Journal of Theoretical Biology 318 (2013) 197–204

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Protein space: A natural method for realizing the nature of protein universe

Chenglong Yu [a], Mo Deng [a], Shiu-Yuen Cheng [b], Shek-Chung Yau [c], Rong L. He [d,*], Stephen S.-T. Yau [e,**]

[a] Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, Chicago, IL, USA
[b] Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
[c] Information Technology Services Center, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
[d] Department of Biological Sciences, Chicago State University, Chicago, IL, USA
[e] Department of Mathematical Sciences, Tsinghua University, Beijing, PR China

HIGHLIGHTS

- The protein universe can be realized in a 60-dimensional Euclidean space termed as the protein space.
- The distance between two points in protein space represents the biological distance of the corresponding two proteins.
- We propose a natural and unique graphical representation for inferring protein phylogenies.
- Our new approach will solve the fundamental question of how proteins are distributed in the protein universe.

ARTICLE INFO

Article history:
Received 7 July 2012
Received in revised form
1 November 2012
Accepted 2 November 2012
Available online 12 November 2012

Keywords:
Natural vector
Phylogeny
Protein universe

ABSTRACT

Current methods cannot tell us what the nature of the protein universe is concretely. They are based on different models of amino acid substitution and multiple sequence alignment which is an NP-hard problem and requires manual intervention. Protein structural analysis also gives a direction for mapping the protein universe. Unfortunately, now only a minuscule fraction of proteins' 3-dimensional structures are known. Furthermore, the phylogenetic tree representations are not unique for any existing tree construction methods. Here we develop a novel method to realize the nature of protein universe. We show the protein universe can be realized as a protein space in 60-dimensional Euclidean space using a distance based on a normalized distribution of amino acids. Every protein is in one-to-one correspondence with a point in protein space, where proteins with similar properties stay close together. Thus the distance between two points in protein space represents the biological distance of the corresponding two proteins. We also propose a natural graphical representation for inferring phylogenies. The representation is natural and unique based on the biological distances of proteins in protein space. This will solve the fundamental question of how proteins are distributed in the protein universe.

1. Introduction

The protein universe is the collection of all known proteins (Ladunga, 1992). It is a large and mysterious entity, which is an essential underpinning of all biology (Levitt, 2009). The current methods (Levitt, 2009; Dokholyan et al., 2002; Jaroszewski et al., 2009; Koonin, 2007; Koonin et al., 2002; Povolotskaya and Kondrashov, 2010) to reveal the nature of the protein universe cluster sequences into families by similarities. However, these methods cannot even tell us what the nature of the protein universe is concretely. Moreover, they are based on different models of amino acid substitution (Yang, 2006) and require manual intervention, and therefore the results are often controversial. On the other hand, the methods generate protein families using multiple sequence alignment (Altschul et al., 1990; Lipman., Pearson (1985); Smith and Waterman, 1981) which is an NP-hard problem. Analysis of 3-dimensional structures of proteins also gives a direction for mapping the protein universe (Holm and Sander, 1996). Unfortunately, up to now, only a minuscule fraction of proteins' 3-dimensional structures are known (Berman et al., 2000). Therefore, the current methods are impossible to annotate the huge protein universe of 8 million members.

Detecting homology may help in partially realizing the nature of protein universe. Domains play an important role in studying the homology of proteins. Domains in protein sequences and structures can evolve, function, and exist independently of the rest of the protein chain. Because they are independently stable, domains become the important bases for inferring homology and classifying proteins. Many studies (Bateman et al., 2004; Corpet et al., 2000) show that protein domains are powerful in the analysis of newly discovered protein sequences. However, many proteins consist of at

* Corresponding author.
** Corresponding author. Tel.: + 86 10 62787874; fax: + 86 10 62798033.
E-mail addresses: rhe@csu.edu (R.L. He), yau@uic.edu (S.-T. Yau).

0022-5193/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2012.11.005

least two domains. These domains and nature of their interactions determine the function of the protein. Therefore, multidomain proteins make the deduction of homology very difficult. For example, if protein 1 contains domains A and B, protein 2 contains domains B and C, protein 3 contains domains C and D, then are protein 1 and protein 3 homologous? This simple example indicates the inadequacy of domain-based analysis methods.

The total number of all protein sequences, i.e., the size of the protein universe, is very large. A question of fundamental and practical interest is how these sequences are distributed in this universe (Koonin et al., 2002). The answer of this question may reveal important aspects of the evolution of proteins from a diverse range of organisms. To answer this question, we need to have a space where all proteins live so that this question makes sense.

We develop a novel method, protein space, to realize the nature of protein universe that is motivated from our previous studies of genomes (Yu et al., 2010; Deng et al., 2011). Unlike the current protein databases, our proposed protein space is a space where all proteins live, which supports simultaneous comparative study for all available proteins. We can accomplish the "impossible mission" of characterizing the huge protein universe in a relatively short time period. Furthermore, we propose a novel graphical representation for protein phylogeny. The representation is natural and unique based on the Euclidean distances of proteins in protein space. This will solve the fundamental question of how proteins are distributed in the protein universe.

2. Materials and methods

We will construct our protein space as a subspace in R^{20N+20} ($N \geq 2$) by means of the natural vector mapping which is based on the global distributions of the protein sequences. Every protein is in one-to-one correspondence with a point in this protein space. The Euclidean distance between two points truly represents the biological distance of the corresponding two proteins. We can perform phylogenetic and cluster analysis for all the existing proteins. A key finding is that this protein space is a 60-dimensional space. We emphasize that our natural vectors depend only on the numbers and distributions of amino acids in the protein sequences. They do not depend on any model assumption. There are two reasons that the protein is represented as a point in the protein space without losing inherent biological information. First, the 60-dimensional natural vector mapping on all the data-sets we examined is one-to-one. Second, we do not gain any more information using the 80-dimensional natural vector mapping. Our new approach of classifying proteins is not a domain-based method. Our protein space is constructed based on the global sequence information of proteins, and thus natural and convincing.

