DNA AND CELL BIOLOGY
Volume 27, Number 5, 2008

Pp. 241–250
DOI: 10.1089/dna.2007.0676

A Protein Map and Its Application

Stephen S.-T. Yau,[1,2] Chenglong Yu,[3] and Rong He[4]

Graphical representation of gene sequences provides a simple way of viewing, sorting, and comparing various gene structures. Here we first report a two-dimensional graphical representation for protein sequences. With this method, we constructed the moment vectors for protein sequences, and mathematically proved that the correspondence between moment vectors and protein sequences is one-to-one. Therefore, each protein sequence can be represented as a point in a map, which we call protein map, and cluster analysis can be used for comparison between the points. Sixty-six proteins from five protein families were analyzed using this method. Our data showed that for proteins in the same family, their corresponding points in the map are close to each other. We also illustrate the efficiency of this approach by performing an extensive cluster analysis of the protein kinase C family. These results indicate that this protein map could be used to mathematically specify the similarity of two proteins and predict properties of an unknown protein based on its amino acid sequence.

Introduction

MANY METHODS HAVE BEEN REPORTED to analyze the huge amounts of gene data. One of them is the graphical representation of gene sequences, which is a very powerful tool for visual comparison of gene sequences. Hamori (1985) first used a three-dimensional H curve to represent a gene sequence. Gates (1985) later published a two-dimensional graphical representation that is simpler than the H curve. However, Gates' graphical representation has high degeneracy. We reported previously a new two-dimensional graphical representation of gene sequences (Yau *et al.*, 2003), which has no circuit or degeneracy, so the correspondence between gene sequences and gene graphs is one-to-one. Lately, many graphical representation methods for gene sequences have been proposed (Randic *et al.*, 2003a, 2003b); however, the method to make a protein sequence graph has never been shown. Unlike dealing with a gene or DNA sequence, from only four nucleotides, dealing with a protein sequence, from 20 amino acids, is more complicated. Here we report that a protein or amino acid sequence can be graphically represented and a universal protein map can be generated. This protein map can be used to predict the properties of proteins whose functions are not yet determined. We have analyzed 66 proteins from 5 protein families and an exhaustive set of 127 proteins from the protein kinase C (PKC) family using this method, and found it to be a useful predictive tool.

Protein Sequence Graphical Representation

Following our previous work (Yau *et al.*, 2003), we construct a protein sequence graph on two quadrants of the Cartesian coordinate system. The vectors corresponding to the 20 amino acids are lying in the line segment whose x-coordinate value is 1 and whose y-coordinate values are between -1 and 1. The y-coordinate values of the 20 amino acid vectors are all distinct. The ordering of these y-coordinate values is based on amino acid hydrophobicity scale values (Fauchere and Pliska, 1983) because amino acid hydrophobicity plays an important role in protein folding. The 12 amino acids with positive hydrophobicity scale values were assigned in the first quadrant, and the difference of y-coordinate values between two amino acids next to each other is 1/13. Because the hydrophobicity scale value of Gly is zero, its y-coordinate value was assigned to be zero. The other seven amino acids with negative hydrophobicity scale values were assigned in the fourth quadrant, and the difference of y-coordinate values between two amino acids next to each other is 1/8. Thus, all y-coordinates of 20 amino acids are less than 1 and more than -1. The y-coordinates for 20 amino acids are listed in Table 1, and 20 vectors are shown in Figure 1. The points in the graphical representation are obtained by the sum of vectors representing amino acids in the sequence. In Figure 2, we give the graphical representation of the first 10 vectors of human beta-globin amino acid sequence on the vector system shown in Figure 1, and the graphical representation of the whole

[1]Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, Illinois.
[2]Institute of Mathematics, East China Normal University, Shanghai, China.
[3]Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong Kong.
[4]Department of Pharmacology, University of Illinois at Chicago, Chicago, Illinois.

Table 1. Hydrophobicity Scale Values and y-Coordinate of 20 Amino Acids

Amino acid	*Hydrophobicity scale*	*y-Coordinate*
Trp (W)	+2.25	12/13
Ile (I)	+1.80	11/13
Phe (F)	+1.79	10/13
Leu (L)	+1.70	9/13
Cys (C)	+1.54	8/13
Met (M)	+1.23	7/13
Val (V)	+1.22	6/13
Tyr (Y)	+0.96	5/13
Pro (P)	+0.72	4/13
Ala (A)	+0.31	3/13
Thr (T)	+0.26	2/13
His (H)	+0.13	1/13
Gly (G)	0	0
Ser (S)	−0.04	−1/8
Gln (Q)	−0.22	−2/8
Asn (N)	−0.60	−3/8
Glu (E)	−0.64	−4/8
Asp (D)	−0.77	−5/8
Lys (K)	−0.99	−6/8
Arg (R)	−1.01	−7/8

human beta-globin amino acid sequence based on the same vector system is also shown in Figure 3. This protein sequence graphical representation has no circuits or degeneracy, and the correspondence between the sequence and the graphical curve can be mathematically proven to be one-to-one, as follows:

To prove there is no circuit or degeneracy in our two-dimensional graphical representation, we assume that (1) the total number of amino acids is N; (2) the number of amino acids R, K, D, E, N, Q, S, G, H, T, A, P, Y, V, M, C, L, F, I, and W is r, k, d, e, n, q, s, g, h, t, a, p, y, v, m, c, l, f, i, and w, respectively. Therefore, we have $r+k+d+e+n+q+s+g+h+t+a+p+y+v+m+c+l+f+i+w=N$.

