Distances and classification of amino acids for different protein secondary structures

Xin Liu,[1] Li-mei Zhang,[2] Shan Guan,[3] and Wei-Mou Zheng[1]
[1]*Institute of Theoretical Physics, Beijing 100080, China*
[2]*School of Science at North Jiaotong University, Beijing 100044, China*
[3]*Center of Bioinformations at Peking University, Beijing 100871, China*

(Received 18 December 2002; published 27 May 2003)

Window profiles of amino acids in protein sequences are used to describe the amino acid environment. The relative entropy or Kullback-Leibler distance derived from these profiles is used as a measure of dissimilarity for comparison of amino acids and secondary structure conformations. Distance matrices of amino acid pairs at different conformations are obtained, which display a non-negligible dependence of amino acid similarity on conformations. Based on the conformation specific distances, a clustering analysis for amino acids is conducted.

DOI: 10.1103/PhysRevE.67.051927 PACS number(s): 87.10.+e, 02.50.−r

I. INTRODUCTION

The similarity of amino acids is the basis of protein sequence alignment, protein design, and protein structure prediction. Several scoring schemes have been proposed based on amino acid similarity. The mutation data matrices of Dayhoff [1] and the substitution matrices of Henikoff [2] are standard choices of scores for sequence alignment and amino acid similarity evaluation. However, these matrices, focusing on the whole protein database, pay little attention to protein secondary structures. How the amino acid similarity is influenced by different secondary structures is an interesting question. Furthermore, understanding the differences can help us in the protein sequence analysis.

Despite the efforts in uncovering the information encoded in the primary structure, we still cannot read the language describing the final three-dimensional fold of an active biological macromolecule. Compared with the DNA sequence, a protein sequence is generally much shorter, but the size of the alphabet is five times larger. A proper coarse graining of the 20 amino acids into fewer clusters for different conformations is important for improving the signal-to-noise ratio when extracting information by statistical means.

It is our purpose to propose a simple scheme to study amino acid similarity from amino acid string statistics. Information about the environment for an amino acid at a certain conformation state may be provided by the statistics of residue strings or windows centered at the amino acid. The success of window-based approaches such as GOR [3] for secondary structure prediction validates the use of such statistics. We shall derive a measure for the difference of amino acid pairs based on the distance of probability distributions, and investigate how the difference is dependent on conformations.

II. AMINO ACID DISTANCES

Our discussion will be strongly based on the distance between two probability distributions. A well-defined measure of the distance is the Kullback-Leibler (KL) distance or relative entropy [4–6], which, for two distributions $\{p_i\}$ and $\{q_i\}$, is given by

$$d(\{p_i\},\{q_i\})=\sum_i p_i \ln(p_i/q_i). \tag{1}$$

It corresponds to a likelihood ratio, and, if p_i is expanded around q_i, its leading term is the χ^2 distance:

$$d_\chi(\{p_i\},\{q_i\})=\sum_i (p_i-q_i)^2/p_i. \tag{2}$$

It is often used in the following symmetrized form for the KL distance:

TABLE I. Sample sizes of each amino acid residue in different protein secondary structures.

	h	*e*	*c*	*t*
C	690	732	822	224
S	2841	1764	3538	1179
T	2350	2288	3112	762
P	1173	624	3648	1302
A	5950	2019	2651	1122
G	1795	1633	4328	3090
N	1904	922	2692	1388
D	2841	1029	3621	1424
E	4773	1514	2325	1172
Q	2757	1008	1532	653
H	1132	794	1148	426
R	3108	1469	1948	771
K	3861	1579	2645	1187
M	1390	693	679	223
I	3169	3333	1719	368
L	6262	3307	2952	850
V	3233	4461	2330	487
F	2225	1948	1545	444
Y	1806	1773	1303	459
W	827	632	536	173