2.1. Natural vector

To avoid losing any important information hidden in protein sequences, the pseudo amino acid composition (PseAAC) was proposed (Chou, 2001, 2005) to replace the simple amino acid composition (AAC) for representing the sample of a protein. For a summary about its recent development and applications, see a comprehensive review (Chou, 2009). Ever since the concept of PseAAC was proposes by Chou (2001), its has rapidly penetrated into almost all the fields of protein attribute prediction, such as identifying bacterial virulent proteins (Nanni et al., 2011), predicting homo-oligomeric proteins (Qiu et al., 2011), predicting protein secondary structure content (Chen et al., 2006, 2009), predicting supersecondary structure (Zou et al., 2011), predicting protein structural classes (Lin and Li, 2007; Li et al., 2009; Sahu and Panda, 2010), predicting protein quaternary structure (Zhang et al., 2008), predicting enzyme family and sub-family classes (Zhou et al., 2007; Qiu et al., 2010; Wang et al., 2010), predicting protein subcellular location (Li and Li, 2008; Zhang et al., 2008; Du et al., 2009; Fan and Li, 2012), predicting subcellular localization of apoptosis proteins (Ding and Zhang, 2008; Jiang et al., 2008; Li et al., 2009; Kandaswamy et al., 2010), predicting protein subnuclear location (Jiang et al., 2008; Xiao et al., 2012), predicting protein submitochondria locations (Lin et al., 2008; Nanni and Lumini, 2008; Zeng et al., 2009), identifying cell wall lytic enzymes (Ding et al., 2009), identifying risk type of human papillomaviruses (Esmaeili et al., 2010), identifying DNA-binding proteins (Fang et al., 2008; Lin et al., 2011), predicting G-Protein-Coupled Receptor Classes (Qiu et al., 2009; Gu et al., 2010), predicting protein folding rates (Guo et al., 2011), predicting outer membrane proteins (Lin, 2008; Gao et al., 2010; Mahdavi and Jahandideh, 2011; Hayat and Khan, 2012), predicting cyclin proteins (Mohabatkar, 2010), predicting GABA(A) receptor proteins (Mohabatkar et al., 2011), identifying bacterial secreted proteins (Yu et al., 2010), identifying the cofactors of oxidoreductases (Zhang and Fang, 2008), identifying lipase types (Zhang et al., 2008), identifying protease family (Hu et al., 2011), predicting Golgi protein types (Ding et al., 2011), classifying amino acids (Georgiou et al., 2009), mapping protein sequences (Yau et al., 2008; Wu et al., 2010; Yu et al., 2011), combining with cellular automata (Xiao and Chou, 2011; Xiao et al., 2011; Xiao et al., 2011), among many others.

According to Eq. (6) of a recent comprehensive review (Chou, 2011), the general form of Chou's PseAAC can be formulated as $P=[\psi_1,\psi_2,\ldots,\psi_\Omega]^T$, where T is a transpose operator, while the subscript Ω reflects the dimension of the vector and its value as well as the components $\psi_1,\psi_2,\ldots$ will be defined by a series of feature extractions as elaborated below.

Let us first introduce the definition of normalized central moments which is the most important part of natural vector. Normalized central moments are defined as follows:

$$D_j^k=\sum_{i=1}^{n_k}\frac{(s[k][i]-\mu_k)^j}{n_k^{j-1}n^{j-1}},\quad j=1,2,\ldots,n_k$$

where $k=20$ amino acids (A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V). n_k denotes the number of amino acid k in the protein sequence and n is the length of the protein sequence. $s[k][i]$ is the distance from the first amino acid (regarded as origin) to the ith amino acid k in the protein sequence. $T_k=\sum_{i=1}^{n_k}s[k][i]$ denotes the total distance of each set of 20 amino acids to the origin. $\mu_k=\frac{T_k}{n_k}$, which is the mean value of the distances of the amino acids from the origin. Therefore, we have the sequence of normalized central moments: $\langle D_1^A, D_2^A,\ldots, D_{n_A}^A, D_1^R, D_2^R,\ldots, D_{n_R}^R,\ldots, D_1^V, D_2^V,\ldots, D_{n_V}^V\rangle$.

Observe that these are natural parameters associated to a protein sequence.

Our method described below gives a complete understanding of the distribution of 20 amino acids.

(1) The quantities of the 20 amino acids of a protein sequence are chosen as the first 20 parameters of the natural vector. The 20 integers n_A, n_R, n_N, …, n_V denote the numbers of 20 amino acids in a protein sequence.

(2) The second group of 20 numerical parameters which are a part of the natural vector are the arithmetic mean values of total distance for each of the 20 amino acids:

$$\mu_k=\frac{T_k}{n_k},\quad k=\text{A, R, N,}\ldots\text{,V.}$$

(3) The final group of parameters that we include in the natural vector are composed of normalized central moments as defined earlier. If the distribution of each amino acid is different, protein

sequences cannot be similar even though they may have the same amino acid contents and the same total distance measurement. Therefore, the information about distribution has also been included in the natural vector. As described above, each subset of numerical parameters is not sufficient to annotate protein sequences. However, the combined numerical parameters are sufficient to characterize each protein sequence. So the natural vector is given as follows:

$$\left\langle n_A,\mu_A,D_1^A,\ldots,D_{n_A}^A,n_R,\mu_R,D_1^R,\ldots,D_{n_R}^R,\ldots,n_V,\mu_V,D_1^V,\ldots,D_{n_V}^V\right\rangle$$

If a specific amino acid k does not exist, we define n_k, μ_k, and D_j^k to be zero. In order to express the vector elegantly, we rewrite it as follows:

$$\left\langle n_A,n_R,\ldots,n_V,\mu_A,\mu_R,\ldots,\mu_V,D_1^A,D_1^R,\ldots,D_1^V,\ldots,D_{n_A}^A,D_{n_R}^R,\ldots,D_{n_V}^V\right\rangle \quad (1)$$

Alternatively, the natural vector can be written as

$$\left\langle n_A,n_R,\ldots,n_V,\mu_A,\mu_R,\ldots,\mu_V,D_1^A,D_1^R,\ldots,D_1^V,\ldots,D_{n_\pi}^A,D_{n_\pi}^R,\ldots,D_{n_\pi}^V\right\rangle \quad (2)$$

where $n_\pi=\max\{n_A,n_R,\ldots,n_V\}$. By definition, $D_j^k=0$, if $j>n_k$.

