Journal of Protein Chemistry, Vol. 4, No. 1, 1985

Statistical Analysis of the Physical Properties of the 20 Naturally Occurring Amino Acids

Akinori Kidera,[1] Yasuo Konishi,[1] Masahito Oka,[1] Tatsuo Ooi,[2] and Harold A. Scheraga[1,3]

Received January 25, 1985

In order to describe the conformational and other physical properties of the 20 naturally occurring amino acid residues with a minimum number of parameters, several multivariate statistical analyses were applied to 188 of their physical properties and ten orthogonal properties (factors) were obtained for the 20 amino acids without losing the information contained in the original physical properties. The analysis consisted of three main steps. First, 72 of the physical properties were eliminated from further consideration because they did not pass statistical tests that they follow a normal distribution. Second, the remaining 116 physical properties of the amino acids were classified by a cluster analysis to eliminate duplications of highly correlated physical properties. This led to nine clusters, each of which was characterized by an average characteristic property, namely bulk, two hydrophobicity indices for free amino acids, one hydrophobicity index for amino acid residues in a protein, two types of β-structure preference, α-helix preference, and two types of bend-structure preference. The physical properties within a given cluster were highly correlated with each other, but the correlation between clusters was low. Third, a factor analysis was applied to the nine average classified properties and 16 additional physical properties to obtain a small number of orthogonal properties (ten factors). Four of these factors arise from the nine characteristic properties, and the remaining six factors were obtained from the 16 physical properties not included in the nine characteristic properties. Finally, most of the 188 physical properties could be expressed as a sum of these ten orthogonal factors, with appropriate weighting factors. Since these factors contain information relating almost all properties of all 20 amino acids, it is possible to estimate the numerical values of a property for one or two amino acids for which experimental data for this property are not available. For example, the estimated values for the Zimm-Bragg parameter s *at 20°C are 0.66 and 0.92 for proline and cysteine, respectively, computed from the first four factors.*

KEY WORDS: amino acid physical properties; characteristic properties; bulk; hydrophobicity; β-structure preference; α-helix preference; bend-structure preference; statistical analysis; cluster analysis; factor analysis.

1. INTRODUCTION

One of the ultimate goals in studying protein structure is to understand how information about the three-dimensional structure is encoded in the one-dimensional sequence of its amino acids. A quantitative elucidation of such information must begin with a parametrization of the amino acid sequence to incorporate: (1) informa-

[1] Baker Laboratory of Chemistry, Cornell University, Ithaca, New York 14853-1301.
[2] Institute for Chemical Research, Kyoto University, Uji 611, Japan.
[3] To whom requests for reprints should be addressed at Cornell University.

0277-8033/85/0200-0023$04.50/0

tion about intraresidue interactions and the interactions between a residue and the surrounding solvent molecules, and (2) information about interresidue interactions and cooperative interactions between two or more residues and the solvent molecules. The former can easily be parametrized in terms of the physical properties of the amino acids. However, the latter has too many degrees of freedom to be parametrized from a data base of finite size (e.g., we must consider 400 possible pairs of 20 amino acids for only the nearest-neighbor interaction). To avoid this difficulty in this study, and thereby to include information about long-range interactions implicitly, we introduce the following assumption: all of the information contained in the amino acid sequence (both of types 1 and 2 above) can be represented in the form of a *pattern* appearing in the amino acid sequence, which, in turn, is parametrized in terms of the physical properties of its constituent amino acids, i.e., the sequence profiles of the physical properties of its amino acids. Under this assumption, the problem is then simply to identify those physical properties that are the most important ones for determining protein structure. This study comprises two steps: (1) the assignment of as much information as possible to the 20 naturally occurring amino acids; and (2) the definition of the most important physical properties from the sequence profiles of those physical properties assigned in the first step.

We can use many different kinds of physical properties of the amino acids as the parameters; for example, the propensity to form ordered backbone structures has been used to predict so-called secondary structures [reviewed by Némethy and Scheraga (1977)], hydrophobicity indices have been used to predict the surface-inside profiles of amino acid sequences (Rose and Roy, 1980; Kyte and Doolittle, 1982), and so on. In this paper and in one presented elsewhere (Kidera *et al.*, 1985), we are faced with the problem of identifying those properties (from a set of a large number of physical properties) that are the most important ones for determining protein structure. Therefore, it is necessary that the parameters describing the amino acids include as much information as possible. However, when we try to use all of the available properties as parameters, we encounter the following two problems: (1) there are many duplications of highly correlated physical properties, and (2) the number of physical properties is unmanageably large.

In this paper, by applying several multivariate statistical analyses to a number of physical properties of the 20 naturally occurring amino acids, we parametrize the latter to obtain a set of a small number of properties that has a sufficient amount of information but no duplication (the first step of this study). We use a factor analysis (Lawley and Maxwell, 1971) to obtain ten orthogonal properties (factors) for the 20 amino acids. This is accomplished by first classifying the amino acid properties by a cluster analysis to reduce duplications (Hartigan, 1975) and then by carrying out a factor analysis of the reduced number of properties by a method that is developed in this paper.

Elsewhere (Kidera *et al.*, 1985), we apply the results obtained here (i.e., the ten factors), and use evolutionary relationships to define the optimal linear combination of these factors that is most important for determining the three-dimensional structures of proteins (the second step of this study). Then we demonstrate the possibility that the sequence profile of such an optimal combination of factors corresponds uniquely to the three-dimensional structure of a protein; i.e., we can

use this sequence profile to predict an initial protein structure (for subsequent energy minimization).

2. METHODS

In the factor analysis method, we describe the physical properties of each amino acid in a multidimensional space by using several multivariate statistical analyses (Morrison, 1976). The model in these analyses must have two characteristics: (1) it should contain a sufficient amount of information to be physically meaningful, and (2) it should contain a minimum number of parameters to be manageable. To satisfy these requirements, we collected a large number of physical properties of the amino acids and, by applying a factor analysis model (Lawley and Maxwell, 1971), we converted them into a small number NF of orthogonal (independent) properties (factors) without losing the information contained in the large set of properties. Then an amino acid sequence is describable in terms of the factors for its constituent amino acids, each of which may be represented in this NF-dimensional space. Ideally, as in a normal coordinate analysis of the vibrational motions of molecules, each factor should correspond as closely as possible to a given physical property. This is possible for *some* properties, but for others, the factors are linear combinations of several physical properties.

2.1. Physical Properties of the Amino Acids

We collected 188 physical properties of the 20 naturally occurring amino acids from the literature. These include not only the properties of the amino acid residues in proteins, but also those of the free amino acids. Their designations are listed in Table I. The criterion for including a property in this collection is that the values for at most two amino acids could be missing. The following comments about some of these properties should be noted:

1. Oobatake and Ooi (1977) reported the short- and medium-range energy per atom (defined as interactions within ten residues), and multiplied this by the number of heavy atoms to obtain the energy per amino acid residue (no. 4).

2. The Zimm-Bragg parameters (Sueki *et al.*, 1984) (nos. 100 and 138) pertain to *p*H 7, i.e., with Arg, Asp, Glu, and Lys in the charged state and His in the uncharged state.

3. The normalized frequency of occurrence of the α-helix (Crawford *et al.*, 1973) (no. 75) was calculated by summing their data for the N-helix, M-helix, and C-helix.

4. The normalized frequencies of Levitt (1978) (nos. 59, 78, and 96) are his data set II, and those "with weights" (Levitt, 1978) (nos. 63, 79, and 97) are his data set I.

5. The normalized frequencies of Tanaka and Scheraga (1977) (nos. 70, 82, and 98) were recalculated in the form

$$N_{ij}N_t / N_{it}N_{tj} \tag{1}$$

where N_{ij} is the number of occurrences of residue j in conformational state i, N_t

Table I. Physical Properties of Amino Acids and Results of Their Statistical Analysis[a]