1063-651X/2003/67(5)/051927(7)/$20.00

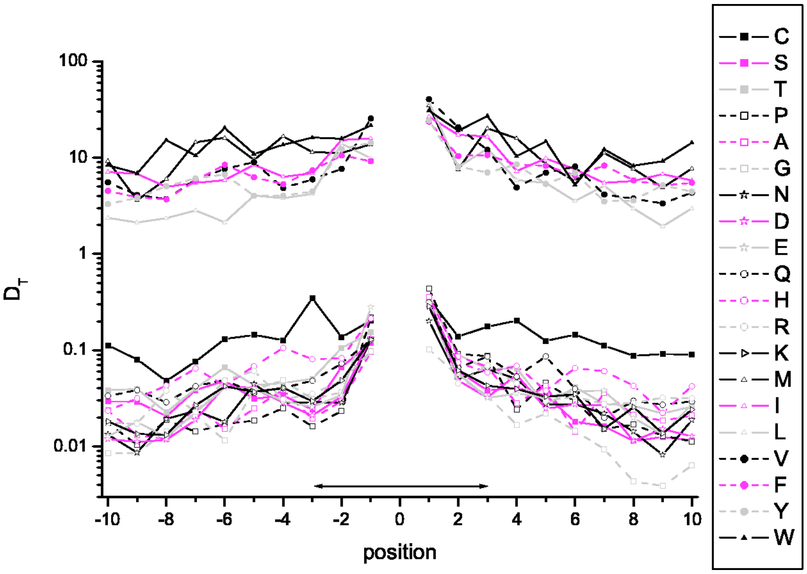

FIG. 1. (Color online) KL distances (doubled) of outer sites from their corresponding noise background. Each curve is for an amino acid at the center labeled as 0, whose conformation is turn. For clarity, the curves for *M*, *I*, *L*, *V*, *F*, *Y*, and *W* have been shifted up by multiplying an extra factor 100.

$$D(\{p_i\},\{q_i\}) = \tfrac{1}{2}[d(\{p_i\},\{q_i\}) + d(\{q_i\},\{p_i\})]. \quad (3)$$

The distributions to be considered here come from window statistics. For a given amino acid residue $a_i = x$ at the conformation state α in a sequence $a_1 a_2 \cdots a_i \cdots$, we take the string $a_{-n+i} a_{-n+i+1} \cdots a_i \cdots a_{i+n}$ of width $(2n+1)$ as a window. Denote by $N_k(y|x,\alpha)$ the count of residue y at the k th site from the center of such windows. As in GOR, only the conformation of the central residue is concerned. A quantity derived from $N_k(y|x,\alpha)$ is

$$N(x,\alpha) = \sum_y N_k(y|x,\alpha), \quad (4)$$

which, as the total count of residue x at the conformation α, is independent of k. The conditional probability distribution $P_k(y|x,\alpha)$ is estimated as

$$P_k(y|x,\alpha) = \frac{N_k(y|x,\alpha)}{N(x,\alpha)}. \quad (5)$$

The weight matrix $\mathbf{M}_{20\times 2n}$ with its entries being $P_k(y|x,\alpha)$ is the so-called residue profile of x at α. Such profiles are used in window-based approaches, e.g., GOR and artificial neural network algorithm [7].

We expect that, on an average, the correlation between the central residue and an outer site decays when these become far apart in the sequence. To examine the correlation, we consider a large window width of 21, i.e., $n=10$, and take the "noise" background to be the following average:

$$Q(y|x,\alpha) = \tfrac{1}{6}\left[\sum_{k=-10}^{-8} P_k(y|x,\alpha) + \sum_{k=8}^{10} P_k(y|x,\alpha)\right]. \quad (6)$$

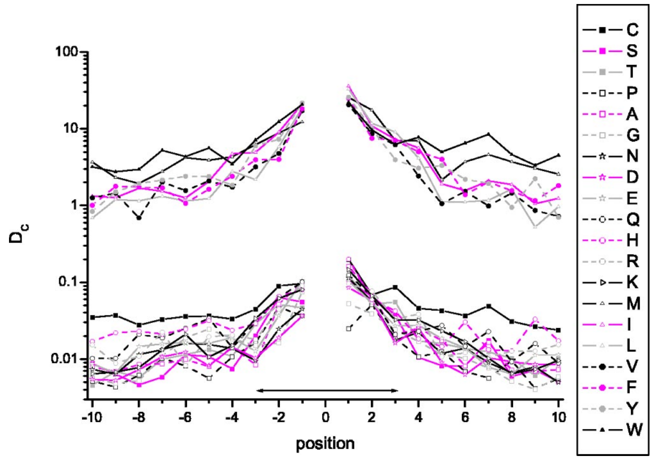

FIG. 2. (Color online) KL distances (doubled) of outer sites from their corresponding noise background. Each curve is for an amino acid at the center labeled as 0, whose conformation is coil. For clarity, the curves for *M*, *I*, *L*, *V*, *F*, *Y*, and *W* have been shifted up by multiplying an extra factor 100.

