Barrell, D. Binns, W. Fleischmann, P. Kersey, N. Mulder, T. Oinn, J. Maslen, A. Cox and R. Apweiler, *Genome Res.*, 2003, **13**, 662–672.
51 D. Barrell, E. Dimmer, R. P. Huntley, D. Binns, C. O'Donovan and R. Apweiler, *Nucleic Acids Res.*, 2009, **37**, D396–D403.
52 M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, *Nat. Genet.*, 2000, **25**, 25–29.
53 K. Nakai and P. Horton, *Trends Biochem. Sci.*, 1999, **24**, 34–36.
54 K. C. Chou and D. W. Elrod, *Protein Eng.*, 1999, **12**, 107–118.
55 O. Emanuelsson, H. Nielsen, S. Brunak and G. von Heijne, *J. Mol. Biol.*, 2000, **300**, 1005–1016.
56 G. P. Zhou and K. Doctor, *Proteins: Struct., Funct., Genet.*, 2003, **50**, 44–48.
57 S. Matsuda, J. P. Vert, H. Saigo, N. Ueda, H. Toh and T. Akutsu, *Protein Sci.*, 2005, **14**, 2804–2813.
58 J. L. Gardy, M. R. Laird, F. Chen, S. Rey, C. J. Walsh, M. Ester and F. S. Brinkman, *Bioinformatics*, 2005, **21**, 617–623.
59 A. Hoglund, P. Donnes, T. Blum, H. W. Adolph and O. Kohlbacher, *Bioinformatics*, 2006, **22**, 1158–1165.
60 P. Mundra, M. Kumar, K. K. Kumar, V. K. Jayaraman and B. D. Kulkarni, *Pattern Recognit. Lett.*, 2007, **28**, 1610–1615.
61 P. Horton, K. J. Park, T. Obayashi, N. Fujita, H. Harada, C. J. Adams-Collier and K. Nakai, *Nucleic Acids Res.*, 2007, **35**, W585–W587.
62 A. A. Schaffer, L. Aravind, T. L. Madden, S. Shavirin, J. L. Spouge, Y. I. Wolf, E. V. Koonin and S. F. Altschul, *Nucleic Acids Res.*, 2001, **29**, 2994–3005.
63 Y. Loewenstein, D. Raimondo, O. C. Redfern, J. Watson, D. Frishman, M. Linial, C. Orengo, J. Thornton and A. Tramontano, *Genome Biol.*, 2009, **10**, 207.
64 M. Gerstein and J. M. Thornton, *Curr. Opin. Struct. Biol.*, 2003, **13**, 341–343.
65 K. C. Chou, *Curr. Med. Chem.*, 2004, **11**, 2105–2134.
66 K. C. Chou and H. B. Shen, *Nat. Protocols*, 2008, **3**, 153–162.