Name of property	COR	ST	US	VAR4	VAR10	LCOR	P
1 Molecular weight (Fasman, 1976)	A		$	0.97	0.99	0.97	1
2 Residue accessible surface area in tripeptide (Chothia, 1976)	A		$	0.99	0.99	0.99	1
3 Molecular volume (Grantham, 1974)	A		$	0.98	0.99	0.97	1
4 Short- and medium-range energy per residue (Oobatake&Ooi, 1977)	A		$	0.93	0.96	-0.94	1
5 +Residue volume (Krigbaum&Komoriya, 1979) (Gly, Pro:8.0)	A		$	0.98	0.99	0.97	1
6 Distance between C^{α} and centroid of side-chain (Levitt, 1976)	A		$	0.98	0.99	0.96	1
7 +Radius of gyration of side-chain (Levitt, 1976) (Gly:0.09)	A		$	0.98	0.99	0.96	1
8 +Volume (Bigelow, 1967) (Asn, Gln:5.0)	A		$	0.98	0.99	0.98	1
9 +Residue volume (Goldsack&Chalifoux, 1973) (Asn, Gln:9.0)	A		$	0.98	0.99	0.98	1
10 Hydrophobic contribution to transfer free energy (von Heijne&Blomberg, 1979)	A		$	0.99	0.99	-0.99	1
11 +Polarizability parameter (Charton&Charton, 1982) (Pro:0.02)	A		$	0.98	0.99	0.98	1
12 Absolute entropy (Hutchens, 1970)	B			0.72	0.89	0.82	1
13 Refractivity (Jones, 1975)	B			0.76	0.92	0.83	1
14 Size (Dawson, 1972)	A	STP(A, 1)		0.85	0.96	0.90	1
15 Hydrophobicity (Jones, 1975)	A		$	0.77	0.83	0.99	2
16 Retention coefficient in $NaClO_4$ (Meek&Rossetti, 1981)	A		$	0.91	0.98	0.94	2
17 Transfer free energy (Simon, 1976)	A		$	0.72	0.86	0.97	2
18 +Hydrophobicity factor (Goldsack&Chalifoux, 1973) (Asn, Gln:!)	A			0.72	0.91	0.93	2
19 Free energy of transfer to surface (Bull&Breese, 1974)	A	STP(B, 2)		0.79	0.91	-0.91	2
20 Van der Waals parameter R_0 (Levitt, 1976)	B	STP(B, 2)		0.91	0.94	0.88	2
21 +Free energy of hydration of side-chain (Robson&Osguthorpe, 1979) (Gly:!)	B	STP(B, 2)		0.91	0.95	0.85	2
22 Polarity (Grantham, 1974)	A		$	0.94	0.98	-0.94	3
23 Hydrophobic parameter (Levitt, 1976)	A		$	0.89	0.94	-0.98	3
24 Hydrophilicity value (Hopp&Woods, 1981)	A		$	0.86	0.91	-0.98	3
25 R_F-rank (Zimmerman et al., 1968)	B			0.71	0.90	0.85	3
26 Polar requirement (Woese, 1973)	A	STP(A, 3)		0.87	0.95	-0.93	3
27 +Partition coefficient (Pliska et al., 1981) (Arg:!)	B	STP(A, 3)		0.93	0.94	0.88	2
28 Retention coefficient in NaH_2PO_4 (Meek&Rossetti, 1981)	A	STP(B, 3)		0.89	0.97	0.91	2
29 Partition coefficient (Garel et al., 1973)	E	STP(B, 3)		0.69	0.90	0.71	2
30 Proportion of residues 95 percent buried (Chothia, 1976)	A		$	0.91	0.94	0.93	4
31 Percentage of buried residues (Janin et al., 1978)	A		$	0.94	0.96	0.96	4
32 Percentage of exposed residues (Janin et al., 1978)	A		$	0.90	0.96	-0.91	4
33 Free energy of transfer (Janin, 1979)	A		$	0.92	0.98	0.93	4
34 Side-chain interaction parameter (Krigbaum&Komoriya, 1979)	A		$	0.83	0.93	-0.92	4

35	Average surrounding hydrophobicity (Manavalan&Ponnuswamy,1978)	A			0.92	0.94	0.93	4
36	Average number of surrounding residues (Ponnuswamy et al.,1980)	A		$	0.90	0.95	0.93	4
37	Average gain in surrounding hydrophobicity (Ponnuswamy et al.,1980)	A		$	0.88	0.96	0.96	4
38	Average gain ratio in surrounding hydrophobicity (Ponnuswamy et al.,1980)	A		$	0.88	0.95	0.96	4
39	+Average reduced distance of side-chain (Rackovsky&Scheraga,1977) (Gly:0.08)	A		$	0.85	0.93	-0.94	4
40	Average reduced distance (Meirovitch et al.,1980)	A		$	0.90	0.95	-0.91	4
41	+Average reduced distance of side-chain (Meirovitch et al.,1980) (Gly:0.07)	A		$	0.88	0.94	-0.96	4
42	+Average orientation angle of side-chain (Meirovitch et al.,1980) (Gly:7.4)	A		$	0.85	0.92	0.94	4
43	Single-residue parameter for relative contact number (Nishikawa&Ooi,1980)	A		$	0.87	0.96	0.93	4
44	Total range non-bonded energy per atom (Oobatake&Ooi,1977)	A		$	0.95	0.99	-0.94	4
45	Hydropathy index (Kyte&Doolittle,1982)	A		$	0.94	0.97	0.93	4
46	Average reduced distance (Rackovsky&Scheraga,1977)	B			0.84	0.94	-0.85	4
47	+Average interactions per side-chain atom (Warme&Morgan,1978) (Gly:0.8)	B			0.86	0.89	0.87	4
48	Long range non-bonded energy per atom (Oobatake&Ooi,1977)	B			0.92	0.98	-0.88	4
49	Fraction of site occupied by water (Krigbaum&Komoriya,1979)	D			0.68	0.87	-0.88	4
50	Surrounding hydrophobicity in bend structure (Ponnuswamy et al.,1980)	D	STP(A,4)		0.63	0.81	0.81	4
51	Average surrounding hydrophobicity (Ponnuswamy et al.,1980)	A	STP(A,4)		0.88	0.96	0.94	4
52	Average accessibility surface area (Janin et al.,1978)	B	STP(A,4)		0.94	0.98	-0.88	4
53	Residue accessible surface area in folded protein (Chothia,1976)	B	STP(A,4)		0.89	0.96	-0.87	4
54	+Side-chain orientational angle (Rackovsky&Scheraga,1977) (Gly:1)	B	STP(A,4)		0.70	0.85	-0.88	4
55	Proportion buried inside protein (Wertz&Scheraga,1978)	B	STP(B,4)		0.86	0.95	0.89	4
56	+Hydration number (Hopfinger,1977) (Cys:1)	C	STP(B,4)		0.71	0.82	-0.77	3
57	Normalized frequency of β-structure (Nagano,1973)	A		$	0.71	0.78	0.90	5
58	Conformation parameter for β sheet (Chou&Fasman,1978)	A		$	0.91	0.94	0.98	5
59	Normalized frequency for β sheet (Levitt,1978)	A		$	0.85	0.92	0.94	5
60	Information measure for extended state (Robson&Suzuki,1976)	A		$	0.86	0.91	0.94	5
61	Information measure in pleated-sheet (Robson&Suzuki,1976)	A		$	0.89	0.90	0.97	5
62	Conformational preference of β-structure (Lifson&Sander,1979)	A		$	0.93	0.97	0.96	5
63	Normalized frequency of β-structure with weights (Levitt,1978)	B			0.72	0.90	0.83	5
64	Conformational preference of anti-p β-structure (Lifson&Sander,1979)	D			0.85	0.89	0.89	5
65	Normalized frequency of β-structure (Crawford et al.,1973)	D			0.60	0.72	0.83	5
66	Normalized frequency in N-terminal β sheet (Chou&Fasman,1978)	D			0.69	0.92	0.87	5
67	Average relative probability of β sheet (Kanehisa&Tsong,1980)	A	STP(A,5)		0.89	0.92	0.97	5
68	β-coil equilibrium constant (Ptitsyn&Finkelstein,1983)	A	STP(A,5)		0.87	0.92	0.94	5
69	Average relative probability of inner β sheet (Kanehisa&Tsong,1980)	B	STP(B,5)		0.89	0.94	0.89	5
70	Normalized frequency of extended structure (Tanaka&Scheraga,1977)	A		$	0.79	0.95	0.97	6
71	Normalized frequency of extended structure (Isogai et al.,1980)	A		$	0.77	0.95	0.98	6
72	Normalized frequency of extended structure (Maxfield&Scheraga,1976)	A		$	0.82	0.93	0.97	6

Table I. (continued)