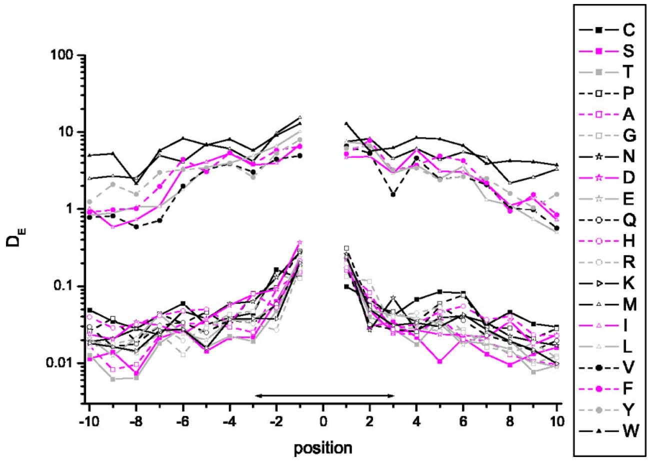

FIG. 3. (Color online) KL distances (doubled) of outer sites from their corresponding noise background. Each curve is for an amino acid at the center labeled as 0, whose conformation is sheet. For clarity, the curves for *M*, *I*, *L*, *V*, *F*, *Y*, and *W* have been shifted up by multiplying an extra factor 100.

The KL distance $D_{k;x,\alpha}(\{P_k(y|x,\alpha)\},\{Q(y|x,\alpha)\})$ provides a measure of the correlation between the central site and site k. As we shall see, for our purpose of amino acid comparison, a narrow window of a strong correlation with width of 7 is used to describe amino acid enviroment.

Using distribution $P_k(y|x,\alpha)$ from window statistics to characterize amino acid residues, we define the distance of residue pair x and y at the same conformation α as the following sum of KL distances:

$$D_{xy;\alpha}=\sum_{k=\pm 1,\pm 2,\pm 3} D(\{P_k(z|x,\alpha)\},\{P_k(z|y,\alpha)\}). \quad (7)$$

Similarly, to explore the difference of the same residue x at different conformations α and β, we may define the distance

$$D_{\alpha\beta;x}=\sum_{k=\pm 1,\pm 2,\pm 3} D(\{P_k(z|x,\alpha)\},\{P_k(z|x,\beta)\}). \quad (8)$$

By means of the residue pair distances, we can further study the classification of amino acids. With the KL distance, we may define the cluster distance in a way consistent with that for residue pairs. For example, we characterize the cluster consisting of residues x and y by the ‘‘coarse-grained’’ probability

$$P_k(z|x\&y,\alpha)=\frac{N_k(z|x,\alpha)+N_k(z|y,\alpha)}{N(x,\alpha)+N(y,\alpha)}. \quad (9)$$

Then we may define the distance between this cluster and some other residues or clusters. With cluster distance defined, the cluster analysis can be used to reduce amino acid alphabets.

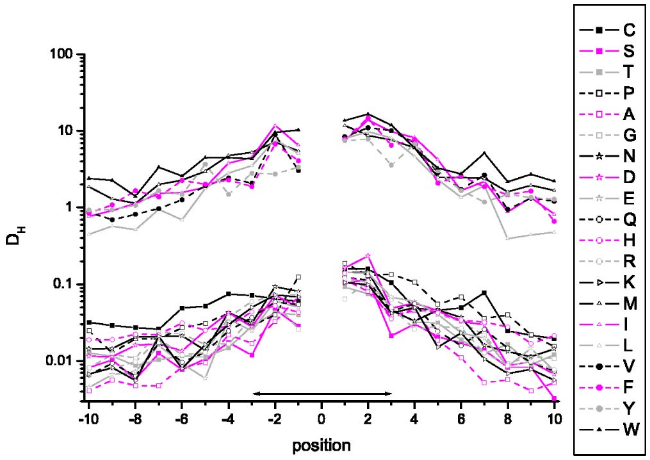

FIG. 4. (Color online) KL distances (doubled) of outer sites from their corresponding noise background. Each curve is for an amino acid at the center labeled as 0, whose conformation is helix. For clarity, the curves for *M*, *I*, *L*, *V*, *F*, *Y*, and *W* have been shifted up by multiplying an extra factor 100.