We prove mathematically that the correspondence between a protein sequence and its natural vector is one-to-one. Actually, all the 1st order central moments $D_1^A,D_1^R,\ldots,D_1^V$ are zero, so we do not need to compute them in the natural vector. (See the theorem in Supplementary material)

2.2. Construction of protein space

The natural vector is obtained by concatenating the first group of parameters (the number of each base) and the second group of parameters (the mean value of total distance of each base) to the normalized central moments. Obviously, higher moments converge to zero for a random generated sequence since for any given k,

$$D_j^k=\sum_{i=1}^{n_k}\frac{(s[k][i]-\mu_k)^j}{n_k^{j-1}n^{j-1}}\le\sum_{i=1}^{n_k}\frac{\max_i|s[k][i]-\mu_k|^j}{n_k^{j-1}n^{j-1}}$$
$$\le n_k\frac{\max_i|s[k][i]-\mu_k|^j}{n_k^{j-1}n^{j-1}}\le\frac{n^j}{n_k^{j-2}n^{j-1}}=\frac{n}{n_k^{j-2}}$$

It is clear that$n_k\ge 2$, otherwise, $s[k][i]-\mu_k=0$, which yields $D_j^k=0$. From the viewpoint of probability, suppose that the expectation number of any amino acid is $n_k=n/20$(uniform distribution) for a sequence with given length n, then

$$\lim_j\frac{n}{n_k^{j-2}}=\lim_j\frac{n}{(n/20)^{j-2}}=\lim_j\frac{n\times 20^{j-2}}{n^{j-2}}=\lim_j\frac{20^{j-2}}{n^{j-3}}$$

clearly, this limit goes to 0 as j approaches n_k. For example, for a kinase C protein from human (GenBank ID: P05771) which has length 671, we can get $D_2^A=69.9855$, $D_3^A=-0.2916$, $D_4^A=0.0127$, $D_5^A=-9.7059\text{e}{-005}$, $D_6^A=2.7284\text{e}{-006}$. That is, the higher central moments converge to zero very quickly.

We will use this natural vector to construct a protein space. A protein space is a moduli space of proteins. In this space, each point corresponds to a protein. The natural distance between two proteins in the protein space reflects the biological distance between these two proteins. For a protein sequence with length n, we can compute its $(n+20)$-dimensional natural vector

$$\left\langle n_A,n_R,\ldots,n_V,\mu_A,\mu_R,\ldots,\mu_V,D_2^A,D_2^R,\ldots,D_2^V,\ldots,D_{n_A}^A,D_{n_R}^R,\ldots,D_{n_V}^V\right\rangle. \quad (3)$$

But the breakthrough of the subject is that we do not need to compute the high central moments in the vector since we have explained that the higher central moments converge to zero very quickly. So, when computing the distance between two natural vectors the high central moments hardly make any contribution. Thus, we can get a low dimensional natural vector by only using the first several central moments:

$$\left\langle n_A,n_R,\ldots,n_V,\mu_A,\mu_R,\ldots,\mu_V,D_2^A,D_2^R,\ldots,D_2^V,\ldots,D_N^A,D_N^R,\ldots,D_N^V\right\rangle \quad (4)$$

This vector is $(20N+20)$-dimensional with $N\ll n$. Using these natural vectors, we can construct the protein space as a subspace in R^{20N+20}. Every protein corresponds to a point in this Euclidean space. In this work, using the Euclidean distance between two points as a metric, we perform phylogenetic and clustering analysis for the protein sequences. We also tried other distances, such as Mahalanobis distance, Manhattan distance, Chebyshev distance and cosine distance. The current results show that the Euclidean distance is the best of them. Actually, the Euclidian distance is the most natural distance in the Euclidian space, but whether it is the best metric for all the universal proteins still needs further study. Here we use $N=2$ because the 60-dimensional natural vectors have allowed us to obtain stable classified results—when higher moments are included, the relationship of being close or farther away remains unchanged. Here we present an example of three PKC proteins to show this. For three PKC proteins (NP_001006133, nPKC; NP_001008716, nPKC; NP_001012707, aPKC), we calculate their 80-dimensional natural vectors ($N=3$): NP_001006133, nPKC: (38, 35, 29, 43, 20, 47, 29, 47, 16, 39, 57, 64, 17, 50, 27, 39, 36, 11, 19, 36, 287.03, 336.06, 379.48, 425.07, 262.55, 351.45, 255.59, 328.55, 341.94, 347.33, 370.61, 355.22, 286.88, 387.28, 376.81, 385.56, 353.03, 377.18, 370.74, 309.61, 49.405, 57.129, 57.741, 43.69, 25.892, 71.847, 53.46, 45.969, 39.234, 61.83, 55.831, 56.245, 59.089, 60.672, 82.67, 61.97, 61.619, 55.095, 54.556, 45.66, 0.15244, 0.045811, 0.0063567, −0.08494, 0.1412, −0.032894, 0.37673, 0.0007835, 0.089498, −0.0092364, −0.086433, 0.02586, 0.26779, −0.10689, −0.28529, −0.086607, −0.03630, −0.43702, −0.12304, 0.039236); NP_001008716, nPKC: (39, 28, 28, 40, 21, 47, 28, 44, 18, 40, 56, 63, 18, 47, 24, 32, 33, 10, 20, 37, 261.23, 320.32, 357.93, 392.73, 263.62, 341.66, 281.25, 346.11, 331.89, 347.07, 364.25, 341.03, 264.33, 384.45, 356.96, 331.19, 344.76, 355.4, 344.8, 312.35, 60.707, 60.429, 45.524, 47.915, 23.464, 65.424, 50.299, 39.828, 37.491, 48.41, 54.22, 54.973, 56.995, 57.679, 79.029, 69.966, 62.278, 54.331, 57.306, 49.799, 0.21819, 0.14251, 0.15492, −0.12262, 0.077574, −0.041408, 0.20865, −0.022511, 0.033844, −0.035522, −0.097136, 0.026068, 0.33381, −0.10588, −0.19271, 0.033953, −0.050089, −0.31804, −0.15787, 0.046949); NP_001012707, aPKC: (24, 37, 3, 44, 17, 46, 26, 37, 20, 31, 49, 33, 19, 33, 33, 36, 22, 7, 16, 34, 294.29, 251.73, 314.78, 315.48, 216.06, 301.17, 320.5, 304.62, 235.45, 288.81, 315.24, 282.33, 255.95, 336, 294.73, 284.81, 265.77, 286, 277.81, 326.24, 44.008, 37.624, 54.748, 55.53, 42.176, 52.532, 60.399, 40.566, 34.654, 48.619, 37.548, 41.273, 53.228, 56.898, 58.504, 49.77, 55.023, 45.686, 35.078, 46.843, 0.060023, 0.089067, −0.1222, −0.10414, 0.6895, 0.036974, −0.22546, −0.024875, 0.088956, 0.015227, −0.01682, 0.06324, 0.081133, −0.23701, 0.013292, −0.023058, −0.012138, 0.021487, 0.015769, −0.071503). We can see that the 3-order central moments have already been approaching zero. The distance matrix for these 80-dimensional vectors is:

The reason for this is because when computing the distance between two natural vectors the high central moments (approaching zero) hardly make any contribution. Moreover, the 60-dimensional natural vector mapping restricted on the dataset we examined is still one-to-one mapping.