Name of property	COR	ST US	VAR4	VAR10	LCOR	P
73 Average relative fractional occurrence in $E_O(i)$ (Rackovsky&Scheraga,1982)	B	STP(B,6)	0.76	0.80	0.84	6
74 Conformational parameter of α-helix (Burgess et al.,1974)	A	$	0.86	0.93	0.92	7
75 Normalized frequency of α-helix (Crawford et al.,1973)	A	$	0.85	0.90	0.92	7
76 Normalized frequency of α-helix (Nagano,1973)	A	$	0.93	0.94	0.95	7
77 Conformation parameter for α-helix (Chou&Fasman,1978)	A	$	0.94	0.97	0.97	7
78 Normalized frequency for α-helix (Levitt,1978)	A	$	0.93	0.96	0.96	7
79 Normalized frequency of α-helix with weights (Levitt,1978)	A	$	0.90	0.95	0.95	7
80 Information measure for α-helix (Robson&Suzuki,1976)	A	$	0.97	0.98	0.98	7
81 Average relative probability of α-helix (Kanehisa&Tsong,1980)	A	$	0.95	0.97	0.97	7
82 Normalized frequency of α-helix (Tanaka&Scheraga,1977)	A	$	0.93	0.97	0.96	7
83 Normalized frequency of α-helix (Isogai et al.,1980)	A	$	0.94	0.96	0.96	7
84 Normalized frequency of α-helix (Maxfield&Scheraga,1976)	A	$	0.94	0.97	0.97	7
85 Average relative fractional occurrence in $A_R(i-1)$ (Rackovsky&Scheraga,1982)	A	STP(A,7)	0.85	0.90	0.91	7
86 Average relative probability of inner helix (Kanehisa&Tsong,1980)	A	STP(A,7)	0.89	0.92	0.90	7
87 Normalized frequency of coil (Nagano,1973)	B	STP(B,7)	0.88	0.92	-0.86	7
88 Average relative fractional occurrence in $E_L(i-1)$ (Rackovsky&Scheraga,1982)	E	STP(B,7)	0.29	0.90	-0.52	7
89 Normalized frequency of turn (Crawford et al.,1973)	A	$	0.87	0.90	0.95	8
90 Information measure for turns (Robson&Suzuki,1976)	A	$	0.76	0.88	0.99	8
91 Information measure in middle turn (Robson&Suzuki,1976)	A	$	0.71	0.88	0.97	8
92 Information measure in loop (Robson&Suzuki,1976)	A	$	0.82	0.90	0.99	8
93 Conformational parameter of β sheet (Beghin&Dirkx,1975)	A	STP(A,8)	0.80	0.93	0.93	8
94 Conformation parameter for β-turn (Chou&Fasman,1978)	A	$	0.93	0.95	0.96	9
95 Normalized frequency in the 2nd and 3rd residue of turn (Chou&Fasman,1978)	A	$	0.96	0.98	0.98	9
96 Normalized frequency for reverse-turn (Levitt,1978)	A	$	0.97	0.97	0.98	9
97 Normalized frequency of reverse turn with weights (Levitt,1978)	A	$	0.95	0.97	0.97	9
98 Normalized frequency of chain reversal (Tanaka&Scheraga,1977)	A	$	0.92	0.95	0.96	9
99 Normalized frequency of bend (Isogai et al.,1980)	A	$	0.93	0.98	0.96	9
100 +Zimm-Bragg parameter s at 20 C (Sueki et al.,1984) (Cys,Pro:0.10)	B		0.88	0.92	-0.89	9
101 Normalized frequency in N-terminal non-β region (Chou&Fasman,1978)	B		0.84	0.89	0.89	9
102 Information measure in coil (Robson&Suzuki,1976)	D		0.69	0.80	0.81	9
103 Information measure in middle helix (Robson&Suzuki,1976)	B	STP(B,9)	0.77	0.90	-0.82	9
104 Accessibility reduction ratio (Ponnuswamy et al.,1980)	B	STP(B,9)	0.79	0.88	0.82	4
105 Partial specific volume (Cohn&Edsall,1943)	E	%	0.46	0.84	0.65	2
106 +Apparent partial specific volume (Bull&Breese,1974) (Tyr:1)	E	%	0.57	0.90	0.77	2
107 Surrounding hydrophobicity in α-helical form (Ponnuswamy et al.,1980)	E	%	0.57	0.93	0.70	4
108 Surrounding hydrophobicity in β-structure (Ponnuswamy et al.,1980)	E	%	0.53	0.95	0.72	4
109 Relative population of conformational state C (Vasquez et al.,1983)	E	%	0.30	0.92	-0.51	1

110	Normalized frequency in C-terminal helix (Chou&Fasman,1978)	E	%	0.68	0.90	0.70 7
111	Normalized frequency in C-terminal non-β region (Chou&Fasman,1978)	E	%	0.65	0.93	0.78 9
112	Information measure in N-terminal turn (Robson&Suzuki,1976)	E	%	0.67	0.98	0.71 9
113	Normalized frequency of double bend (Isogai et al.,1980)	E	%	0.23	1.00	-0.49 3
114	Normalized frequency of α-region (Maxfield&Scheraga,1976)	E	%	0.32	0.98	0.54 9
115	Average relative fractional occurrence in $E_O(i)$ (Rackovsky&Scheraga,1982)	E	%	0.63	1.00	0.63 1
116	pK-C (Fasman,1976)	E	%	0.34	0.98	0.52 3
117	Heat capacity (Hutchens,1970)	E	%	0.54	0.87	0.63 1
118	Relative mutablity (Dayhoff et al.,1978a)	E	%	0.43	0.83	-0.65 2
119	Average percent in protein (Dayhoff et al.,1978b)	E	%	0.52	1.00	-0.67 1
120	Amino acid distribution (Jukes et al.,1975)	E	%	0.47	1.00	-0.62 1
121	Bulkiness (Zimmerman et al.,1968)	C		0.71	0.89	0.79 2
122	+Steric parameter (Charton,1981) (Pro:1)	E	STP(E,9)	0.56	0.76	-0.67 9
123	Polarity (Zimmerman et al.,1968)	E	STP(C,4)	0.66	0.90	-0.78 3
124	Hydrophobicity (Zimmerman et al.,1968)	E		0.38	0.74	0.74 2
125	+Hydration potential (Wolfenden et al.,1981) (Pro:2.9)	C		0.88	0.98	0.76 4
126	Retention coefficient in TFA (Browne et al.,1982)	E	STP(E,3)	0.63	0.87	0.76 2
127	Retention coefficient in HFBA (Browne et al.,1982)	E	STP(C,3)	0.50	0.74	0.72 2
128	R value (Weber&Lacey,1978)	C	STP(C,2)	0.88	0.97	-0.79 2
129	Proportion of residues 100 percent buried (Chothia,1976)	B	STP(C,4)	0.90	0.98	0.81 4
130	Intercept in regression analysis (Prabhakaran&Ponnuswamy,1982)	F		0.35	0.49	0.53 9
131	Slope in regression analysis (Prabhakaran&Ponnuswamy,1982)	E	RES	0.56	0.74	-0.71 1
132	Correlation coefficient in regression analysis (Prabhakaran&Ponnuswamy,1982)	F	STP(F,1)	0.11	0.60	-0.28 5
133	Free energy change of ϵ_i to ϵ_{ex} (Wertz&Scheraga,1978)	F	STP(C,2)	0.19	0.43	-0.42 2
134	+Free energy change of α_{Ri} to α_{Rh} (Wertz&Scheraga,1978)(Met:1)	F	STP(F,1)	0.13	0.56	0.29 3
135	Free energy change of ϵ_i to α_{Rh} (Wertz&Scheraga,1978)	E	RES	0.57	0.73	0.61 4
136	Relative population of conformational state A (Vasquez et al.,1983)	F		0.29	0.40	0.43 7
137	+Relative population of conformational state E (Vasquez et al.,1983) (Pro:1)	E	STP(E,2)	0.55	0.87	0.65 2
138	+Zimm-Bragg parameter σ (Sueki et al.,1984) (Cys,Pro:1)	F	STP(C,2)	0.46	0.55	0.60 2
139	+Conformation parameter of α-helix (Finkelstein&Ptitsyn,1976) (Pro:0.10)	C		0.86	0.91	0.77 7
140	Conformational parameter of bend (Lewis et al.,1971)	E	STP(C,9)	0.54	0.70	0.72 9
141	Conformational parameter of extended structure (Burgess et al.,1974)	E	RES	0.59	0.76	0.72 6
142	Helix-coil equilibrium constant (Ptitsyn&Finkelstein,1983)	B	STP(C,9)	0.73	0.84	-0.83 9
143	Normalized frequency in N-terminal helix (Chou&Fasman,1978)	E	STP(F,5)	0.50	0.70	-0.57 5
144	Normalized frequency in N-terminal non-helical region (Chou&Fasman,1978)	E	RES	0.42	0.70	0.60 9
145	Normalized frequency in C-terminal non-helical region (Chou&Fasman,1978)	E	RES	0.42	0.79	0.62 9
146	Normalized frequency in C-terminal β sheet (Chou&Fasman,1978)	E	STP(E,5)	0.61	0.84	0.75 5
147	Frequency in the 1st residue in bend (Chou&Fasman,1978)	E	STP(E,8)	0.41	0.80	0.66 8

Table I. (*continued*)