If we further assume rR, kK, dD, eE, nN, qQ, sS, gG, hH, tT, aA, pP, yY, vV, mM, cC, lL, fF, iI, and wW form a circuit, the following equation will hold:

$$r\left(1,-\frac{7}{8}\right)+k\left(1,-\frac{6}{8}\right)+d\left(1,-\frac{5}{8}\right)+e\left(1,-\frac{4}{8}\right)+n\left(1,-\frac{3}{8}\right)+q\left(1,-\frac{2}{8}\right)+s\left(1,-\frac{1}{8}\right)+g(1,0)+h\left(1,\frac{1}{13}\right)+t\left(1,\frac{2}{13}\right)+a\left(1,\frac{3}{13}\right)+p\left(1,\frac{4}{13}\right)+y\left(1,\frac{5}{13}\right)+v\left(1,\frac{6}{13}\right)+m\left(1,\frac{7}{13}\right)+c\left(1,\frac{8}{13}\right)+l\left(1,\frac{9}{13}\right)+f\left(1,\frac{10}{13}\right)+i\left(1,\frac{11}{13}\right)+w\left(1,\frac{12}{13}\right)=(0,0)$$

The sum of x-coordinates indicates that $r+k+d+e+n+q+s+g+h+t+a+p+y+v+m+c+l+f+i+w=0$. It follows that $r=k=d=e=n=q=s=g=h=t=a=p=y=v=m=c=l=f=i=w=0$ as the number of amino acids is a nonnegative number. Therefore, no circuit exists in the graphical representation in a nontrivial case where $N>0$.

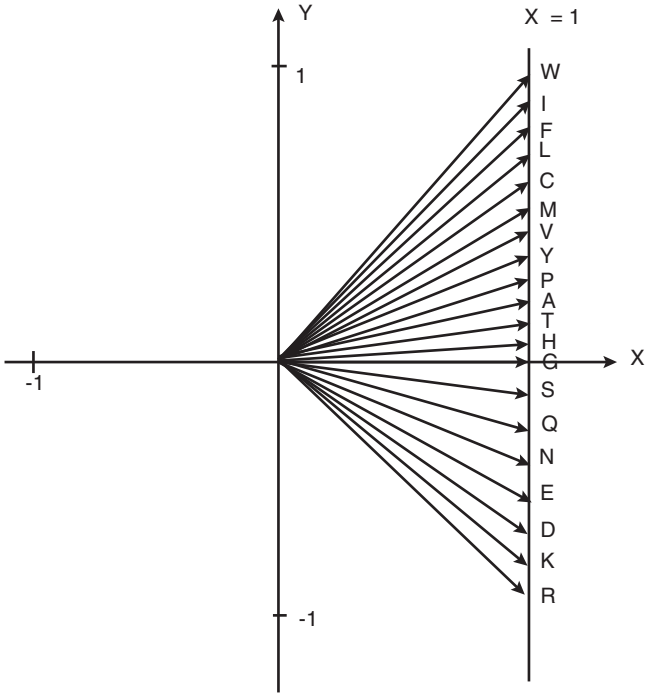

FIG. 1. Amino acid vector system based on Table 1.

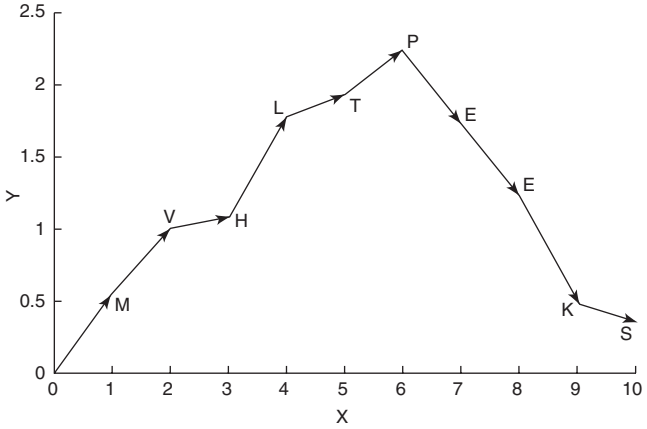

FIG. 2. Graphical representation of first 10 amino acids of human beta-globin sequence based on the vector system of Figure 1.

A Moment Vector for Protein Sequences

Given the graphical curve of a protein sequence that can be represented by a sequence of points $(1, y_1)$, $(2, y_2), \ldots, (n, y_n)$, we can compute a sequence of numbers $1-y_1$, $2-y_2, \ldots, n-y_n$. Conversely, if we know the sequence of numbers $1-y_1$, $2-y_2, \ldots, n-y_n$, we can recover the graph $(1,y_1)$, $(2,y_2), \ldots, (n,y_n)$. Therefore, we would like to find a sequence of numbers, each of which uses the global information of the sequence of numbers $1-y_1$, $2-y_2, \ldots, n-y_n$ in such a way that this new sequence of numbers determines

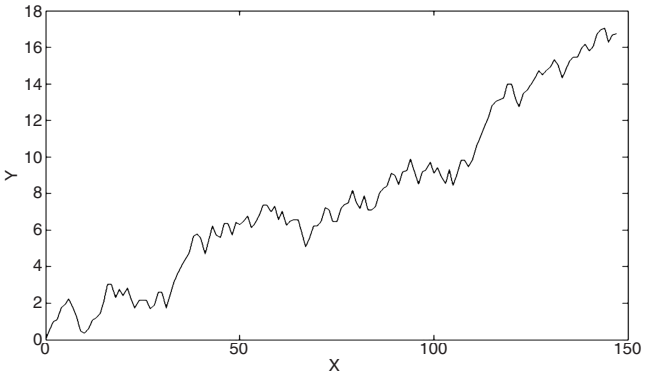

FIG. 3. Graphical representation of human beta-globin amino acid sequence based on the vector system of Figure 1.

and is determined by the sequence of numbers $1-y_1$, $2-y_2, \ldots, n-y_n$. For this purpose, we decided to use moments to characterize a protein graphical curve. The moments are defined as follows:

$$M_j = \sum_{i=1}^{n} \frac{(x_i - y_i)^j}{n^j},\ j = 1, 2, \ldots, n,$$

where n is the number of amino acids contained in a protein sequence, and (x_i, y_i) represents the position of the ith amino acid in the protein graphical curve. According to this definition, each protein sequence has an n-dimensional moment vector $(M_1, M_2, \ldots, M_n)$ associated with it. We should emphasize that two or three moments will already characterize the protein sufficiently well (as it is demonstrated later) and that this is a much simpler representation than the graph with hundreds of points (147 for human beta globin).