III. RESULTS

Our analysis is performed on a dataset taken from the database PDB_SELECT [8,9] of nonredundant protein sequences with known structures. The sequences share amino acid identity less than 25%. We keep only the nonmembrane sequences with their lengths between 80 and 420. The secondary structure assignment is taken from the Database of Secondary Structure in Proteins (DSSP) [10]. As in GOR, we use the following reduction of the eight DSSP states to four states of helix (h), sheet (e), coil (c), and turn (t): $H,G,I \rightarrow h$, $E \rightarrow e$, $X,S,B \rightarrow c$, and $T \rightarrow t$. The counts of each

TABLE II. Amino acid distance matrices for helices (bottom left) and turns (top right). Entries have been multiplied by a factor 200.

C		106	116	135	118	145	134	132	121	129	111	118	124	154	134	121	119	104	123	215
S	64		23	52	29	59	35	33	36	37	54	26	36	78	61	38	49	45	40	100
T	63	13		61	33	74	40	35	39	46	62	33	37	92	63	40	46	45	38	93
P	81	48	49		44	99	71	69	54	62	82	47	55	106	89	71	71	62	66	132
A	45	21	17	63		64	38	39	32	36	58	29	33	63	64	34	46	48	43	98
G	57	15	20	52	25		32	39	54	55	57	47	52	81	79	61	88	75	70	115
N	82	14	22	67	33	26		18	30	31	44	29	31	72	68	38	63	53	36	96
D	101	17	26	56	39	32	16		25	34	44	30	29	77	58	36	54	49	37	91
E	82	20	25	56	27	36	22	14		33	53	37	23	73	65	46	51	59	43	106
Q	70	16	21	60	19	28	17	21	14		51	32	38	79	66	51	62	62	49	100
H	55	23	24	55	26	26	33	35	34	28		51	58	90	79	54	81	70	55	113
R	69	21	22	59	21	30	22	28	24	13	28		30	71	69	38	54	49	48	101
K	80	21	25	67	28	38	22	27	23	19	38	13		81	63	38	49	51	47	102
M	48	57	45	85	23	56	75	82	64	51	50	50	60		93	62	85	93	78	141
I	43	81	65	104	35	78	104	116	88	76	66	73	79	22		47	55	55	54	104
L	35	65	52	90	26	62	83	99	73	59	53	56	67	15	10		49	36	31	85
V	37	59	44	81	22	55	77	90	67	53	52	51	60	16	12	09		46	58	99
F	34	67	53	87	30	61	90	99	79	69	54	66	75	22	17	12	15		48	100
Y	44	43	35	77	23	47	64	71	55	47	34	47	54	26	29	21	22	16		90
W	49	61	53	82	35	64	87	92	72	58	57	60	71	31	35	27	29	25	24	
	C	*S*	*T*	*P*	*A*	*G*	*N*	*D*	*E*	*Q*	*H*	*R*	*K*	*M*	*I*	*L*	*V*	*F*	*Y*	*W*

TABLE III. Amino acid distance matrices for sheets (bottom left) and coils (top right). Entries have been multiplied by a factor 200.

C		43	51	70	36	47	51	66	50	51	46	48	54	55	51	44	42	47	48	66
S	42		10	24	17	17	16	21	20	14	30	15	19	32	38	27	24	30	32	45
T	49	15		28	19	24	16	21	17	16	34	17	17	26	32	24	22	28	31	46
P	68	42	46		37	28	25	22	31	28	48	28	31	52	67	59	49	62	61	71
A	33	20	24	42		16	22	29	16	16	30	17	21	23	27	17	15	24	25	41
G	35	29	37	62	16		18	23	31	28	36	17	31	31	44	17	27	31	32	44
N	51	23	27	46	30	37		14	19	19	30	19	20	34	42	34	31	31	34	54
D	54	24	31	46	32	42	23		22	23	39	23	22	46	56	48	41	46	51	65
E	60	21	19	48	32	47	26	24		14	32	14	11	25	30	25	17	25	24	39
Q	52	20	17	53	28	41	29	30	22		32	14	14	30	33	26	21	26	25	38
H	50	27	26	54	28	33	34	33	30	28		29	36	51	44	41	39	44	40	67
R	46	21	20	44	20	33	32	31	21	23	22		16	31	33	24	21	28	29	47
K	62	29	20	52	30	47	35	34	20	23	35	24		34	34	28	21	29	26	50
M	38	45	44	65	24	33	52	62	50	46	44	38	52		28	22	22	27	30	39
I	32	38	36	62	24	35	56	57	49	41	40	36	43	23		12	15	16	20	34
L	27	37	34	58	19	29	50	55	45	41	37	32	43	20	09		10	12	17	33
V	31	35	32	58	19	27	51	57	46	40	36	32	38	22	09	10		14	17	29
F	29	45	44	71	25	33	62	67	59	47	49	42	56	28	14	12	15		18	33
Y	32	35	32	64	24	33	51	54	47	34	33	31	42	29	13	13	15	14		31
W	46	57	58	71	47	60	69	76	62	52	54	57	66	48	39	39	38	33	37	
	C	*S*	*T*	*P*	*A*	*G*	*N*	*D*	*E*	*Q*	*H*	*R*	*K*	*M*	*I*	*L*	*V*	*F*	*Y*	*W*

amino acid at the reduced four different conformation states are given in Table I.