Name of property	COR	ST US	VAR4	VAR10	LCOR	P
148 Frequency in the 2nd residue in bend (Chou&Fasman, 1978)	E	STP(C,9)	0.47	0.82	0.68	9
149 Frequency in the 3rd residue in bend (Chou&Fasman, 1978)	E	STP(E,8)	0.52	0.87	0.73	8
150 Frequency in the 4th residue in bend (Chou&Fasman, 1978)	E	STP(E,8)	0.33	0.76	0.74	8
151 Information measure in N-terminal helix (Robson&Suzuki, 1976)	E	STP(E,5)	0.45	0.70	-0.54	5
152 Information measure in C-terminal helix (Robson&Suzuki, 1976)	E	STP(C,7)	0.61	0.74	-0.66	9
153 Information measure in extended state without H-bond (Robson&Suzuki, 1976)	F		0.08	0.41	0.46	6
154 Information measure in C-terminal turn (Robson&Suzuki, 1976)	F	STP(F,8)	0.15	0.54	0.57	8
155 Normalized frequency of isolated helix (Tanaka&Scheraga, 1977)	F		0.15	0.41	0.28	1
156 Normalized frequency of chain reversal R (Tanaka&Scheraga, 1977)	E	STP(E,9)	0.40	0.82	0.60	9
157 Normalized frequency of chain reversal S (Tanaka&Scheraga, 1977)	F	RES	0.23	0.68	0.62	8
158 Normalized frequency of chain reversal D (Tanaka&Scheraga, 1977)	F		0.48	0.65	-0.61	4
159 Normalized frequency of left-handed helix (Tanaka&Scheraga, 1977)	E	STP(E,2)	0.33	0.84	-0.54	2
160 Normalized frequency of ζ_R (Tanaka&Scheraga, 1977)	F	STP(F,1)	0.08	0.35	-0.20	3
161 Normalized frequency of coil (Tanaka&Scheraga, 1977)	E	STP(E,6)	0.46	0.93	-0.60	1
162 Conformational preference of p β-structure (Lifson&Sander, 1979)	C	STP(C,4)	0.73	0.83	0.79	5
163 Normalized frequency of bend R (Isogai et al., 1980)	E	STP(C,9)	0.50	0.76	0.70	9
164 Normalized frequency of bend S (Isogai et al., 1980)	E	STP(C,9)	0.51	0.87	0.71	9
165 Normalized frequency of helix end (Isogai et al., 1980)	E	STP(E,5)	0.14	0.93	-0.35	5
166 Normalized frequency of coil (Isogai et al., 1980)	E	STP(E,9)	0.39	0.89	0.59	9
167 Normalized frequency of ζ_R (Maxfield&Scheraga, 1976)	E	STP(E,1)	0.17	0.74	-0.36	6
168 Normalized frequency of left-handed helix (Maxfield&Scheraga, 1976)	E	STP(E,9)	0.41	0.93	0.58	9
169 Normalized frequency of ζ_L (Maxfield&Scheraga, 1976)	E	STP(E,6)	0.32	0.91	-0.51	6
170 Average relative fractional occurrence in A_O(i) (Rackovsky&Scheraga, 1982)	F	STP(F,2)	0.40	0.54	-0.59	4
171 Average relative fractional occurrence in A_R(i) (Rackovsky&Scheraga, 1982)	E	RES	0.52	0.83	0.65	7
172 Average relative fractional occurrence in A_L(i) (Rackovsky&Scheraga, 1982)	E	STP(C,3)	0.50	0.74	-0.62	3
173 Average relative fractional occurrence in E_R(i) (Rackovsky&Scheraga, 1982)	E	STP(E,6)	0.22	0.72	0.44	9
174 Average relative fractional occurrence in A_O(i-1) (Rackovsky&Scheraga, 1982)	F	RES	0.10	0.68	0.44	8
175 Average relative fractional occurrence in A_L(i-1) (Rackovsky&Scheraga, 1982)	E	STP(E,9)	0.54	0.93	0.67	9
176 Average relative fractional occurrence in E_L(i-1) (Rackovsky&Scheraga, 1982)	E	STP(C,8)	0.64	0.74	0.71	9
177 Average relative fractional occurrence in E_R(i-1) (Rackovsky&Scheraga, 1982)	F	STP(F,2)	0.04	0.44	-0.26	2
178 Value of angle θ for ith residue (Rackovsky&Scheraga, 1982)	E	STP(E,8)	0.46	0.87	0.71	8
179 Value of angle θ for (i-1)th residue (Rackovsky&Scheraga, 1982)	F	STP(F,6)	0.53	0.68	-0.62	7
180 pK-N (Fasman, 1976)	F	STP(F,1)	0.20	0.50	0.25	2
181 Entropy of formation (Hutchens, 1970)	C	STP(C,1)	0.82	0.93	0.71	1
182 Melting point (Hutchens, 1970)	E	RES	0.23	0.80	0.47	2
183 Composition (Grantham, 1974)	E	STP(E,2)	0.26	0.71	0.56	8

184	Isoelectric point (Zimmerman et al., 1968)	F	STP(F,2)	0.34	0.56	0.29	1
185	Principal component I (Sneath, 1966)	F		0.55	0.64	0.54	7
186	Principal component II (Sneath, 1966)	E		0.60	0.76	-0.48	1
187	Principal component III (Sneath, 1966)	E	STP(C,1)	0.70	0.79	0.73	1
188	Principal component IV (Sneath, 1966)	F		0.20	0.46	0.46	8

[a] Name: The designation + in front of the name means that this property has one or two missing values for amino acids which are shown in ( ) with the error of the estimated values (! means that the estimated value is not reliable). The error is measured in the same units as used for each physical property.

COR: The results of the correlation analyses.

A: VAR4 > 0.7, LCOR > 0.9 (For 4 factors: $)

B: VAR4 > 0.7, 0.8 < LCOR < 0.9

C: VAR4 > 0.7, LCOR < 0.8

D: VAR4 < 0.7, VAR10 > 0.7, LCOR > 0.8

E: VAR4 < 0.7, VAR10 > 0.7, LCOR < 0.8 (For 6 factors: %)

F: VAR10 < 0.7

ST: Statistical test of distribution of the physical property.

STP: far from normal distribution in the form STP(COR, P) where COR and P were estimated by Kendall's rank correlation which is distribution-free (Kendall and Stuart, 1977).

RES: far from normal distribution for the residual part after subtracting the correlation with the first 4 factors (Only for COR = E or F).

US: Usage.

$: for the first 4 factors.

%: for the last 6 factors.

VAR4: Variance accounted for by the first 4 factors.

VAR10: Variance accounted for by the 10 factors.

LCOR: The largest value of the correlation coefficients with the 9 characteristic properties in Table III.

P: One of the 9 characteristic properties which gives the largest correlation (LCOR) with the physical property.

Table II. Proteins Used to Recalculate Some Normalized Frequencies

Adenylate kinase (2ADK)
Calcium-binding parvalbumin (3CPV)
Carbonic anhydrase B (1CAB)
Carboxypeptidase A (5CPA)
α-Chymotrypsin (2CHA)
Concanavalin A (2CNA)
Cytochrome b5 (2B5C)
Cytochrome c (inner chain) (3CYT)
Ferredoxin (1FDX)
Flavodoxin (3FXN)
Glyceraldehyde-3-p-dehydrogenase (green subunit) (1GPD)
High potential iron protein (1HIP)
Immunoglobulin B-J fragment (A-chain) (1REI)
Lactate dehydrogenase (4LDH)
Lysozyme (6LYZ)
Myoglobin (1MBN)
Neurotoxin B (1NXB)
Papain (8PAP)
Prealbumin (B-chain) (2PAB)
Ribonuclease A (1RN3)
Staphylococcal nuclease (2SNS)
Subtilisin BPN′ (1SBT)
Superoxide dismutase (orange monomer) (2SOD)
Thermolysin (3TLN)
Triose phosphate isomerase (A-chain) (1TIM)
Trypsin inhibitor (4PTI)

is the total number of residues in the whole data base, N_{tj} is the number of occurrences of residue j in all conformational states, and N_{it} is the number of occurrences of conformational state i in the whole data base. These data had to be recalculated because their original definition (which was suitable for their matrix treatment) obtained the frequency of occurrence relative to that of the extended state, and this cannot be regarded as an independent property for the 20 amino acids. The normalized frequency of occurrence of chain reversals (Tanaka and Scheraga, 1977) (no. 98) was calculated by summing the data on their R, S, and D states [also using formula (1)].

6. The normalized frequencies of Isogai *et al.* (1980) (nos. 71, 83, 99, 113, 163–166) and the normalized frequencies of Maxfield and Scheraga (1976) (nos. 72, 84, 114, 167–169) were recalculated by using the X-ray crystallographic coordinates of the 26 nonhomologous proteins listed in Table II [from the Protein Data Bank (Bernstein *et al.*, 1977)]. The total number of amino acid residues in this data base is 4195.

2.2. Statistical Test for Normal Distribution of Physical Properties

First of all, in order to compare the numerical values of various physical properties, they must all be expressed on a common (standardized) numerical basis.

This is accomplished by rewriting each value as

$$Q = \frac{\text{original value} - \text{mean}}{\text{standard deviation}} \tag{2}$$

where the mean and the standard deviation for a given physical property are evaluated for the 20 amino acids. This standardized quantity Q has a mean of 0 and a standard deviation of 1.0. Thus, roughly half of the amino acids have positive values of the standarized physical property and roughly half have negative values. This standardization of each physical property is statistically meaningful only if the distribution of the original values is not far from normal. It is also necessary that the distribution not be far from normal in order to be able to apply the factor analysis model used here, namely the maximum-likelihood method, because the latter is based on a multivariate normal distribution.

In this study, we adopted the following four criteria to determine whether the distribution of each standardized property is normal:

1. The χ^2 statistics test should satisfy a 0.05 significance level when there are seven degrees of freedom.
2. The χ^2 statistics test should satisfy a 0.05 significance level when there are two degrees of freedom.
3. $-0.77 <$ the third moment < 0.77.
4. $0.95 <$ the fourth moment < 4.05.

Since we have two equations describing a distribution (one for the mean and one for the standard deviation), we would have seven degreees of freedom (criterion 1) if the values of the property were divided into ten intervals with nine partitions between these intervals, where the distribution is characterized by the number of values in each interval. Correspondingly, five intervals with four partitions would lead to two degrees of freedom (criterion 2). A standardized property having a sufficiently large number of elements N distributed normally has a third moment equal to $0 \pm (6/N)^{1/2}$ and a fourth moment equal to $3 \pm (24/N)^{1/2}$ (Kendall and Stuart, 1977). However, when N is very small (e.g., if $N = 20$), the errors in the third and fourth moments would be expected to be much larger than the values cited above, and the mean of the fourth moment could be smaller than 3. (The mean of the third moment would be 0 independent of N.) Hence, to allow for an increase in the error, we used the arbitrary values $0 \pm (6/10)^{1/2}$ and $2.5 \pm (24/10)^{1/2}$ as criteria for the third and fourth moments, respectively.

Of the 188 physical properties, only 116 met the four criteria, and only these 116 were used in the following analyses. The 72 properties whose distributions are far from normal are indicated by the designation STP in the fourth column of Table I.