The protein space supports simultaneous comparative study for all available proteins which other methods cannot do it in real time. The results can be used to predict properties of unknown proteins based on their amino acid sequences distribution. Once a protein space has been constructed, it can be stored in a database. There is no need to reconstruct the protein space for any subsequent application, whereas in multiple alignment methods, realignment is needed for add-on new sequences. Furthermore,

one can have global comparison of all proteins simultaneously, which no other existing method can achieve this. Thus, the protein space provides a new powerful tool for analyzing the classification of proteins and their phylogenetic relationships.

	NP_001006133	NP_001008716	NP_001012707
NP_001006133	0	98.91	326.56
NP_001008716	98.91	0	262.52
NP_001012707	326.56	262.52	0

The distance matrix for the 60-dimensional vectors is still:

	NP_001006133	NP_001008716	NP_001012707
NP_001006133	0	98.91	326.56
NP_001008716	98.91	0	262.52
NP_001012707	326.56	262.52	0

2.3. A novel graphical representation for protein phylogeny

Inferring phylogenies from molecular sequences classically has two phases: a biological distance is estimated relying on alignment, and then a tree is produced based on the distance. Unfortunately this distance is not natural as it depends on the alignment parameters. Distance matrices are usually used for phylogenetic analysis of DNA and proteins. Many algorithms (Sokal and Michener, 1958; Fitch and Margoliash, 1967; Saitou and Nei, 1987) may produce either rooted or unrooted phylogenetic trees based on the distance matrices. For example, the neighbor-joining algorithm (Saitou and Nei, 1987) produces unrooted trees, while the UPGMA algorithm (Sokal and Michener, 1958) produces rooted trees. Given a distance matrix, the resulting trees are not unique for any existing tree construction methods (Backeljau et al., 1996). Thus, the phylogenetic results are controversial due to the above two basic problems.

Since we have already constructed natural distance between two proteins, to overcome the disadvantages of existing methods, we propose a natural graphical representation for inferring phylogenies. Specifically, given a distance matrix of finite elements, the algorithm is as follows:

(1) For each element A, find the closest elements $B_1, B_2, \ldots, B_k$ to A. Then draw directed lines from A to$B_1, B_2, \ldots, B_k$.
(2) We then get many graphs after step (1). We compute the distance matrix for these graphs. The distance between two graphs is defined as the minimum of all distance between any element in one graph and any element in the other graph.
(3) We then obtain a new distance matrix, in which the elements are the connected graphs obtained in step (2).
(4) Repeat the process in steps (1) and (2). Finally, we get one connected graph for all elements, which is the final graphical representation.

A detailed example of the construction process is given in Supplementary material. The direction in the graph can show the closest elements of each element based on their biological distances. For example, given a protein (A), biologists would like to know which protein (B) is closest to (A), then an arrow from (A) to (B) in the graph represents this relation. Furthermore, we can let the lengths of lines in the graphical representation be proportional with the biological distances in protein space. Here we need to point out that the natural graphical representation is not necessarily a tree. After realizing the protein universe of 8 million members in our protein space, a cycle may exist in the graphical representation. Actually, when one protein has two equidistant nearest neighbors in the protein space, a cycle may exist in the natural graphical representation (see a simulation example in supplementary Fig. 1). In this case, the existence of cycle may provide new information for the biologists, e.g. the evolutionary relationship of proteins in the cycle may be not straight but circular, say, A evolves into B, B evolves into C, and C evolves into A back.

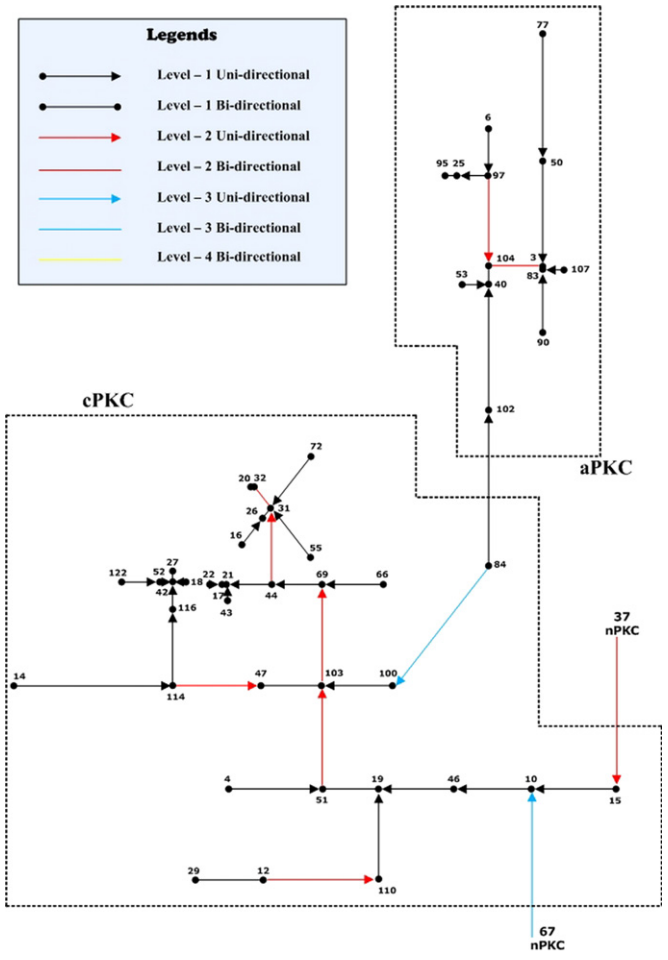

Fig. 1. The natural graphical representation of 124 proteins from PKC like subfamily. We break the large original figure into three pieces: (A)–(C). This figure is Part (A).

3. Results and discussion

We will use a real protein dataset to examine our proposed protein space and its natural graphical representation. Protein kinase C (PKC) is a family of enzymes which are involved in controlling the function of other proteins through the phosphorylation of hydroxyl groups of serine and threonine amino acid residues on these proteins. The structure of all PKC proteins consists of a regulatory domain and a catalytic domain tethered together by a hinge region. The regulatory domain is often the principal determinant of classification, as the catalytic domain tends to be highly conserved. PKC family is divided into three subfamilies: conventional PKCs (cPKCs: α, βI, βII, and γ), novel PKCs (nPKCs: θ, ε, δ, and η), and atypical PKCs (aPKCs: λ/ι, ζ) [Mellor and Parker, 1998]. However, a controversial group of potential PKCs including PKCν and PKCμ/PKD (protein kinase D) has regulatory domains similar to PKCs, but their catalytic domains are more similar to the myosin light-chain kinase of *Dictostelium* (Hurley et al., 1997; Webb et al., 2000). Furthermore, fungi have PKC homologs that characteristically contain more residues than mammalian PKCs with significantly different regulatory domains but similar catalytic domains (Mellor and Parker, 1998). There are also PKC-related kinases (PRKs) that are found in many animals and have features similar to fungal PKCs (Mellor and Parker, 1998). Thus, domain-based classification of this group is controversial because it is not based on the full protein.