We first estimate probability distributions of residues for each central residue at a given conformation. At this step, the window width is 21. We then calculate distances $D_{k;x,\alpha}(\{P_k(y|x,\alpha)\},\{Q(y|x,\alpha)\})$ of these distributions to their corresponding noise distributions. The results are shown in Figs. 1–4, each of which is for one conformation of the central residue. The 20 curves in each figure correspond to 20 central amino acids. Due to the sample size difference, curves are not directly comparable. (Roughly speaking, under the null hypothesis of identical distribution, the χ^2 distance should be scaled with the sample size, so a small sample size would give a relatively large distance.) However, a decay is clearly seen when the site k becomes far away from the center. For more discussions on correlations, we refer reader to Refs. [11,12]. As seen from most curves of the figures, distances at the six sites nearest to the center are significantly larger than those at window border sites. We shall use window width of 7 for further comparison of amino acids.

It is natural to expect that similar residues would have similar window statistics. Thus, the KL distance between two residue profiles provides a measure of their similarity, i.e., a small KL distance implies a large similarity. We calculate the KL distance matrices $D_{xy;\alpha}$ for residue pairs at different conformations with formula (7). The results are given in Tables II and III, where entries have been multiplied by a factor 200. With the distributions (9) defined for clusters, we fur-

TABLE IV. Clustering of amino acid alphabets for helices. The first column indicates the number of amino acid groups.

```
19 A D E K Q R S T N G H C F I LV M Y W P
18 A D E K Q R S T N G H C F ILV  M Y W P
17 A D E K Q R S T N G H C FILV   M Y W P
16 A D E K Q R ST  N G H C FILV   M Y W P
15 A D E K QR  ST  N G H C FILV   M Y W P
14 A D E KQR   ST  N G H C FILV   M Y W P
13 A D E KQRST     N G H C FILV   M Y W P
12 A D E KQRSTN      G H C FILV   M Y W P
11 A D EKQRSTN       G H C FILV   M Y W P
10 A DEKQRSTN        G H C FILV   M Y W P
 9 A DEKQRSTN        G H C FILVM    Y W P
 8 ADEKQRSTN         G H C FILVM    Y W P
 7 ADEKQRSTN         G H C FILVMY     W P
 6 ADEKQRSTNG          H C FILVMY     W P
 5 ADEKQRSTNGH           C FILVMY     W P
 4 ADEKQRSTNGH           C FILVMYW      P
 3 ADEKQRSTNGH           CFILVMYW       P
 2 ADEKQRSTNGHCFILVMYW                  P
```

TABLE V. Clustering of amino acid alphabets for sheets. The first column indicates the number of amino acid groups.

```
19 A G F IL V Y M D E Q S T R K H N C W P
18 A G F ILV  Y M D E Q S T R K H N C W P
17 A G FILV   Y M D E Q S T R K H N C W P
16 A G FILVY    M D E Q S T R K H N C W P
15 A G FILVY    M D E Q ST  R K H N C W P
14 A G FILVY    M D E QST   R K H N C W P
13 A G FILVY    M D EQST    R K H N C W P
12 A G FILVY    M D EQSTR     K H N C W P
11 A G FILVY    M D EQSTRK      H N C W P
10 AG  FILVY    M D EQSTRK      H N C W P
 9 AGFILVY      M D EQSTRK      H N C W P
 8 AGFILVYM       D EQSTRK      H N C W P
 7 AGFILVYM       D EQSTRKH       N C W P
 6 AGFILVYM       D EQSTRKHN        C W P
 5 AGFILVYM       DEQSTRKHN         C W P
 4 AGFILVYMDEQSTRKHN                C W P
 3 AGFILVYMDEQSTRKHNC                 W P
 2 AGFILVYMDEQSTRKHNCW                  P
```