2.3. Classification of Properties

These 116 physical properties of the amino acids have many duplications, i.e., many physical properties with a similar definition (but parametrized differently by different authors) are highly correlated. We can eliminate such duplications from

the data set by a proper classification of these physical properties. For this purpose, we made use of two cluster analysis methods (Hartigan, 1975), a K-mean cluster analysis and a hierarchical cluster analysis. A K-mean cluster analysis, which is a minimization procedure, can give the optimum clustering of the properties into K clusters, but suffers from the multiple-minima problem when the number of properties is large. Therefore, in order to avoid this problem, we first obtain an initial (approximate) clustering by using a hierarchical cluster analysis, which is a binary-tree type clustering, and then carry out a K-mean cluster analysis (starting from the results of the hierarchical cluster analysis).

The various physical properties were first classified roughly, as described in Appendix A, by a heierarchical cluster analysis with a single-linkage algorithm in terms of the absolute value of the correlation coefficients $|C_{ik}|$ between properties i and k, defined by

$$|C_{ik}| = (1/20)\left| \sum_{j=1}^{20} x_{ij}x_{kj} \right| \tag{3}$$

where x_{ij} is the numerical value of the ith standardized property for the jth amino acid; $|C_{ik}|$ varies between 0 and 1.0 for noncorrelated and highly correlated properties, respectively. In this procedure, a cluster of properties is defined by the smallest cluster for which the correlation coefficient with any other cluster satisfies the relation

$$|C_{ik}| < 0.85 \tag{4}$$

where $|C_{ik}|$ is the absolute value of the correlation coefficient between two clusters i and k, which is defined by the single-linkage algorithm (see Appendix A).

Each cluster was then classified further by a nonhierarchical cluster analysis, namely a K-mean cluster analysis (see Appendix A), in terms of the distance D_{ic} between properties i and the centroid of the cluster containing these properties;

$$D_{ic} = \left[(1/20) \sum_{j=1}^{20} (x_{ij} - \langle x_j \rangle)^2 \right]^{1/2} \tag{5}$$

where $\langle x_j \rangle$ is the jth coordinate of the centroid of the cluster. In this step, the combination of properties into clusters was carried out if the following condition was satisfied:

$$C_{ic} > 0.9 \tag{6}$$

where C_{ic} is the correlation coefficient between a property in the cluster and the centroid of the cluster.

The details of the methods are given in Appendix A.

Then each cluster was converted into a new, average property (which we designate as a *characteristic property*) by the principal component method described in Appendix B. (Section 2.4 should be studied before reading Appendix B.)

2.4. Factor Analysis (Partial Correlation Factor Analysis)

The basic assumption of the factor analysis model (Lawley and Maxwell, 1971) is that n standardized properties $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{i20})$ $(i = 1, \ldots, n)$ can be approxi-

mated by a linear combination of a smaller number NF ($<n$) of orthogonal factors $\mathbf{f}_r (r = 1, \ldots, \text{NF})$. In the form of a matrix representation,

$$\mathbf{X} = \mathbf{AF} + \mathbf{U} \tag{7}$$

where $\mathbf{X}$ is an $n \times 20$ matrix $[= (\mathbf{x}_1, \ldots, \mathbf{x}_n)^t]$, $\mathbf{F}$ is an $\text{NF} \times 20$ factor matrix $[= (\mathbf{f}_1, \ldots, \mathbf{f}_{\text{NF}})^t]$, $\mathbf{A}$ is an $n \times \text{NF}$ factor loading matrix (which represents the relative weight given to each factor), and $\mathbf{U}$ is an $n \times 20$ matrix $[= (\mathbf{u}_1, \ldots, \mathbf{u}_n)]$, which represents the residual part. (The term $\mathbf{U}$ arises because NF factors cannot describe n properties *exactly*, since $\text{NF} < n$.) The following orthogonality and normalization relations are assumed:

$$\mathbf{FU}^t = \mathbf{0}, \qquad (1/20)\mathbf{FF}^t = \mathbf{I}, \qquad (1/20)\mathbf{UU}^t = \mathbf{V} \tag{8}$$

where $\mathbf{V}$ is a diagonal matrix $[= \text{diag}(v_1, \ldots, v_n)]$ and v_i is the relative variance in the ith property that cannot be accounted for by NF factors. Geometrically, the $n \times 20$ matrix $\mathbf{X}$ is reduced in size to the $\text{NF} \times 20$ matrix $\mathbf{F}$. *The components of* $\mathbf{X}$, $\mathbf{A}$, $\mathbf{F}$, *and* $\mathbf{U}$ *can be positive, negative or zero.* Multiplication of $\mathbf{X}$ and $\mathbf{X}^t$ of Eq. (7), with the use of Eq. (8), leads to

$$\mathbf{S} = \mathbf{AA}^t + \mathbf{V} \tag{9}$$

where, according to Eq. (3), $\mathbf{S}$ $[= (1/20)\mathbf{XX}^t]$ is the correlation coefficient matrix of $\mathbf{X}$. Equation (9) is the fundamental equation that has to be solved by a factor analysis.

In order to obtain $\mathbf{F}$ (to describe the physical properties of the amino acids with a minimum number of parameters), we carry out a factor analysis in the following three stages (the details are given in Appendix C):

1. The maximum-likelihood method (Jöreskog, 1967) is used to solve Eq. (9) to obtain $\mathbf{A}$ and $\mathbf{V}$.

2. In an NF-dimensional subspace of the n-dimensional space, $\mathbf{F}$ can be defined arbitrarily. For this definition, we use an orthogonal transformation in order to simplify the interpretation of the physical meaning of each factor [each row vector of the matrix $\mathbf{F}$ in Eq. (7)]. This is accomplished by rotating $\mathbf{A}$ with an orthogonal transformation matrix by using the varimax criterion (Kaiser, 1958), which enables each factor to correlate as small a number of properties as possible (as in a normal-mode analysis); i.e., this procedure enables each factor to correspond as closely as possible to a given standardized property.

3. $\mathbf{F}$ of Eq. (7) is then computed from the resulting values of $\mathbf{A}$ and $\mathbf{U}$ (i.e., $\mathbf{V}$) by using the Anderson–Rubin (1956) method, which yields an exactly orthogonal solution for the factor $\mathbf{F}$. The Anderson–Rubin procedure is required because the orthogonal transformation of $\mathbf{A}$ is a necessary but not a sufficient condition that the $\mathbf{F}$'s be orthogonal.

It should be noted, however, that, if $n \geq 20$, the correlation matrix $\mathbf{S}$ in Eq. (9) is not positive-definite, because the maximum rank of $\mathbf{S}$ is only 19 (the rank of $\mathbf{S}$ is the same as the rank of $\mathbf{X}$, which is 19, rather than 20, because of the standardization condition $\sum_{j=1}^{20} x_{ij} = 0$). For such a nondefinite correlation matrix, we would have to use a particular type of factor analysis method, namely the principal factor analysis method (Harman, 1976). However, two unavoidable problems make this method

unsatisfactory for our purpose: (1) the contribution from the negative roots of **S** to the value of **F** leads to an overfactorization (i.e., overestimation of the factor loading **A**) and reduces the reliability of the solution; (2) if n is large, **F** becomes a mixture of all of the n properties even after an orthogonal rotation using the varimax criterion.

In order to circumvent these problems, we developed the partial correlation factor analysis method. This method performs a factor analysis in a stepwise manner. In the analysis carried out here, we used two steps separately for two sets of physical properties given by the cluster analysis; the first and second sets include 79 and 37 physical properties, respectively (see Section 3.1 for details). In the first step, stages 1-3 of the factor analysis method were applied to a set of nine characteristic properties $\mathbf{X}_1$ (which has almost all the information of the first 79 physical properties) to obtain the first factor solution $\mathbf{F}_1$ (which consisted of four factors). In the second step, the factor analysis was applied to another set of 16 physical properties $\mathbf{X}_2$ (which were selected from the second 37 physical properties according to the criterion given in Section 3.2) to yield a factor solution $\mathbf{F}_2$ (which consisted of six factors). In this procedure, the correlation coefficient matrix for $\mathbf{X}_2$ was assessed with the partial correlation matrix in terms of $\mathbf{F}_1$ (i.e., after subtracting the contribution from $\mathbf{F}_1$) instead of with the usual correlation matrix **S** in Eq. (9). Since $\mathbf{F}_1\mathbf{F}_2^t = \mathbf{0}$, where $\mathbf{F}_2$ is the factor solution in the second step, $\mathbf{F}_1$ (4×20) and $\mathbf{F}_2$ (6×20) form an orthogonal set of the factor solution

$$\left[\mathbf{F} = \begin{pmatrix}\mathbf{F}_1\\ \mathbf{F}_2\end{pmatrix}\right]$$

where **F** is a 10×20 matrix. Even if n is very large, this procedure enables us to obtain $\mathbf{F}_1$, using only $\mathbf{X}_1$, and $\mathbf{F}_2$ from $\mathbf{X}_2$ (after subtracting the contribution from $\mathbf{F}_1$), and so on (if desired, although only two steps were used here).