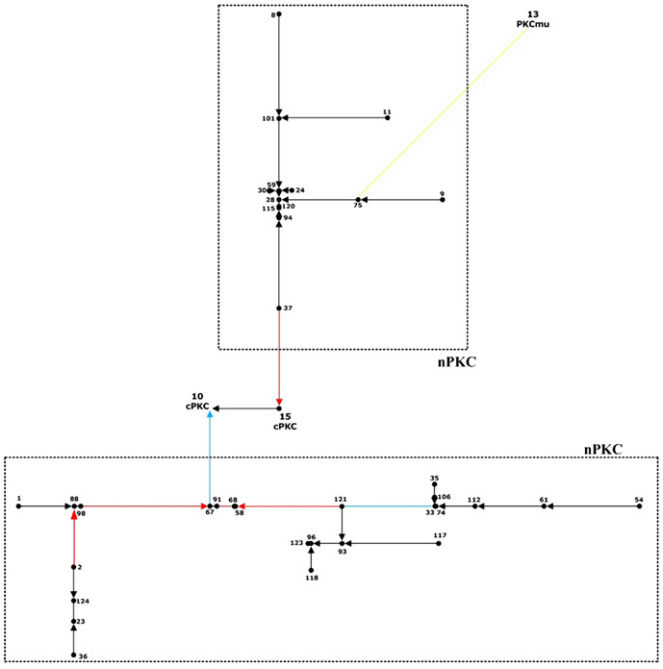

Fig. 2. The natural graphical representation of 124 proteins from PKC like subfamily. This figure is Part (B).

As discussed above, the PKC-like superfamily is composed of six categories of PKCs and PKC-related protein molecules: cPKC, nPKC, aPKC, PKCmu (ν, μ, and D2 types), PKC1 (from fungus), and PRK (similar to PKC1 but from animals). We examined a dataset of 124 proteins from the PKC-like superfamily as shown in Table S3 in Supplementary material. Here we calculate the 60-dimensional natural vector: $\left\langle n_A,n_R,....,n_V,\mu_A,\mu_R,...,\mu_V,D_2^A,D_2^R,...,D_2^V\right\rangle$ for the 124 proteins. By computing the Euclidean distances between these vectors, we obtain the distance matrix. In Figs. 1–3, we give the natural graphical representation for the 124 proteins. The lengths of

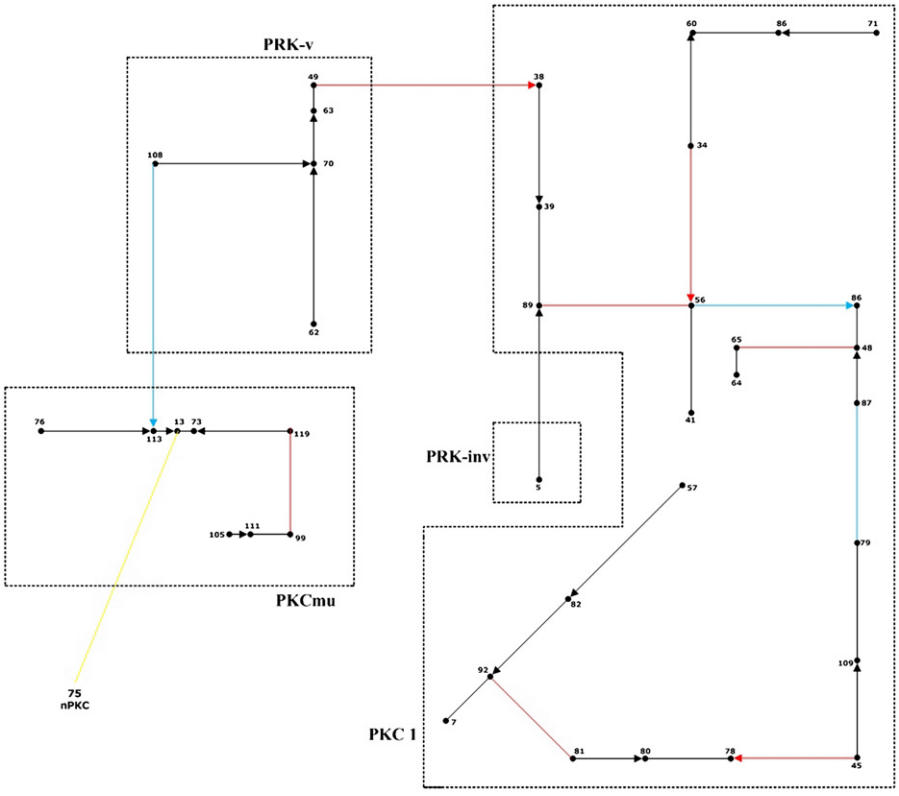

Fig. 3. The natural graphical representation of 124 proteins from PKC like subfamily. This figure is Part (C).

lines in this graphical representation are proportional with the biological distances among the proteins. Our classifying results for these proteins totally agree with those from GenBank (NCBI) descriptions and literature (see Table S3). For No. 5 (GenBank ID: O17874), it is a PKC-like protein from C. elegans, and has a PKC-related kinase homology region 1 domain which is often found in vertebrate PKC and yeast PKC1 proteins. In the natural graphical representation, it is closest to No. 89 (a PKC1 protein). Furthermore, we check all the other PRK proteins in the dataset: No. 49 (from mouse), No. 62 (from human), No. 63 (from rat), No. 70 (from human), and No. 108 (from flog). Clearly, they are all from vertebrate animals. Thus, we believe that PRK subfamily should be divided into two smaller groups, one is from vertebrate animals (PRK-v) and the other is from invertebrate animals (PRK-inv). Thus, No. 5 belongs to a new subfamily PRK-inv, which is more close to PKC1 subfamily than PRK-v. For No. 84 (GenBank ID: Q69G16), GenBank describes it as cPKC. According to our result, the closest protein to it is No. 102 (an aPKC protein), but the next closest to it is No. 100 (a cPKC protein). Thus, our theory predicts that there are some cPKC members missing in our dataset, lying between No. 84 and No. 100 in our protein space. It is the job for biologists to find these new cPKC members. This unique natural graphical representation gives a whole picture of phylogenetic relationships of the PKC like superfamily. It allows us to have global comparison of proteins simultaneously, which no other existing method can achieve. For a better direct comparison between our method and the traditional method on protein sequence classification, we provide a phylogenetic tree of protein kinase C with the traditional graph representation as shown in Fig. 4. This figure is directly from the work of Mellor and Parker (1998), and made by Clustal V software with PAM 250 residue tables.