Theorem: Consider the set of protein sequences having the fixed number (n) of amino acids. Then the correspondence between a protein sequence and its n-dimensional moment vector $(M_1, M_2, \ldots, M_n)$ is one-to-one.

Proof: We have demonstrated that the correspondence between a protein sequence and its graphical curve is one-to-one (Yau *et al.*, 2003). To prove the theorem, we will need to prove that the correspondence between a protein graphical curve and its moment vector is one-to-one.

By the definition, one protein sequence graph has an n-dimensional moment vector $(M_1, M_2, \ldots, M_n)$. Hence, we need to demonstrate that from any given protein moment vector, we can recover the protein curve, which means all (x_i, y_i) $(i = 1, 2, \ldots, n)$ can be recovered from any given protein moment vector.

x_i is the x-coordinate value of ith amino acid on a protein graph. Based on our assignment, x_i should be equal to i. y_i is the y-coordinate value of ith amino acid on a protein graph. The next step is to obtain y_i from moment vector. Let $z_i = x_i - y_i$, then the moments can be simplified as:

$$M_j = \sum_{i=1}^{n} \frac{z_i^j}{n^j},\ j = 1, 2, \ldots, n.$$

To solve for z_i, let $\delta_j = M_j n^j$, then the δ_j can be obtained by M_j and n. δ_j and z_i have the relation given below:

$$\begin{cases} \delta_1 = z_1 + z_2 + \cdots + z_n \\ \delta_2 = z_1^2 + z_2^2 + \cdots + z_n^2 \\ \qquad \cdots \qquad \cdots \\ \delta_n = z_1^n + z_2^n + \cdots + z_n^n \end{cases}$$

The $z_1, z_2, \ldots, z_n$ can be the roots of a symmetric polynomial $a_0 + a_1 z + a_2\ z^2 + \ldots + a_n\ z^n = (z - z_1)\ (z - z_2) \ldots (z - z_n)$. By using Newton's identities (Jacobson, 1974):

$$\delta_d - s_1\delta_{d-1} + \cdots + (-1)^{d-1} s_{d-1}\delta_1 + (-1)^d d s_d = 0,$$

where $d = 1, 2, \ldots, n$; s_d is the elementary symmetric polynomials in $z_1, z_2, \ldots, z_n$; a_i can be obtained by δ_j as shown below:

$$\begin{cases} a_n = 1 \\ a_{n-1} = (-1)\delta_1 \\ a_{n-2} = \dfrac{1}{2}(\delta_1^2 - \delta_2) \\ a_{n-3} = (-1)^3 \dfrac{1}{6}(\delta_1^3 - 3\delta_1\delta_2 + 2\delta_3) \\ a_{n-4} = \dfrac{1}{24}(\delta_1^4 - 6\delta_1^2\delta_2 + 3\delta_2^2 + 8\delta_1\delta_3 - 6\delta_4) \\ \vdots \\ \vdots \end{cases}$$

As a result, the coefficients of the symmetric polynomial $a_0 + a_1 z + a_2\ z^2 + \ldots + a_n\ z^n = (z - z_1)\ (z - z_2) \ldots (z - z_n)$ can be confirmed, and the set of all roots can be obtained. Next we need to identify each root $z_1, z_2, \ldots, z_n$.

Because the y-coordinate values of all 20 amino acids are between -1 and 1, $x_i - y_i$ is greater than zero. As we have defined that the position of kth amino acid on a graph is (x_k, y_k) or (k, y_k), the position of $(k+1)$th amino acid on a graph (x_{k+1}, y_{k+1}) can be represented as $(k+1, y_k + u_{k+1})$, where u_{k+1} may be any of y-coordinate value of these 20 amino acids. Thus, $z_{k+1} = x_{k+1} - y_{k+1} = (k+1) - (y_k + u_{k+1}) = (k - y_k) + (1 - u_{k+1}) > (k - y_k)$. Because $z_k = x_k - y_k = k - y_k$, z_{k+1} must be greater than z_k. As a consequence, z_i is strictly increasing and each root can be identified by this property, which means each value of y_i can be obtained. With all (x_i, y_i), a protein graph can be recovered.

Therefore, we have successfully proven that the correspondence between a protein sequence and its moment vector obtained from its sequence graph is one-to-one.

Protein Map and Cluster Analysis

By using moments of our protein graphical curve, we change beta-globin sequences of human, gorilla, cod, duck, chicken, and tortoise (the length of these sequences are 147) into 147-dimensional moment vectors. In Table 2, we give the distances between human and the other five species for some different dimensional moment vectors. According to

Table 2. Distances between Beta-Globin Sequences of Human and Other Five Species for Different Dimensional Moment Vectors

	Gorilla	*Cod*	*Duck*	*Chicken*	*Tortoise*
2-dim	0.0667	1.9216	4.6743	4.4788	3.5836
3-dim	0.0927	2.3493	5.5714	5.2891	4.3612
4-dim	0.1141	2.6188	6.1769	5.8255	4.9223
5-dim	0.1314	2.7961	6.6031	6.1984	5.3410
10-dim	0.1761	3.1439	7.5224	6.9921	6.3494
100-dim	0.1947	3.2472	7.8132	7.2425	6.7329
147-dim	0.1947	3.2472	7.8132	7.2425	6.7329

this table, we find that the first two or three moments are most important because when higher moments are included the relationship of being close or further away remains unchanged. For example, distances between human and gorilla are always the smallest for any dimensional moment vector.

Thus, we can use the first two components of the moment vector (M_1, M_2) of a protein sequence graph to represent a protein as a point in a two-dimensional space and generate a universal protein map. Using the distance between two points as an index for comparison, we can perform cluster analysis for protein sequences on this protein map. If two sequences are similar, the distance between two corresponding points should be small. Therefore, we may use this universal protein map to predict properties of newly found proteins by performing clustering analysis.