In order to calculate $\mathbf{F}_2$ in the second step, Eq. (7) is rewritten as

$$\mathbf{X}_2 = \mathbf{A}_1\mathbf{F}_1 + \mathbf{A}_2\mathbf{F}_2 + \mathbf{U}_2 \tag{10}$$

where **A** and **F** in Eq. (7) were decomposed into two parts, i.e.,

$$\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2) \qquad \text{and} \qquad \mathbf{F} = \begin{pmatrix}\mathbf{F}_1\\ \mathbf{F}_2\end{pmatrix}$$

By multiplying the terms of Eq. (10) into $\mathbf{F}_1^t$ and making use of Eq. (8), we obtain the following expression for $\mathbf{A}_1$ (9×4):

$$\mathbf{A}_1 = \mathbf{X}_2\mathbf{F}_1^t \tag{11}$$

$\mathbf{A}_2$ (16×6) is the factor loading for $\mathbf{F}_2$ (6×20), and $\mathbf{U}_2$ is a (16×20) matrix representing the residual part. Then, instead of Eq. (9), the equation that has to be solved by the maximum-likelihood method (see Appendix C) is

$$\mathbf{S}_2 = (1/20)\mathbf{X}_2\mathbf{X}_2^t - \mathbf{A}_1\mathbf{A}_1^t = \mathbf{A}_2\mathbf{A}_2^t + \mathbf{V}_2 \tag{12}$$

where $\mathbf{S}_2$ is the partial correlation matrix of $\mathbf{X}_2$ in terms of $\mathbf{F}_1$, and $\mathbf{V}_2 = (1/20)\mathbf{U}_2\mathbf{U}_2^t$.

In order to carry out the partial correlation factor analysis, the partial correlation of $\mathbf{X}_2$ in terms of $\mathbf{F}_1$ must be statistically significant; i.e., $\mathbf{X}_1$ and $\mathbf{X}_2$ should have a low correlation with each other. This is another reason for carrying out the classification of properties prior to the factor analysis.

The procedures of Sections 2.2–2.4 were carried out first by omitting those physical properties that had one or two missing values (properties preceded by a plus sign in Table I). These missing values were then estimated by the method of Section 2.5, based on the result of this preliminary factor analysis. Then, the physical properties with the estimated missing values were included in a second application of the procedures of Sections 2.2–2.4 to all 116 physical properties. These procedures were iterated until the values of F converged.

2.5. Missing Values of Some Properties

The 24 properties that are reported for only 18 or 19 of the 20 amino acids are indicated by a plus sign in front of their designations in Table I. Some of these missing values were estimated by a multidimensional regression analysis based on the results of the factor analysis. The details of the method are given in Appendix D. The errors in these estimated values are also given in Table I.

3. RESULTS AND DISCUSSION

3.1. Classifications of Amino Acid Properties

Figure 1 shows a flow chart to obtain the final factor solution from the original physical properties. The details of each step are explained below.

The hierarchical cluster analysis for the 116 physical properties that satisfied the statistical test of the distribution gave only four major clusters, containing 13, 42, 11, and 13 physical properties, respectively:

1. Bulk-related (nos. 1–13 in Table I).
2. Hydrophobicity + β-structure preference-related (nos. 15–18, 22–25, 30–50, 57–66, and 70–72).
3. α-Helix preference-related (nos. 74–84).
4. Bend-structure preference-related (nos. 89–92, 94–102).

Of the remaining 37 physical properties, the partial specific volume (nos. 105, 106 in Table I) and average composition (nos. 119, 120) each form a cluster within which the properties are highly correlated. The remaining 33 physical properties have low correlations with one another; this was also confirmed by the hierarchical cluster analysis with partial correlation, i.e., after subtracting the contributions from the four major clusters cited above.

The K-mean cluster analysis further classified these four clusters into nine clusters and some outliers, namely

1. Bulk (nos. 1–11 with two outliers)
2. Hydrophobicity 1 (nos. 15–18)
3. Hydrophobicity 2 (nos. 22–24 with one outlier)
4. Hydrophobicity 3 (nos. 30–45 with five outliers)
5. β-Structure preference 1 (nos. 57–62 with four outliers)
6. β-Structure preference 2 (nos. 70–72)
7. α-Helix preference (nos. 74–84)
8. Bend-structure preference 1 (nos. 89–92)
9. Bend-structure preference 2 (nos. 94–99 with three outliers)

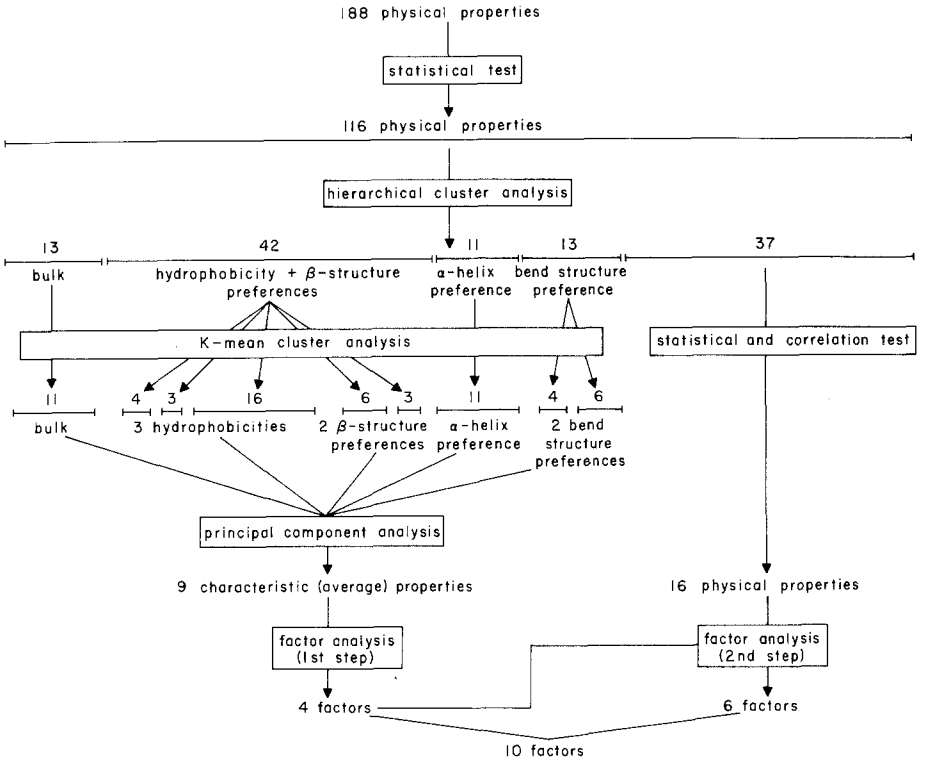

Fig. 1. Flow chart of the statistical analysis to obtain the ten factors. Values above the horizontal lines represent the number of physical properties. The gaps in the line under "*K*-mean cluster analysis" pertain to those physical properties referred to as "outliers" in the text.

where the outliers are the physical properties that do not pass the criterion of the *K*-mean cluster analysis but do satisfy the condition: $0.8 <$ (correlation coefficient between an outlier and the centroid of the *K*-mean cluster) < 0.9; the outliers are designated as B or D in the column COR of Table I, but were not used in the further analysis. These nine clusters were converted into the nine corresponding characteristic properties by the principal component analysis, and are shown in Table III. The nine characteristic properties accounted for a 92% variance of the 64 physical properties in the nine clusters. Since the correlation coefficient cannot be a good index for the classification of the 72 physical properties having problems in their distributions (designated as STP in Table I), these physical properties were also classified in terms of their rank correlation (Kendall and Stuart, 1977) with the nine characteristic properties; for properties that distribute far from normal, a rank correlation takes into account only the order of the amino acids, and not the distribution of the numerical values of any given property (i.e., a rank correlation is distribution-free), and provides a better index for the classification than the usual correlation coefficient.

The result of the classification by the *K*-mean cluster analysis shows the following characteristics of some of the amino acid properties.

1. All the physical properties designated as bulk are quite similar, and likewise for those designated as α-helix preference. These two kinds of physical properties seem to depend little on differences in their particular definitions.

2. The hydrophobicity properties can be classified into three types; the first two (1 and 2) are defined for free amino acids and the third one (3) is for amino acid residues in proteins. The correlation among types 1-3 is seen in Fig. 2. The difference between 1 and 2 is seen clearly in the amino acids having ionizable side chains (Arg, Asp, Glu, and Lys); type 1 suppresses the charge contribution of the ionizable side-chain groups (e.g., by adding salt), while type 2 includes them. Hydrophobicity 3 is defined as the depth to which an amino acid residue is buried inside a protein (from X-ray crystallographic data), and the differences from 1 and 2 are: (a) The charge contribution from ionizable side-chain groups is almost the average between those of 1 and 2. (b) In type 3, the amino acids having bulky side chains with polar groups (Lys, Trp, and Tyr) tend to be located closer to the surface of a protein (lower scores) than expected from the other two hydrophobicity scales, while the amino acids having bulky hydrocarbon side chains (Ile, Leu, Met, Phe, and Val) have almost the same scores as the other two scales. This means that there is an effect of bulkiness of those side chains containing polar groups on hydrophobicity 3; the polar part of a side chain, even if the other part is hydrophobic, can be

Table III. Nine Characteristic Properties Obtained by Principal Component Analysis