TABLE VI. Clustering of amino acid alphabets for coils. The first column indicates the number of amino acid groups.

```
19 A E K Q R ST N G D F L V I Y M H P W C
18 A E K Q R ST N G D F LV  I Y M H P W C
17 A E K Q R ST N G D FLV   I Y M H P W C
16 A E K Q R ST N G D FLVI    Y M H P W C
15 A EK  Q R ST N G D FLVI    Y M H P W C
14 A EKQ   R ST N G D FLVI    Y M H P W C
13 A EKQR    ST N G D FLVI    Y M H P W C
12 A EKQRST     N G D FLVI    Y M H P W C
11 AEKQRST      N G D FLVI    Y M H P W C
10 AEKQRSTN       G D FLVI    Y M H P W C
 9 AEKQRSTNG        D FLVI    Y M H P W C
 8 AEKQRSTNG        D FLVIY     M H P W C
 7 AEKQRSTNGD         FLVIY     M H P W C
 6 AEKQRSTNGDFLVIY              M H P W C
 5 AEKQRSTNGDFLVIYM               H P W C
 4 AEKQRSTNGDFLVIYMH                P W C
 3 AEKQRSTNGDFLVIYMHP                 W C
 2 AEKQRSTNGDFLVIYMHPW                  C
```

TABLE VII. Clustering of amino acid alphabets for turns. The first column indicates the number of amino acid groups.

```
19 A DN E K S T R Q L Y F V H G I P M W C
18 A DN E K ST  R Q L Y F V H G I P M W C
17 A DN EK  ST  R Q L Y F V H G I P M W C
16 A DNEK   ST  R Q L Y F V H G I P M W C
15 A DNEKST     R Q L Y F V H G I P M W C
14 A DNEKSTR      Q L Y F V H G I P M W C
13 ADNEKSTR       Q L Y F V H G I P M W C
12 ADNEKSTRQ        L Y F V H G I P M W C
11 ADNEKSTRQL         Y F V H G I P M W C
10 ADNEKSTRQLY          F V H G I P M W C
 9 ADNEKSTRQLYF           V H G I P M W C
 8 ADNEKSTRQLYFV            H G I P M W C
 7 ADNEKSTRQLYFVH             G I P M W C
 6 ADNEKSTRQLYFVHG              I P M W C
 5 ADNEKSTRQLYFVHGI               P M W C
 4 ADNEKSTRQLYFVHGIP                M W C
 3 ADNEKSTRQLYFVHGIPM                 W C
 2 ADNEKSTRQLYFVHGIPMW                  C
```

ther perform the simplest bottom-up approach of hierachical clustering for residues, by starting from 20 clusters of single residues and then joining the two nearest clusters step by step until a single cluster is obtained. The results of clustering are given in Tables IV–VII. Since the dendritic trees returned from clustering are less informative, for visualization we introduce graphs where vertices are the 20 amino acids and an edge exists between a pair of amino acids if and only if their distance is below some preset threshold. Graphs obtained from the distance matrices are shown in Figs. 5–8, where vertices with no connecting edges are neglected.

In sequence pair alignment, we often do not have structure information of both the sequences. With the structure information ignored, we have the mixed counts

$$N_k(y|x) = \sum_{\alpha} N_k(y|x,\alpha), \tag{10}$$

from which we calculate the residue pair distances averaged over conformations. The distance matrix obtained is given in Table VIII. We have also calculated distances (8) to compare different conformations. Distances between any two conformations for various residues are listed in Table IX.

IV. DISCUSSIONS

Figures 1–4 illustrate the dependence of outer sites in a window on the center. Although in the KL distance, we sum up effects on individual residues from the center, we still can

TABLE VIII. Table 8. Amino acid distances ignoring conformation.

C																				
S	21																			
T	25	5																		
P	25	9	11																	
A	29	12	12	16																
G	21	8	11	11	11															
N	25	7	9	13	12	8														
D	32	9	9	15	10	11	6													
E	40	18	18	21	11	18	14	9												
Q	34	12	12	18	8	14	10	9	8											
H	21	13	14	17	18	14	12	15	23	17										
R	31	11	13	16	7	13	11	10	9	5	15									
K	35	15	14	18	12	16	10	9	8	10	22	8								
M	33	19	16	20	10	17	18	18	19	16	24	15	18							
I	25	16	13	16	12	14	16	17	20	18	19	16	15	10						
L	26	16	14	17	9	14	16	17	19	15	20	14	15	8	4					
V	24	10	9	13	8	9	11	12	15	13	17	12	12	10	6	6				
F	22	13	11	16	13	11	14	16	20	18	18	16	15	12	6	6	6			
Y	24	9	9	13	13	10	11	14	19	15	14	15	14	13	8	9	7	5		
W	32	20	19	20	21	17	22	25	29	23	24	24	27	18	14	13	13	10	12	
	C	S	T	P	A	*G*	N	D	*E*	Q	H	R	K	M	I	L	V	F	Y	W