Fifty beta-globin sequences of different species were extracted from Swiss-Prot (http://au.expasy.org/). Using the amino acid vectors shown in Figure 1, the sequence graphs of these beta-globins were obtained. We used our two-dimensional moment vector (M_1, M_2) system to characterize these 50 protein graphical curves and calculated the 50 points shown in Figure 4. From this figure, we note that these 50 beta-globins are separated into two main clusters. 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 two-dimensional moment vector. For the same reason, dog and coyote beta-globin sequences have the same moment vector, as do black bear and polar bear beta-globin sequences. Figure 4 also shows that the distances between beta-globin sequences from several primatal species (human, grivet, gorilla, langur, gibbon, and chimpanzee) are very small, and they form a subcluster.

We have also developed a three-dimensional moment vector (M_1, M_2, M_3) system to characterize graphical curves of these 50 proteins using the first three moments of their protein graph. As a result, these 50 beta-globins can be represented as 50 points in a three-dimensional map. By computing distances between these points, we did cluster analysis. The data are shown in a hierarchical tree in Figure 5, which indicates the evolutionary relationships between these proteins.

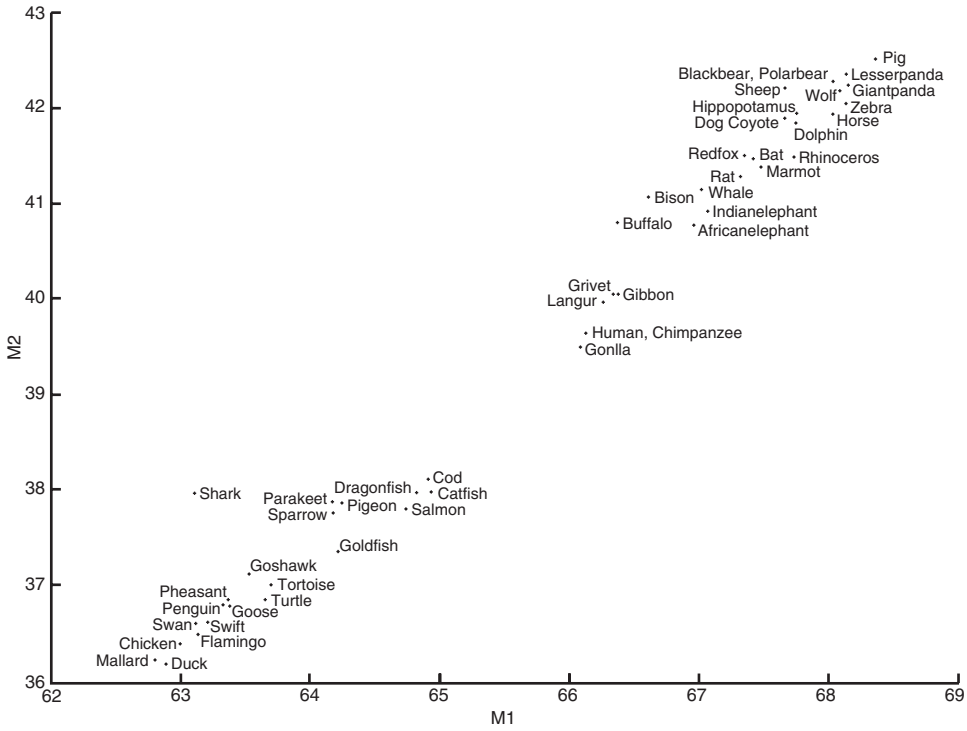

FIG. 4. Two-dimensional moment vector points of beta-globin sequences of 50 species.