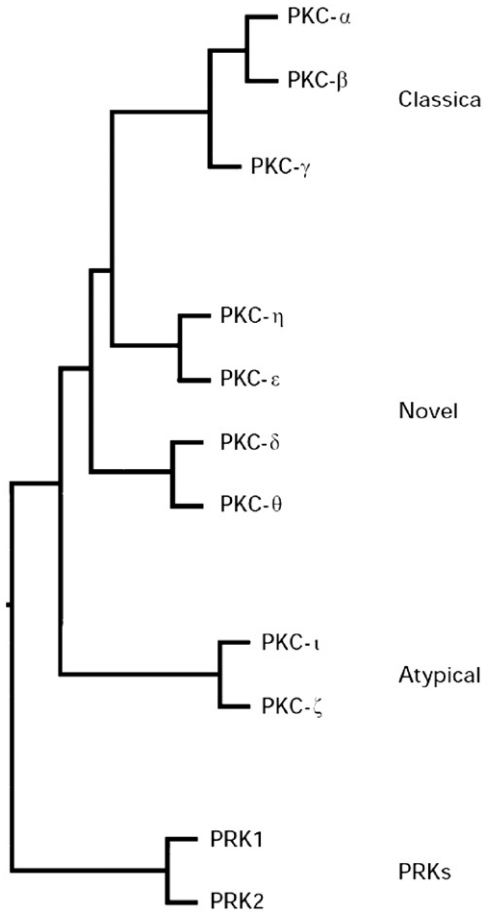

Fig. 4. A phylogenetic tree of protein kinase C with the traditional graph representation which is directly from the work of Mellor and Parker (1998).

In order to further illustrate the efficiency of our method we examine a dataset of beta-globins of 25 animals. The corresponding natural graphical representation is shown in Fig. 5. From this figure, we note that the 25 beta-globins are separated into two main clusters by level-3 line. One cluster contains mammalian beta-globins, and the other contains beta-globins from avian, fish, and reptilian species. Because the chimpanzee beta-globin sequence is the same as the human beta-globin sequence, these two proteins have the same natural vector. For the same reason, black bear and polar bear beta-globin sequences have the same natural vector. Here we should point out that to get an accurate evolutionary tree for organisms, the complete genome sequences may be necessary. In this paper, we focus on the protein sequences. Despite of this, this figure still clearly shows the similarity of these 25 protein sequences.

In the above examples, we only use the 2^{nd} order central moments in the natural vector in Eq. (4). That is, the vector is 60-dimensional with $N=2$: $\left\langle n_A,n_R,....,n_V,\mu_A,\mu_R,...,\mu_V,D_2^A,D_2^R,...,D_2^V\right\rangle$. Here we should emphasize that we do not need to calculate all the moments to determine the biological information of proteins. The 60-dimensional natural vectors have allowed us to obtain the stable classified results because higher central moments converge to 0 very quickly. In addition, we also check that the 60-dimensional natural vector gives a one-to-one map on this real dataset. Furthermore, our approach surpasses the multiple alignment method for both computational efficiency and biological results (See the comparison with MSA method in Supplementary material). Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors (Chou and Shen, 2009), we will make efforts in our future work to provide a web-server for the method presented. Currently the source codes and dataset of this work are freely available at http://homepages.math.uic.edu/~clyu/codes/NaturalVector.rar.

The proposed graphical representation of protein space is natural for two reasons. Firstly, the distances in the graphical representation are based on natural vectors. The natural vectors are naturally obtained from the original sequence, not any artificial parameters. Secondly, the graphical representation is naturally based on the minimum distances among any two proteins. The direction in the graph can show the closest elements of each element based on their distances. It is not based any tree-construction algorithms which usually produce not unique resulting trees for the same distance matrix (Buneman, 1974; Backeljau et al., 1996).

Our results suggest the following possible law of molecular biology: the normalized distribution of amino acids may determine the property of the protein. Our universe is a 4-dimensional space. Physicists speculate the Big Bang theory which states that all the galaxies must have originated from the same point. The protein universe is the collection of all known proteins. Here realizing the nature of the protein universe means that, we first find out how proteins are distributed in the protein universe, and then try to find the origin of protein phylogeny and explain the evolutionary process of proteins. Actually, for the Big Bang theory of protein universe, there have been many works on this field (Dokholyan et al., 2002; Povolotskaya and Kondrashov, 2010). Unfortunately, these works are not based on a concrete space of proteins. For the first time, we are able to construct the protein universe concretely and show that the protein universe is a 60-dimensional Euclidean space. In our protein universe, the galaxies may correspond to a superfamily of proteins while the star systems may correspond to a subfamily of proteins. So we can speculate the Big Bang theory for the protein universe may possibly be true. We can simply compute the centre of