TABLE IX. Conformation pair distances for each amino acid. Entries have been multiplied by a factor 200. (*h*: Helix, *e*: Sheet, *c*: coil, and *t*: Turn.)

	he	*hc*	*ht*	*ec*	*et*	*ct*
C	133	185	163	127	197	139
S	93	129	124	93	148	73
T	98	120	131	103	175	96
P	172	118	121	89	233	116
A	112	148	127	122	149	73
G	79	101	80	91	107	57
N	126	145	118	106	152	76
D	149	137	149	93	174	81
E	159	152	138	109	192	73
Q	130	157	133	93	143	93
H	100	150	110	117	152	98
R	131	146	128	91	144	85
K	137	149	128	93	155	88
M	130	161	147	126	156	135
I	138	180	134	118	130	110
L	143	162	113	127	148	98
V	114	151	151	98	147	101
F	120	150	111	107	115	88
Y	95	147	96	111	117	80
W	120	181	201	123	173	111

see the tendency that the center is generally more strongly correlated with the *C*-terminal sites than *N*-terminal sites. Furthermore, we may divide the 20 amino acids into two groups with *M*, *I*, *L*, *V*, *F*, *Y*, and *W* in one, and the remaining in the other. These roughly correspond to hydrophobic and hydrophilic groups. It is seen that for the coil and turn conformations, a hydrophobic center exhibits a stronger correlation with outer sites than a hydrophilic center, while for the sheet conformation a hydrophilic center exhibits a stronger correlation.

It is interesting to make a comparison between the distance matrices obtained here with the commonly used Block substitution matrix (BLOSUM62) similarity score matrix. A small distance implies a large similarity score. There are many evidences showing the consistency between the distances and the scores. For example, residue pairs *VI*, *IL*, *VL*, and *ST* have positive BLOSUM scores and at the same time small distances. The graphs in Figs. 5–8 contain two

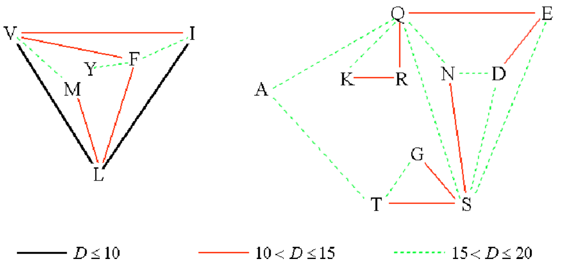

FIG. 5. (Color online) Connecting graph of amino acids in helix. Edges exist only between vertices with a scaled distance not greater than 20. Vertices without any connecting edges are not shown.

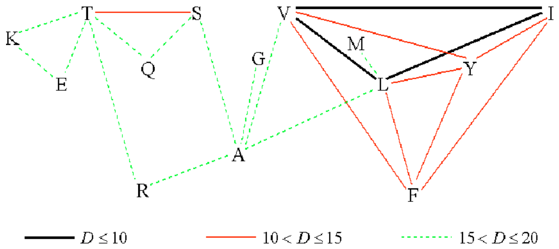

FIG. 6. (Color online) Connecting graph of amino acids in sheet. Edges exist only between vertices with a scaled distance not greater than 20. Vertices without any connecting edges are not shown.

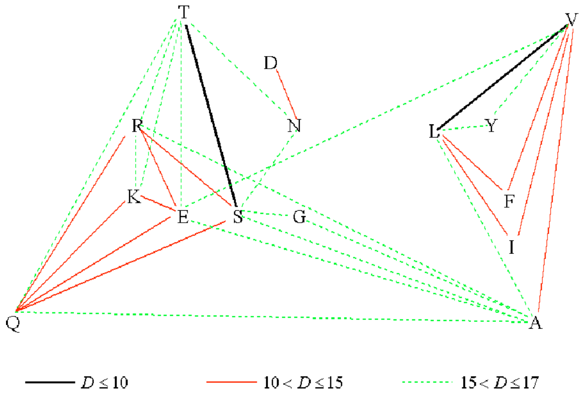

FIG. 7. (Color online) Connecting graph of amino acids in coil. Edges exist only between vertices with a scaled distance not greater than 17. Vertices without any connecting edges are not shown.