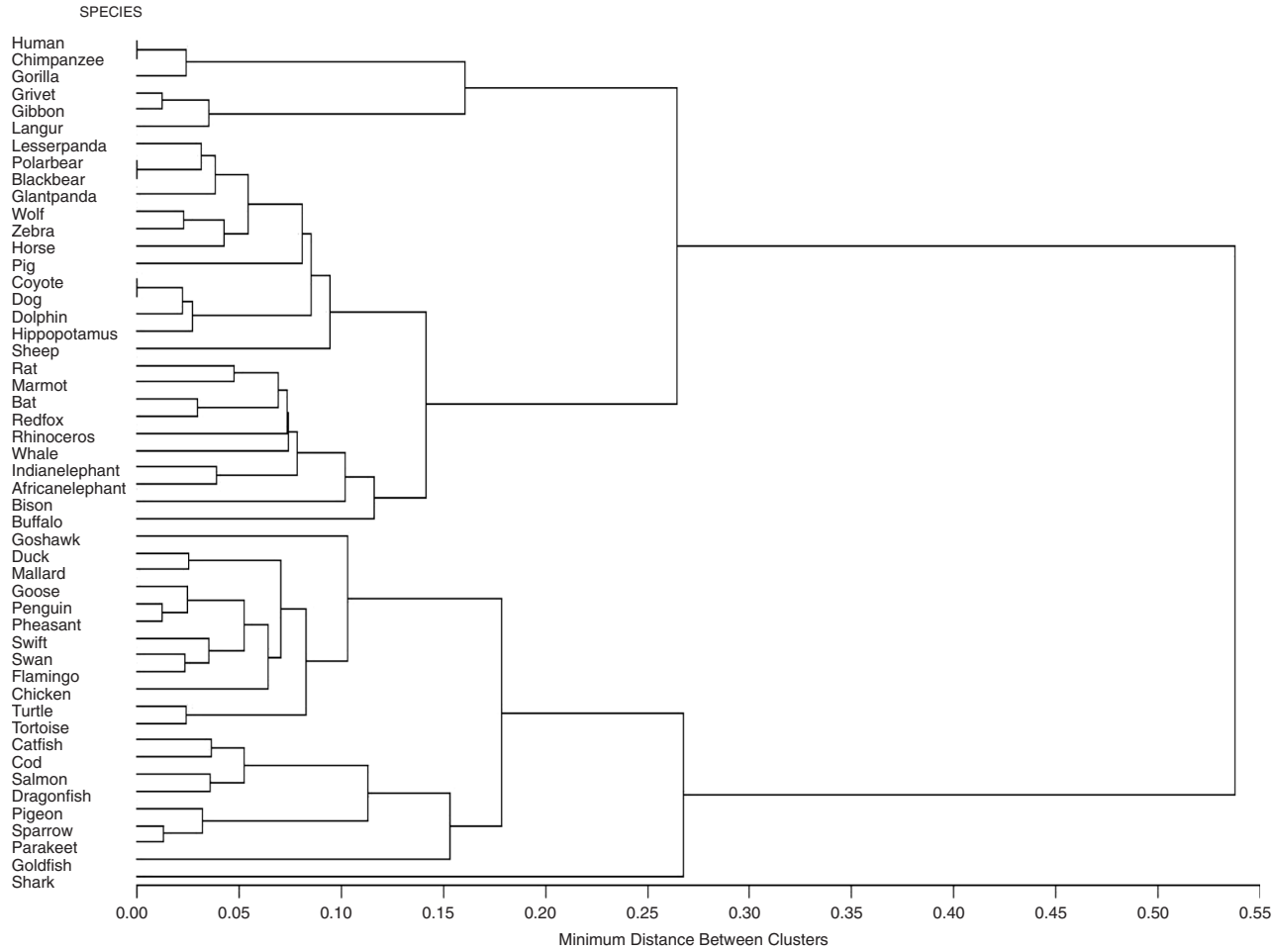

FIG. 5. Clustering hierarchical tree of beta-globin sequences of 50 species by three-dimensional moment vector.

To demonstrate the effectiveness of our clustering approach, we extracted another 16 proteins in 4 families (insulin family, catalase family, CCN family, and thiolase family) from Swiss-Prot. By using our two-dimensional moment vector system, these 16 proteins and the 50 beta-globins are represented as 66 points shown in Figure 6. The clusters of five families are identified in the figure.

We also applied our protein map to a set of 127 proteins from the PKC family (Table 3). PKCs can be divided into two functional parts, the regulatory and catalytic domains. The regulatory domain is the principal determinant of classification, as the catalytic domain tends to be highly conserved. Since PKCs were originally identified in mammals, their classification system is strongly characterized by varieties of PKCs found in mammals. In particular, mammalian PKCs are generally 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 and ζ). A controversial group of potential PKCs including PKCν and PKCμ/PKD (protein kinase D) (mouse) (Webb *et al.*, 2000) has similar regulatory domains to PKCs, but catalytic domains are more similar to the myosin light-chain kinase of *Dictostelium* (Hurley *et al.*, 1997). Fungi have PKC homologs that characteristically contain more residues than mammalian PKCs with significantly different regulatory domains but similar catalytic domains. There are also PKC-related kinases (PRKs) that are found in many animals and have features similar to fungal PKCs. Thus, we identified six categories of structural architecture for PKCs and PKC-related molecules: cPKC, nPKC, aPKC, PKCmu (ν, μ, and D2 types), PKC1 (fungal PKCs), and PRK. We used our two-dimensional protein map to characterize the regulatory domains of these 127 proteins from PKC family and calculated the 127 points in Figure 7. Four clusters, PKC, PKCmu, PKC1, and PRK, can be seen in this figure. To study the PKC cluster clearly, we expanded it and obtained Figure 8. In this figure, we classified these points into three categories (cPKC, nPKC, and aPKC). In Table 3, we list our clustering results and made a comparison with the categories established in the literature.

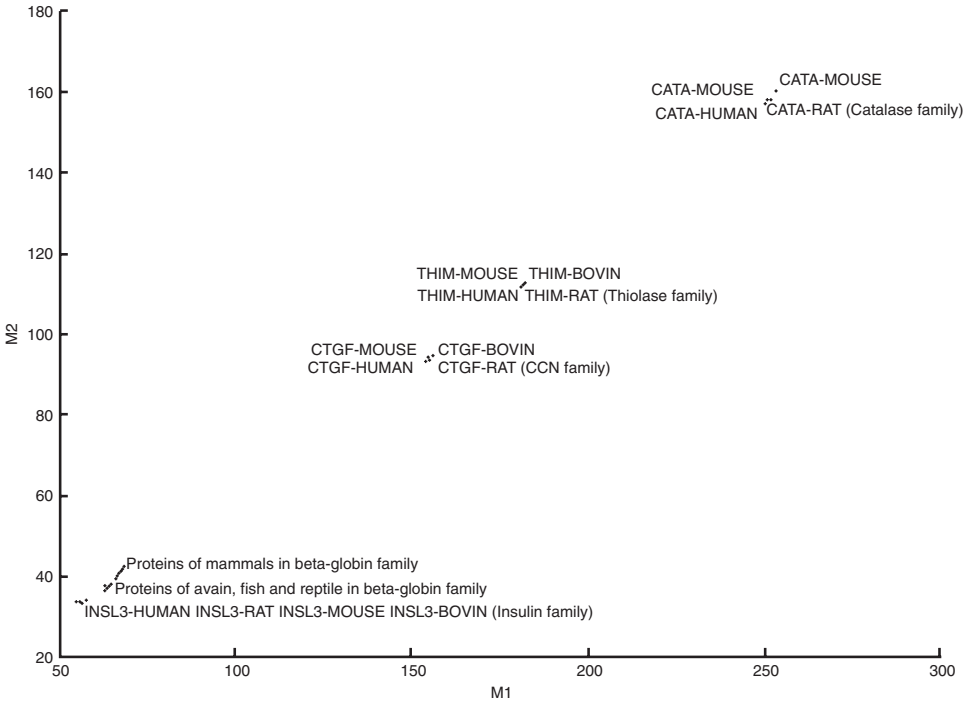

FIG. 6. Two-dimensional moment vector points of 66 proteins from 5 families.