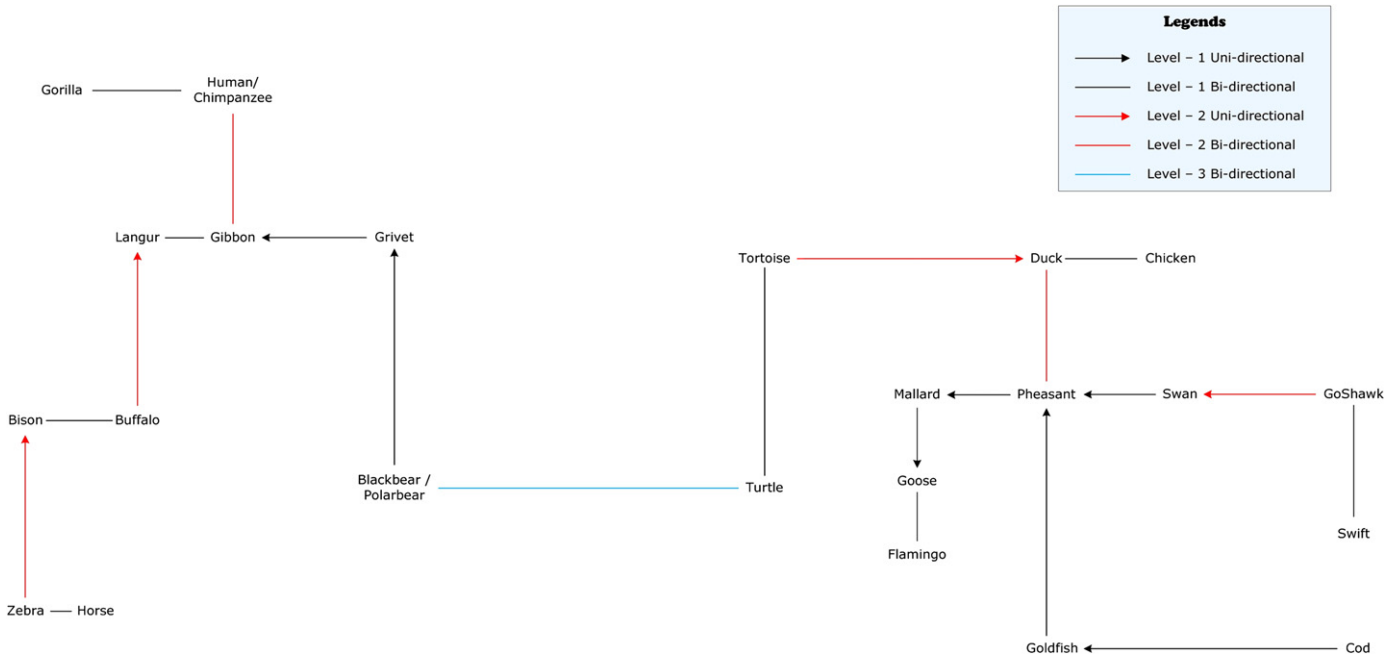

Fig. 5. The natural graphical representation of 25 animal beta-globins.

our protein universe. Of course, this conjecture needs further in-depth study. We will present the results in our future work.

Acknowledgments

We thank Dr. Max Benson and Dr Fenny Cheng for critically reading and editing the manuscript. This research is supported by the U.S. NSF grant DMS-1120824, China NSF grant 31271408, Tsinghua University, and Hong Kong University of Science and Technology. The authors have declared no conflict of interest.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2012.11.005.

References

Altschul, S.F., Carrol, R.J., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.
Backeljau, T., De Bruyn, L., De Wolf, H., Jordaens, K., Van Dongen, S., et al., 1996. Multiple UPGMA and neighbor-joining trees and the performance of some computer packages. Mol. Biol. Evol. 13 (2), 309–313.
Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., et al., 2004. The Pfam protein families database. Nucleic Acids Res. 32 (suppl 1), D138–D141, http://dx.doi.org/10.1093/nar/gkm960.
Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., et al., 2000. The protein data bank. Nucleic Acids Res. 28 (1), 235–242.
Buneman, P., 1974. A note on metric properties of trees. J. Combin. Theory Ser. B 17, 48–50.
Chen, C., et al., 2006. Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal. Biochem. 357, 116–121.
Chen, C., et al., 2009. Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. Protein Pept. Lett. 16, 27–31.
Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo amino acid composition. PROTEINS: Struct. Funct. Genet. (Erratum: ibid., Vol. 44, 60) 43, 246–255.
Chou, K.C., 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21, 10–19.
Chou, K.C., Shen, H.B., 2009. Review: recent advances in developing web-servers for predicting protein attributes. Nat. Sci. 2, 63–92.
Chou, K.C., 2009. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr. Proteomics 6, 262–274.
Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). J. Theor. Biol. 273, 236–247.
Corpet, F., Servant, F., Gouzy, J., Kahn, D., 2000. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. Nucleic Acids Res. 28, 267–269.
Deng, M., Yu, C., Liang, Q., He, R.L., Yau, S.S.-T., 2011. A novel method of characterizing genetic sequences: genome space with biological distance and applications. PLoS One 6 (3), e17293.
Ding, H., et al., 2009. Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. Protein Pept. Lett. 16, 351–355.
Ding, H., et al., 2011. Identify Golgi protein types with modified mahalanobis discriminant algorithm and pseudo amino acid composition. Protein Pept. Lett. 18, 58–63.
Ding, Y.S., Zhang, T.L., 2008. Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. Pattern Recognit. Lett. 29, 1887–1892.
Dokholyan, N.V., Shakhnovich, B., Shakhnovich, E.I., 2002. Expanding protein universe and its origin from the biological Big Bang. Proc. Natl. Acad. Sci. USA 99, 14132–14136.
Du, P., Cao, S., Li, Y., 2009. SubChlo: predicting protein subchloroplast locations with pseudo-amino acid composition and the evidence-theoretic K-nearest neighbor (ET-KNN) algorithm. J. Theor. Biol. 261, 330–335.
Esmaeili, M., et al., 2010. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. J. Theor. Biol. 263, 203–209.
Fan, G.L., Li, Q.Z., 2012. Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition. J. Theor. Biol. 304, 88–95.
Fang, Y., et al., 2008. Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. Amino Acids 34, 103–109.
Fitch, W.M., Margoliash, E., 1967. Construction of phylogenetic trees. Science 155, 279–284.
Gao, Q.B., et al., 2010. Improving discrimination of outer membrane proteins by fusing different forms of pseudo amino acid composition. Anal. Biochem. 398, 52–59.
Georgiou, D.N., et al., 2009. Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. J. Theor. Biol. 257, 17–26.
Gu, Q., et al., 2010. Prediction of G-Protein-coupled receptor classes in low homology using Chou's pseudo amino acid composition with approximate entropy and hydrophobicity patterns. Protein Pept. Lett. 17, 559–567.
Guo, J., et al., 2011. Predicting protein folding rates using the concept of Chou's pseudo amino acid composition. J. Comp. Chem. 32, 1612–1617.
Hayat, M., Khan, A., 2012. Discriminating outer membrane proteins with fuzzy K-nearest neighbor algorithms based on the general form of Chou's PseAAC. Protein Pept. Lett. 19, 411–421.
Holm, L., Sander, C., 1996. Mapping the protein universe. Science 273, 595–602.