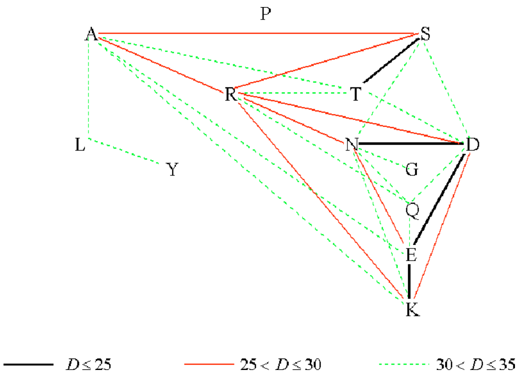

FIG. 8. (Color online) Connecting graph of amino acids in turn. Edges exist only between vertices with a scaled distance not greater than 35. Vertices without any connecting edges are not shown.

connected subgraphs: one consists of *I*, *L*, *V*, *F*, *Y*, and the other consists of *S*, *T*. This is another evidence of the consistency. Generally, the averaged distance matrix is closer to BLOSUM62 than the conformation specific ones. However, there do exist some remarkable differences. For example, residue pairs GT, QA, FV with negative scores have rather small distances in either the conformation helix, or sheet or coil, while pairs YH and NH with positive scores have rather large distances in the helix conformation. Moreover, *YH* has a large distance in all the four conformations.

BLOSUM matrices are derived from conserved amino acid patterns called blocks. It is expected that for most score entries, we should see the consistency in at least one conformation specific distance matrix. For a given residue pair, if residue profiles of an amino acid center are very dissimilar for different conformations, after averaging over conformations the pair distance would generally become smaller. In this case, BLOSUM scores and conformation specific distance need not be consistent since the former contains no structure information.

Our results show some strong dependence of residue behavior on conformations. For example, the distances of pairs CD and SI in helix are about two times higher than in sheet. There are many residue pairs displaying strong dependence of distances on conformations. Table IX views the conformation dependence from conformation pair comparison. Indeed, the table indicates that for any conformation pairs, there are certain residues, which behave very differently in the two conformations. However, generally speaking, coil and turn are quite similar.

In a comparison of physicochemical properties of amino acids, the abundance of amino acids is not taken into consideration. This is also the case for the above defined distances. Other statistical variables including the effect of sample size may be introduced. One candidate is the χ^2 statistic for identical distributions. The analysis using this new statistic is under study.

We expect that algorithms using multiple conformation specific matrices should work better in sequence alignment. The popular Needleman-Wunsch algorithm can be modified to include putative conformation for each residue. This will be discussed elsewhere.

This work was supported in part by the Special Funds for Major National Basic Research Projects and the National Natural Science Foundation of China.

[1] M.O. Dayhoff and R.V. Eck, *Atlas of Protein Sequence and Structure* (Natl. Biolmed. Res. Found., Silver Springs, MD, 1968), Vol. 3, p. 33.

[2] S. Henikoff and J.G. Henikoff, Proc. Natl. Acad. Sci. U.S.A. **89**, 10 915 (1992).

[3] J. Garnier, D. Osguthorpe, and B. Robson, J. Mol. Biol. **120**, 97 (1978).

[4] S. Kullback, J.C. Keegel, and J.H. Kullback, *Information Theory and Statistics* (Wiley, New York, 1959).

[5] S. Kullback, *Topics in Statistical Information Theory* (Springer, Berlin, 1987).

[6] T. Sakamoto, M. Ishiguro, and G. Kitagawa, *Akaike Information Criterion Statistics* (KTK Scientific, Tokyo, 1986).

[7] P. Stolorz, A. Lapedes, and Y. Xia, J. Mol. Biol. **225**, 363 (1992).

[8] U. Hobohm and C. Sander, Protein Sci. **3**, 522 (1994).

[9] See URL:ftp.embl-heidelberg.de/pub/databases/protein_extras/pdb_select/old/2001_sep.25

[10] W. Kabsch and C. Sander, Biopolymer **22**, 2577 (1983).

[11] O. Weiss and H. Herzel, J. Theor. Biol **190**, 341 (1998).

[12] O. Weiss, M.A. Jiménez-Montaǹo, and H. Herzel, J. Theor. Biol **206**, 379 (2000).