Our clustering of the regulatory domains corresponds closely to the traditional classification of PKCs. These results confirm a compact relationship between nPKCs and cPKCs. A group of cPKCs including *Sycon raphanus* 7, *Apis mellifera* 113, *Caenorhabditis elegans* 45, *Bombyx mori* 64, *Drosophila melanogaster* 18, *Drosophila melanogaster* 28, and *Danio rerio* 124, which is very close to nPKCs in Figure 8, is not mammalian. This implies that nPKCs and cPKCs may be less differentiated outside of mammals. Moreover, our clustering of fungal PKCs suggests a division between yeasts that have

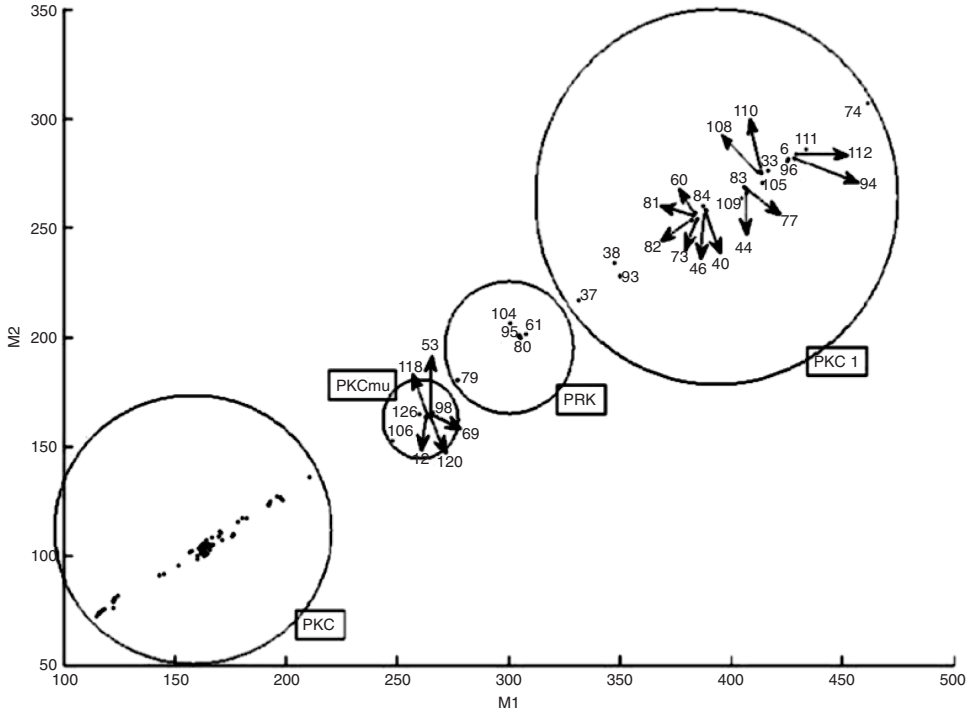

FIG. 7. Two-dimensional moment vector points of regulatory domains of 127 proteins from PKC family shown in Table 3.

Table 3. Clustering Results of 127 Proteins from PKC Family by Using Protein Map

Number	*Species*	*Sequence ID*	*Category*	*Architecture*
1	*Gallus gallus*	NP001006133	delta	nPKC
2	*Canis familiaris*	NP001008716	delta	nPKC
3	*Xenopus tropicalis*	NP001012707	iota	aPKC
4	*Hydra vulgaris*	O01715	cPKC	cPKC
5	*Oryctolagus cuniculus*	O19111	zeta	aPKC
6	*Cochliobolus heterostrophus*	O42632	PKC1	PKC1
7	*Sycon raphanus*	O61224	cPKC	cPKC
8	*Sycon raphanus*	O61225	nPKC	nPKC
9	*Suberites domuncula*	O62567	cPKC	cPKC
10	*Suberites domuncula*	O62569	nPKC	nPKC
11	*Calliphora vicina*	O76850	cPKC	cPKC
12	*Homo sapiens*	O94806	mu	PKCmu
13	*Rhabdocalyptus dawsoni*	O96942	cPKC	cPKC
14	*Geodia cydonium*	O96997	cPKC	cPKC
15	*Bos taurus*	P04409	alpha	cPKC
16	*Bos taurus*	P05126	beta	cPKC
17	*Homo sapiens*	P05129	gamma	cPKC
18	*Drosophila melanogaster*	P05130	cPKC	cPKC
19	*Rattus norvegicus*	P05696	alpha	cPKC
20	*Homo sapiens*	P05771	beta	cPKC
21	*Oryctolagus cuniculus*	P05772	beta	cPKC
22	*Rattus norvegicus*	P09215	delta	nPKC
23	*Rattus norvegicus*	P09216	epsilon	nPKC
24	*Rattus norvegicus*	P09217	zeta	aPKC
25	*Oryctolagus cuniculus*	P10102	alpha	cPKC
26	*Oryctolagus cuniculus*	P10829	gamma	cPKC
27	*Oryctolagus cuniculus*	P10830	epsilon	nPKC
28	*Drosophila melanogaster*	P13677	cPKC	cPKC
29	*Mus musculus*	P16054	epsilon	nPKC
30	*Homo sapiens*	P17252	alpha	cPKC
31	*Mus musculus*	P20444	alpha	cPKC
32	*Mus musculus*	P23298	eta	nPKC
33	*Saccharomyces cerevisiae*	P24583	PKC1	PKC1
34	*Homo sapiens*	P24723	eta	nPKC
35	*Mus musculus*	P28867	delta	nPKC
36	*Caenorhabditis elegans*	P34885	nPKC	nPKC
37	*Schizosaccharomyces pombe*	P36582	PKC1	PKC1
38	*Schizosaccharomyces pombe*	P36583	PKC1	PKC1
39	*Homo sapiens*	P41743	iota	aPKC
40	*Candida albicans*	P43057	PKC1	PKC1
41	*Mus musculus*	P63318	gamma	cPKC
42	*Rattus norvegicus*	P68403	beta	cPKC
43	*Mus musculus*	P68404	beta	cPKC
44	*Neurospora crassa*	P87253	PKC1	PKC1
45	*Caenorhabditis elegans*	P90980	cPKC	cPKC
46	*Aspergillus niger*	Q00078	PKC1	PKC1
47	*Mus musculus*	Q02111	theta	nPKC
48	*Homo sapiens*	Q02156	epsilon	nPKC
49	*Mus musculus*	Q02956	zeta	aPKC
50	*Homo sapiens*	Q04759	theta	nPKC
51	*Homo sapiens*	Q05513	zeta	aPKC
52	*Homo sapiens*	Q05655	delta	nPKC
53	*Homo sapiens*	Q15139	mu	PKCmu
54	*Aplysia californica*	Q16974	cPKC	cPKC
55	*Aplysia californica*	Q16975	nPKC	nPKC
56	*Caenorhabditis elegans*	Q19266	aPKC	aPKC
57	*Lytechinus pictus*	Q25378	cPKC	cPKC
58	*Xenopus tropicalis*	Q28EN9	iota	aPKC
59	*Mus musculus*	Q2NKI4	cPKC	cPKC
60	*Aspergillus oryzae*	Q2U6A7	PKC1	PKC1
61	*Mus musculus*	Q3UEA6	PRK	PRK
62	*Xenopus laevis*	Q498G7	nPKC	nPKC
63	*Bombyx mori*	Q4AED5	aPKC	aPKC