Hu, L., et al., 2011. Using pseudo amino acid composition to predict protease families by incorporating a series of protein biological features. Protein Pept. Lett. 18, 552–558.
Hurley, J.H., Newton, A.C., Parker, P.J., Blumberg, P.M., Nishizuka, Y., 1997. Taxonomy and function of C1 protein kinase C homology domains. Protein Sci. 6, 477–480.
Jaroszewski, L., Li, Z., Krishna, S.S., Bakolitsa, C., Wooley, J., et al., 2009. Exploration of uncharted regions of the protein universe. PLoS Biol. 7 (9), e1000205, http://dx.doi.org/10.1371/journal.pbio.1000205.
Jiang, X., et al., 2008. Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. Protein Pept. Lett. 15, 392–396.
Jiang, X., et al., 2008. Using Chou's pseudo amino acid composition based on approximate entropy and an ensemble of AdaBoost classifiers to predict protein subnuclear location. Amino Acids 34, 669–675.
Kandaswamy, K.K., et al., 2010. Prediction of apoptosis protein locations with genetic algorithms and support vector machines through a new mode of pseudo amino acid composition. Protein Pept. Lett. 17, 1473–1479.
Koonin, E.V., 2007. Metagenomic sorcery and the expanding protein universe. Nat. Biotech. 25, 540–542.
Koonin, E.V., Wolf, Y.I., Karev, G.P., 2002. The structure of the protein universe and genome evolution. Nature 420, 218–223.
Ladunga, I., 1992. Phylogenetic continuum indicates galaxies in the protein universe: preliminary results on the natural group structures of proteins. J. Mol. Evol. 4, 358–375.
Levitt, M., 2009. Nature of the protein universe. Proc. Natl. Acad. Sci. USA 106, 11079–11084.
Li, F.M., Li, Q.Z., 2008. Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. Protein Pept. Lett. 15, 612–616.
Li, Z.C., et al., 2009. Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis. Amino Acids 37, 415–425.
Lin, H., Li, Q.Z., 2007. Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. J. Comp. Chemi. 28, 1463–1466.
Lin, H., 2008. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J. Theor. Biol. 252, 350–356.
Lin, H., et al., 2008. Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein Pept. Lett. 15, 739–744.
Lin, W.Z., et al., 2011. iDNA-Prot: identification of DNA binding proteins using random forest with grey model. PLos One 6 (9), e24756.
Lipman., D.J., Pearson, W.R., 1985. Rapid and sensitive protein similarity. Science 227, 1435–1441.
Mahdavi, A., Jahandideh, S., 2011. Application of density similarities to predict membrane protein types based on pseudo-amino acid composition. J. Theor. Biol. 276, 132–137.
Mohabatkar, H., 2010. Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein Pept. Lett. 17, 1207–1214.
Mohabatkar, H., et al., 2011. Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. J. Theor. Biol. 281, 18–23.
Mellor, H., Parker, P.J., 1998. The extended protein kinase C superfamily. Biochem. J. 332, 281–292.
Nanni, L., Lumini, A., 2008. Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. Amino Acids 34, 653–660.
Nanni, L., et al., 2011. Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of chou's pseudo amino acid composition and on evolutionary information. IEEE/ACM Trans. Comput. Biol. Bioinform. , http://dx.doi.org/10.1109/TCBB.2011.1117.
Povolotskaya, I.S., Kondrashov, F.A., 2010. Sequence space and the ongoing expansion of the protein universe. Nature 465, 922–926.
Qiu, J.D., et al., 2009. Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform. Anal. Biochem. 390, 68–73.
Qiu, J.D., et al., 2010. Using the concept of Chou's pseudo amino acid composition to predict enzyme family classes: an approach with support vector machine based on discrete wavelet transform. Protein Pept. Lett. 17, 715–722.
Qiu, J.D., et al., 2011. OligoPred: a webserver for predicting homo-oligomeric proteins by incorporating discrete wavelet transform into Chou's pseudo amino acid composition. J. Mol. Graph. Model 30, 129–134.
Sahu, S.S., Panda, G., 2010. A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. Comput. Biol. Chem. 34, 320–327.
Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4 (4), 406–425.
Smith, T.F., Waterman, M.S., 1981. Comparison of biosequences. Adv. Appl. Math. 2, 482–489.
Sokal, R., Michener, C., 1958. A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull. 38, 1409–1438.
Wang, Y.C., et al., 2010. Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature. Protein Pept. Lett. 17, 1441–1449.
Webb, B.L.J., Hirst, S.J., Giembycz, M.A., 2000. Protein kinase C isoenzymes: a review of their structure, regulation and role in regulating airways smooth muscle tone and mitogenesis. Br. J. Pharmacol. 130, 1433–1452.
Wu, Z.C., Xiao, X., Chou, K.C., 2010. 2D-MH: a web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. J. Theor. Biol. 267, 29–34.
Xiao, X., Chou, K.C., 2011. Using pseudo amino acid composition to predict protein attributes via cellular automata and others approaches. Curr. Bioinf. 6, 251–260.
Xiao, X., Wang, P., Chou, K.C., 2011. GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions. Mol. BioSyst. 7, 911–919.
Xiao, X., Wang, P., Chou, K.C., 2011. Cellular automata and its applications in protein bioinformatics. Curr. Protein Pept. Sci. 12, 508–519.
Xiao, X., Wang, P., Chou, K.C., 2012. iNR-PhysChem: a sequence-based predictor for identifying nuclear receptor and their subfamilies via physical–chemical property matrix. PLoS One 7 (2), e30869.
Yang, Z., 2006. Computational Molecular Evolution. Oxford University Press New York.
Yau, S.S.-T., Yu, C., He, R., 2008. A protein map and its application. DNA Cell. Biol. 27, 241–250.
Yu, C., et al., 2011. Protein map: an alignment-free sequence comparison method base on various properties of amino acids. Gene 486, 110–118.
Yu, C., Liang, Q., Yin, C., He, R.L., Yau, S.S.-T., 2010. A novel construction of genome space with biological geometry. DNA Res. 17, 155–168.
Yu, L., et al., 2010. SecretP: identifying bacterial secreted proteins by fusing new features into Chou's pseudoamino acid composition. J. Theor. Biol. 267, 1–6.
Zeng, Y.H., et al., 2009. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. J. Theor. Biol. 259, 366–372.
Zhang, G.Y., Fang, B.S., 2008. Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition. J. Theor. Biol. 253, 310–315.
Zhang, G.Y., et al., 2008. Predicting lipase types by improved Chou's pseudo-amino acid composition. Protein Pept. Lett. 15, 1132–1137.
Zhang, S.W., et al., 2008. Using Chou's pseudo amino acid composition to predict protein quaternary structure: a sequence-segmented PseAAC approach. Amino Acids 35, 591–598.
Zhang, S.W., et al., 2008. Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies. Amino Acids 34, 565–572.
Zhou, X.B., et al., 2007. Using Chou's amphiphilic pseudoamino acid composition and support vector machine for prediction of enzyme subfamily classes. J. Theor. Biol. 248, 546–551.
Zou, D., et al., 2011. Supersecondary structure prediction using Chou's pseudo amino acid composition. J. Comput. Chem. 32, 271–278.