Amino acid	Characteristic property[a]								
	1	2[b]	3	4	5	6	7	8	9
ALA	−1.44	−0.47	0.11	0.32	−0.51	−0.86	1.35	−1.29	−0.60
ARG	1.16	−0.57	−1.52	−1.07	−0.28	−0.13	−0.16	0.28	−0.03
ASN	−0.34	−1.25	−0.60	−0.96	−1.00	−1.19	−0.97	1.19	1.27
ASP	−0.54	−0.75	−1.74	−1.07	−1.17	−1.72	−0.06	0.74	1.39
CYS	−0.75	0.06	0.63	1.50	0.60	1.14	−0.53	1.18	−0.19
GLN	0.22	−1.24	−0.46	−1.05	0.19	−0.42	0.57	−0.14	−0.12
GLU	0.17	−0.62	−1.65	−1.03	−1.74	−1.78	1.96	−1.21	−0.27
GLY	−2.16	−1.02	−0.19	−0.03	−0.84	−0.99	−1.72	1.43	1.73
HIS	0.52	−0.46	−0.18	−0.13	−0.56	−0.10	0.59	−0.27	−0.27
ILE	0.21	1.37	0.97	1.52	1.91	1.27	0.06	−1.30	−1.49
LEU	0.25	1.06	1.01	1.14	0.69	0.02	0.93	−1.36	−1.14
LYS	0.68	−0.16	−1.62	−1.76	−0.86	−1.19	0.71	0.40	0.15
MET	0.44	0.20	0.72	1.00	0.45	0.24	1.39	−1.24	−1.29
PHE	1.09	1.46	1.24	1.16	0.88	0.48	0.37	−0.46	−0.75
PRO	−0.71	0.90	0.21	−0.72	−1.26	0.86	−1.72	1.03	1.98
SER	−1.21	−1.19	−0.33	−0.46	−0.54	0.22	−0.99	0.74	1.02
THR	−0.67	−0.97	0.01	−0.36	0.57	0.86	−0.68	0.11	0.14
TRP	2.08	2.06	1.55	0.67	0.61	0.42	0.23	0.83	−0.52
TYR	1.34	1.16	1.04	−0.07	1.02	1.21	−1.25	0.94	0.30
VAL	−0.34	0.42	0.77	1.38	1.84	1.66	−0.09	−1.63	−1.32

[a] 1: bulk; 2,3,4: hydrophobicity; 5,6: β-structure preference; 7: α-helix preference; 8,9: bend-structure preference (See text for details). The values of each property are standardized.

[b] To obtain property 2, hydrophobicity factor (No. 18) (Goldsack & Chalifoux, 1973) was not used because the estimations of the two missing values are not reliable.

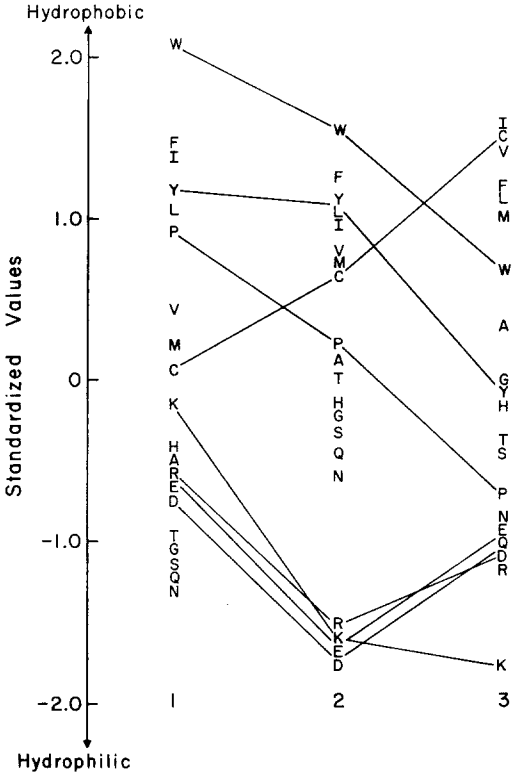

Fig. 2. Comparison of the three hydrophobicity indices of Table III. (1) Hydrophobicity 1; (2) hydrophobicity 2; (3) hydrophobicity 3. The amino acids are represented by the one letter code. It should be noted that 1-3 here correspond to properties 2-4 in Table III.

exposed to water. (c) The high score for Cys and low score for Pro reflect the specific structural characteristics of proteins, i.e., the disulfide bond tends to be buried inside a protein and the bend structure (which frequently contains proline) tends to be found on the outside.

3. Two kinds of β-structure preferences arise from two different definitions; β-structure preference 1 is defined mainly by the interstrand distance, and that of 2 is defined by the dihedral angles. The difference between β-structure preferences 1 and 2 is especially large for Pro (see Table III). The high score for Pro in β-structure preference 2 arises because this definition regards the collagen helix-like structure of Pro ($\phi \approx -75°$ and $100° < \psi < 180°$) as a β-structure.

4. There are three kinds of definitions for the bend-structure preference: the distance between C_i and $C_{i+3} < 5.7$ Å (Crawford *et al.*, 1973); the rotational angle defined by the three virtual bonds connecting four C^α's $\sim 0°$ (Robson and Suzuki, 1976); the distance between C_i and $C_{i+3} < 7$ Å (all the other authors). The first two

Table IV. Correlation Coefficient [S of Eq. (9)] between Nine Characteristic Properties

1	1.00								
2	0.59	1.00							
3	0.22	0.73	1.00						
4	0.05	0.61	0.85	1.00					
5	0.33	0.58	0.76	0.77	1.00				
6	0.19	0.57	0.76	0.66	0.80	1.00			
7	0.33	0.06	−0.11	0.10	0.00	−0.33	1.00		
8	−0.13	−0.21	−0.19	−0.42	−0.38	−0.11	−0.75	1.00	
9	−0.45	−0.46	−0.46	−0.64	−0.70	−0.38	−0.69	0.82	1.00
	1	2	3	4	5	6	7	8	9

1: bulk; 2,3,4: hydrophobicity; 5,6: β-structure preference; 7: α-helix preference; 8,9: bend-structure preference.

lead to the characteristic property of bend-structure preference 1, and the third leads to that of 2. The first two definitions seem to be much more restricted than the third.

Further, two important relations among the nine characteristic properties may be deduced from their correlation coefficients, which are shown in Table IV: (1) there is a positive correlation between hydrophobicity and β-structure preference; i.e., those residues with high hydrophobicity have high β-structure preference. Because of the interstrand hydrogen bonds, an amino acid residue in a β structure cannot be completely exposed to water. Hence, a hydrophobic residue is more stable in a β structure. (2) There is a negative correlation between either α-helix or β-structure preference and bend-structure preference. Bend structures are frequently located at the end of an α-helix or a β-structure. Thus, an amino acid that has a high bend-structure preference tends to avoid α- or β-structures, and hence to terminate an α-helix or a β-structure.

3.2. Factor Analysis

In order to avoid the problems mentioned in Section 2, we used the partial correlation factor analysis, which consists of two steps. In the first, four factors were calculated from the nine characteristic properties in Table III. In the second step, six factors, each of which is orthogonal with respect to the first four factors and with respect to each other, were obtained from the other 16 physical properties (see Fig. 1). The number of factors in each step (four and six, respectively) was determined so as to be the least number that satisfies the criterion of the χ^2 statistics test given in Appendix C [Eq. (C6)]. The 16 physical properties for the last six factors were chosen from 37 by the following criteria:

1. As described in Section 2, the physical property considered in the second step must have a low correlation with the first four factors or the nine characteristic

Table V. Final Ten Factors Used to Describe 86% of the Original 188 Physical Properties

Amino acid	Factor[a]									
	1	2	3	4	5	6	7	8	9	10
ALA	−1.56	−1.67	−0.97	−0.27	−0.93	−0.78	−0.20	−0.08	0.21	−0.48
ARG	0.22	1.27	1.37	1.87	−1.70	0.46	0.92	−0.39	0.23	0.93
ASN	1.14	−0.07	−0.12	0.81	0.18	0.37	−0.09	1.23	1.10	−1.73
ASP	0.58	−0.22	−1.58	0.81	−0.92	0.15	−1.52	0.47	0.76	0.70
CYS	0.12	−0.89	0.45	−1.05	−0.71	2.41	1.52	−0.69	1.13	1.10
GLN	−0.47	0.24	0.07	1.10	1.10	0.59	0.84	−0.71	−0.03	−2.33
GLU	−1.45	0.19	−1.61	1.17	−1.31	0.40	0.04	0.38	−0.35	−0.12
GLY	1.46	−1.96	−0.23	−0.16	0.10	−0.11	1.32	2.36	−1.66	0.46
HIS	−0.41	0.52	−0.28	0.28	1.61	1.01	−1.85	0.47	1.13	1.63
ILE	−0.73	−0.16	1.79	−0.77	−0.54	0.03	−0.83	0.51	0.66	−1.78
LEU	−1.04	0.00	−0.24	−1.10	−0.55	−2.05	0.96	−0.76	0.45	0.93
LYS	−0.34	0.82	−0.23	1.70	1.54	−1.62	1.15	−0.08	−0.48	0.60
MET	−1.40	0.18	−0.42	−0.73	2.00	1.52	0.26	0.11	−1.27	0.27
PHE	−0.21	0.98	−0.36	−1.43	0.22	−0.81	0.67	1.10	1.71	−0.44
PRO	2.06	−0.33	−1.15	−0.75	0.88	−0.45	0.30	−2.30	0.74	−0.28
SER	0.81	−1.08	0.16	0.42	−0.21	−0.43	−1.89	−1.15	−0.97	−0.23
THR	0.26	−0.70	1.21	0.63	−0.10	0.21	0.24	−1.15	−0.56	0.19
TRP	0.30	2.10	−0.72	−1.57	−1.16	0.57	−0.48	−0.40	−2.30	−0.60
TYR	1.38	1.48	0.80	−0.56	−0.00	−0.68	−0.31	1.03	−0.05	0.53
VAL	−0.74	−0.71	2.04	−0.40	0.50	−0.81	−1.07	0.06	−0.46	0.65

[a] These values are standardized. 1: α-helix or bend-structure preference-related; 2: bulk-related; 3: β-structure preference-related; 4: hydrophobicity-related; 5 to 10: Mixture of several physical properties. See Section 3.2 for details.

properties in Table III. We used the criteria:

[variance (expressing the correlation) accounted for by the first four factors]<0.7
(VAR4 in Table I)

and

(correlation coefficient with the nine characteristic properties)<0.8
(LCOR in Table I)

2. Not only the physical property itself, but also the residual part, after subtracting the contribution from the first four factors [i.e., $\mathbf{X}_2 - \mathbf{A}_1\mathbf{F}_1$ in Eq. (10)], has to pass the criteria in the statistical test on its distribution (Section 2.2).