(*continued*)

TABLE 3. (*CONTINUED*)

Number	*Species*	*Sequence ID*	*Category*	*Architecture*
64	*Bombyx mori*	Q4AED6	cPKC	cPKC
65	*Macaca fascicularis*	Q4R4U2	cPKC	cPKC
66	*Pongo pygmaeus*	Q5R4K9	aPKC	aPKC
67	*Danio rerio*	Q5TZD4	nPKC	nPKC
68	*Mus musculus*	Q62074	iota	aPKC
69	*Mus musculus*	Q62101	mu	PKCmu
70	*Rattus norvegicus*	Q64617	eta	nPKC
71	*Schistosoma mansoni*	Q69G16	cPKC	cPKC
72	*Xenopus laevis*	Q6AZF7	cPKC	cPKC
73	*Debaryomyces hansenii CBS767*	Q6BI27	PKC1	PKC1
74	*Yarrowia lipolytica*	Q6C292	PKC1	PKC1
75	*Xenopus laevis*	Q6DCJ8	delta1	nPKC
76	*Rattus norvegicus*	Q6DUV1	epsilon	nPKC
77	*Candida glabrata CBS138*	Q6FJ43	PKC1	PKC1
78	*Xenopus laevis*	Q6GNZ7	nPKC	nPKC
79	*Homo sapiens*	Q6P5Z2	PRK	PRK
80	*Rattus norvegicus*	Q6P748	PRK	PRK
81	*Cryptococcus neoformans* var.	Q6UB96	PKC1	PKC1
82	*Cryptococcus neoformans* var.	Q6UB97	PKC1	PKC1
83	*Eremothecium gossypii*	Q75BT0	PKC1	PKC1
84	*Aspergillus nidulans*	Q76G54	PKC1	PKC1
85	*Xenopus laevis*	Q7LZQ8	cPKC	cPKC
86	*Xenopus laevis*	Q7LZQ9	cPKC	cPKC
87	*Anopheles gambiae*	Q7QCP8	nPKC	nPKC
88	*Danio rerio*	Q7SY24	cPKC	cPKC
89	*Xenopus laevis*	Q7SZH7	delta2	nPKC
90	*Xenopus laevis*	Q7SZH8	delta1	nPKC
91	*Danio rerio*	Q7T2C5	cPKC	cPKC
92	*Homo sapiens*	Q86XJ6	delta	nPKC
93	*Pichia pastoris*	Q86ZV2	PKC1	PKC1
94	*Leptosphaeria maculans*	Q873Y9	PKC1	PKC1
95	*Homo sapiens*	Q8IUV5	PRK	PRK
96	*Kluyveromyces lactis*	Q8J213	PKC1	PKC1
97	*Takifugu rubripes*	Q8JFZ9	cPKC	cPKC
98	*Mus musculus*	Q8K1Y2	mu	PKCmu
99	*Mus musculus*	Q8K2K8	eta	nPKC
100	*Limulus polyphemus*	Q8MXB6	nPKC	nPKC
101	*Homo sapiens*	Q8NE03	eta	nPKC
102	*Danio rerio*	Q90XF2	iota	aPKC
103	*Xenopus laevis*	Q91569	iota	aPKC
104	*Xenopus* sp.	Q91948	PRK	PRK
105	*Hypocrea jecorina*	Q99014	PKC1	PKC1
106	*Homo sapiens*	Q9BZL6	D2	PKCmu
107	*Drosophila melanogaster*	Q9GSZ3	aPKC	aPKC
108	*Blumeria graminis*	Q9HF10	PKC1	PKC1
109	*Tuber borchii*	Q9HGK8	PKC1	PKC1
110	*Botryotinia fuckeliana*	Q9UVJ5	PKC1	PKC1
111	*Sporothrix schenckii*	Q9Y792	PKC1	PKC1
112	*Magnaporthe grisea*	Q9Y7C1	PKC1	PKC1
113	*Apis mellifera*	XM391874	cPKC	cPKC
114	*Rattus norvegicus*	XP001066028	theta	nPKC
115	*Macaca mulatta*	XP001116804	gamma	cPKC
116	*Pan troglodytes*	XP001147999	theta	nPKC
117	*Bos taurus*	XP001250401	delta	nPKC
118	*Rattus norvegicus*	XP234108	PKCmu	PKCmu
119	*Gallus gallus*	XP421417	eta	nPKC
120	*Canis familiaris*	XP540151	PKCmu	PKCmu
121	*Canis familiaris*	XP541432	gamma	cPKC
122	*Bos taurus*	XP583587	epsilon	nPKC
123	*Bos taurus*	XP602125	gamma	cPKC
124	*Danio rerio*	XP683138	nPKC	cPKC
125	*Canis familiaris*	XP849292	theta	nPKC
126	*Canis familiaris*	XP851386	PKCmu	PKCmu
127	*Canis familiaris*	XP851861	epsilon	nPKC

The number column provides identifying numbers used in Figures 7 and 8. Sequence IDs refer to NCBI or SwissProt accession numbers. Categories are those established in the literature, whereas Architectures are our clustering results.

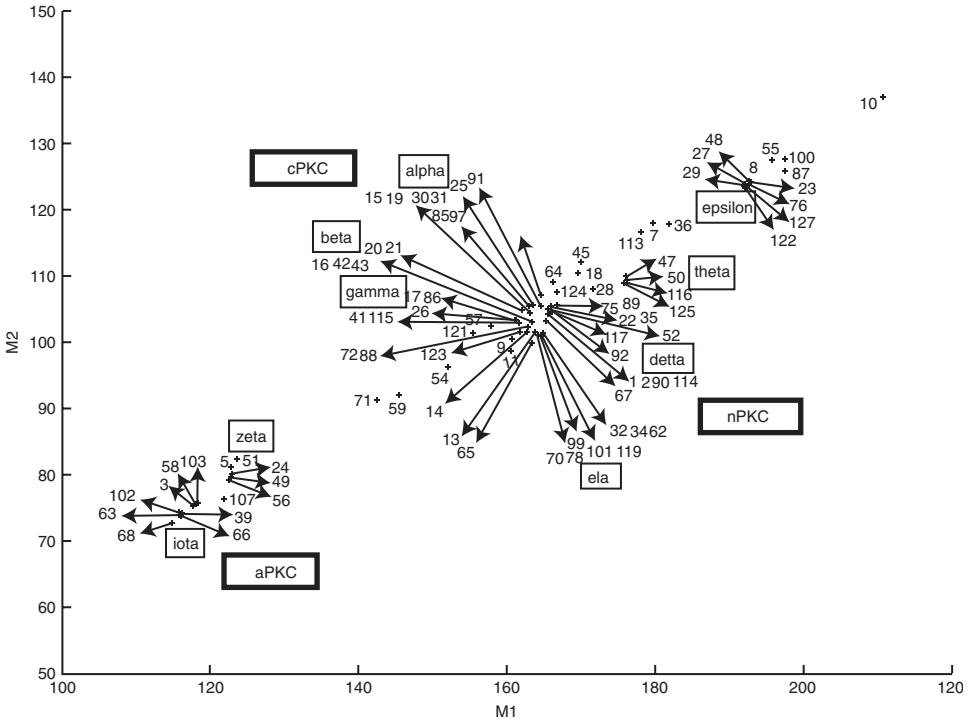

FIG. 8. Two-dimensional moment vector points of regulatory domains of proteins from PKC cluster in Figure 7.

specific yeast-like growth patterns and filamentous fungi that display mycelial growth patterns. This distinction is supported by the evidence of fungal PKCs involvement in control of growth patterns (Aquino-Piñero and Valle, 2002). We believe that our protein map provides a rapid and accurate way of identifying the likely category of a PKC family member.

Discussion and Conclusion

In this paper, we report a two-dimensional graphical representation for protein sequences. A moment vector system to represent a protein sequence is introduced, and the correspondence between a protein sequence and its moment vector is mathematically proven to be one-to-one. With this moment vector system, a protein can be represented as a point on a map, which is called protein map. Proteins with similar properties plot close together. Thus, the protein map allows us to analyze protein sequences using clustering methods and to predict properties of newly found proteins. This method will provide a new tool for protein functional studies.

To represent a protein sequence as a two-dimensional graph, we assigned 20 vectors to 20 amino acids. The x-coordinate value of a vector was chosen greater than zero to avoid circuit or degeneracy. The y-coordinate value of a vector was assigned based on amino acid hydrophobicity scale values. Because many other amino acid physicochemical properties, such as polarity (Grantham, 1974) and refractivity (Jones, 1975), should be considered, further studies will be needed to decide which combination of amino acid properties is most biologically meaningful for determining its y-coordinate value.

With our moment vector system, a protein sequence graph can be represented as a point in a two-dimensional or three-dimensional map depending on whether first two or first three moments of this sequence graph are used as its moment vector. By applying clustering methods to either of these maps, protein sequences, protein domains, and even arbitrary amino acid sequences can be efficiently analyzed.

Acknowledgment

We thank Dr. Max Benson for critically reading and editing the manuscript. We also thank Kareem Carr for providing us the data of PKC family.

References

Aquino-Piñero, E., and Valle, N.R.-D. (2002). Characterization of a protein kinase C gene in *Sporothrix schenckii* and its expression during the yeast-to-mycelium transition. Med Mycol **40,** 185–199.

Fauchere, J., and Pliska, V. (1983). Hydrophobic parameters of amino-acid side-chains from the partitioning of N-acetyl-amino-acid amides. Eur J Med Chem **18,** 369–375.

Gates, M.A. (1985). Simpler DNA sequence representations. Nature **316,** 219.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. Science **185,** 862–864.

Hamori, E. (1985). Novel DNA sequence representation. Nature **314,** 585–586.

Hurley, J.H., Newton, A.C., Parker, P.J., Blumberg, P.M., and Nishizuka, Y. (1997). Taxonomy and function of C1 protein kinase C homology domains. Protein Sci **6,** 477–480.

Jacobson, N. (1974). *Basic Algebra*, vol. 1. India: Hindustan Publishing Corporation, p. 135.

Jones, D.D. (1975). Amino acid properties and side-chain orientation in proteins: a cross correlation approach. J Theor Biol **50,** 167–184.

Randic, M., Vracko, M., Lers, N., and Plavsic, D. (2003a). Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chem Phys Lett **368,** 1–6.

Randic, M., Vracko, M., Lers N., and Plavsic, D. (2003b). Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. Chem Phys Lett **371,** 202–207.

Webb, B.L.J., Hirst, S.J., and 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.

Yau, S.S.-T., Wang, J., Niknejad, A., Lu, C., Jin, N., and Ho, Y. (2003). DNA sequence representation without degeneracy. Nucleic Acids Res. **31,** 3078–3080.

Address reprint requests to:
Stephen S.-T. Yau, Ph.D.

Distinguished Professor
Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
851 S. Morgan St.
Chicago, IL 60607-7045

Director
Institute of Mathematics
East China Normal University
Shanghai 200062
China

E-mail: yau@uic.edu

Received for publication August 23, 2007; received in revised form December 7, 2007; accepted December 24, 2007.