3. When a physical property is used in the second step, the factors obtained in the analysis should account well for the property. The criterion used was

(variance accounted for by the ten factors)>0.7
(VAR10 in Table I)

The 16 properties (nos. 105-120) used in the second step are designated by a % mark in Table I. Here, we used two *average* properties, partial specific volume (nos. 105, 106) and average composition (nos. 119, 120), respectively, i.e., two

Table VI. Correlation Coefficient [S of Eq. (9)] between Nine Properties and Four Factors

Characteristic property	Factor 1	2	3	4
1. Bulk	−0.17	0.98	0.10	−0.09
2. Hydrophobicity 1	−0.06	0.51	0.11	−0.73
3. Hydrophobicity 2	−0.04	0.11	0.30	−0.89
4. Hydrophobicity 3	−0.33	−0.12	0.36	−0.83
5. β-Structure preference 1	−0.24	0.17	0.73	−0.55
6. β-Structure preference 2	0.11	0.10	0.68	−0.59
7. α-Helix preference	−0.93	0.21	−0.29	0.09
8. Bend-structure preference 1	0.88	0.05	−0.15	0.13
9. Bend-structure preference 2	0.83	−0.25	−0.37	0.33

average properties instead of the original four physical properties. Hence, we used 14 properties to obtain six factors. The values of the ten factors are listed in Table V. Table VI shows the correlation coefficients between the first four factors and the nine properties. From these data, the first four factors can be identified as follows:

Factor 1. α-Helix or bend-structure preference-related (it should be noted, however, that these two have a negative correlation[4] with each other as shown in Table IV).
Factor 2. Bulk-related.
Factor 3. β-Structure preference-related.
Factor 4. Hydrophobicity-related.

The last six factors are more or less mixtures of several physical properties. If they are identified by the property having the highest correlation, we obtain the following [the numbers are the correlation coefficient [S of Eq. (9)] of the given property with the given factor]:

Factor 5. Normalized frequency of double bend (no. 113) (−0.83).
Factor 6. Average value of average composition (nos. 119, 120) (0.66) or average value of partial specific volume (nos. 105, 106) (0.55).
Factor 7. Average relative fractional occurrence in E_0 (ith) (no. 115) (−0.50).
Factor 8. Normalized frequency of α-region (no. 114) (−0.73).
Factor 9. pK-C (no. 116) (−0.57).
Factor 10. Surrounding hydrophobicity in β-structure (no. 108) (−0.61).

[4] Factor 1 expresses the highly positive correlation for bend-structure preference and the highly negative correlation for α-helix preference. "Strong correlation" (whether positive or negative) means that the given factor contributes strongly to the given property. For example, for a given property, x_j for the jth amino acid may be written as

$$x_j = a_1 F_{1j} + a_2 F_{2j} + \cdots + a_{10} F_{10j} + u_j$$

The sign of x_j (and the signs of the a_i's) may be positive or negative because the sign of x_j (as a standardized property with a mean of 0 and a standard deviation of 1.0) depends on the (arbitrary) designation of the ends of the standardized scale as positive or negative. Thus, the absolute value, and not the sign of the correlation coefficient S, is the important quantity.

The variances accounted for by four or ten factors and the correlation coefficent with one of nine characteristic properties (Table III) are also shown (as LCOR) in Table I.

Finally, the first four factors account for a 91% variance of the nine characteristic properties and a 68% variance of the 188 physical properties; the last six factors account for an 85% variance of the residual part of the 16 physical properties after subtracting the correlation with the first four factors and an 18% variance of the 188 properties; thus, these ten factors account for an 86% variance of the 188 physical properties. Now we have two sets of parameters, the nine characteristic properties in Table III, which are typical properties related to bulk, hydrophobicity, and the propensity to form ordered backbone structures, and the ten factors, which are a set of orthogonal properties that describe almost all kinds of properties. The application of these parameters to a consideration of protein structure is made elsewhere (Kidera *et al.*, 1985).

3.3. Estimation of Missing Values

The procedure described in Section 2.5 was used to obtain the missing values of some physical properties. The results of these computations are given in Appendix D.

APPENDIX A. CLUSTER ANALYSIS

In this Appendix, we describe a procedure for combining several correlated physical properties into one cluster. For this purpose, we made use of two cluster analysis methods (Hartigan, 1975), first a hierarchical cluster analysis with a single-linkage algorithm in terms of the absolute value of the correlation coefficients defined in Eq. (3), and then a K-mean cluster analysis in terms of the distance defined in Eq. (5).

A1. Hierarchical Cluster Analysis

A hierarchical cluster analysis makes use of a binary-tree type of clustering. Such an analysis is based on a stepwise algorithm using an $n \times n$ correlation matrix for n properties (n initial clusters), the $n \times n$ elements of this matrix being the absolute values of the correlation coefficients [Eq. (3)]. We first combine the two closest clusters (i.e., the two with the highest correlation coefficient), thereby forming a new cluster.[5] We then produce a new $(n-1) \times (n-1)$ correlation coefficient matrix by calculating the correlation coefficients between the new cluster and each of the remaining clusters. Here, we adopted a single-linkage algorithm to update the

[5] The correlation coefficient $|C_{km}|$ can be defined by either a single-linkage algorithm ($|C_{km}|$ = the larger of $|C_{im}|$ and $|C_{jm}|$) or a complete-linkage algorithm ($|C_{km}|$ = the smaller of $|C_{im}|$ and $|C_{jm}|$). We use a single-linkage algorithm here in order to produce rather large clusters so that the properties in such a cluster may have rather low correlations with each other, because in this initial procedure we do not want to restrict the clusters too severely; i.e., we want to reduce the number of properties before using the optimal method of cluster analysis, the K-mean cluster analysis.

correlation matrix; i.e., when a cluster k is formed by merging clusters i and j, the single-linkage algorithm takes the absolute value of the correlation coefficient between cluster k and one of the remaining clusters m [$|C_{km}|$, defined by Eq. (3)] to be the larger of $|C_{im}|$ and $|C_{jm}|$. This process is continued until the absolute value of the correlation coefficient between any two clusters defined by the single-linkage algorithm does not exceed 0.85 [see Eq. (4)]. The Institute for Mathematical and Statistical Subroutine Library (IMSL, 1982) OCLINK was used for the computation.

A2. *K*-Mean Cluster Analysis

Since the hierarchical cluster analysis is based on the arbitrary choice between the single- and complete-linkage alogorithms (neither of which is ideal), and does not give a sufficiently refined set of noncorrelated clusters, we carry out the additional procedure of a K-mean cluster analysis. This latter analysis is a minimization procedure for a given number of clusters K and a given initial clustering, in which n properties form K ($<n$) clusters. We begin by assuming the existence of an arbitrary number K of clusters, which were generated randomly. An object function, taken as the sum of the D_{ic}'s [of Eq. (5)] over these K clusters, is then minimized, and each resulting cluster is tested to see if it satisfies Eq. (6). The minimization is carried out by transferring properties from one cluster to another until the object function is minimized. If Eq. (6) is satisfied, the number of clusters K is reduced; if it is not satisfied, K is increased. In either case, the minimization of the resulting object function is repeated to obtain, finally, the minimum number of clusters K that satisfy Eq. (6). The procedure generally increases the number of clusters, but the properties within a given cluster are now more highly correlated. Since the results of these minimizations depend on the initial clustering (the multiple-minima problem), the computations were carried out with many initial clusterings, which were generated randomly, to make sure that the global minimum of the object function was attained. For this purpose, we used 100 starting cluster arrangements, and found that more than 50 of them attained the same (global) minimum when the algorithm of Späth (1980) was used for the minimization.

APPENDIX B. PRINCIPAL COMPONENT ANALYSIS

For the purpose of converting a cluster of similar physical properties into an average property, a one-factor model [Eq. (7), with NF = 1], i.e., a model with a one-to-one correspondence between a factor and the properties of a given cluster, cannot be used, because this model is valid only if the correlation coeffients satisfy the relation

$$C_{jk}C_{lm} = C_{lk}C_{jm} \tag{B1}$$

(Harman, 1976), where j, k, l, and m are physical properties under consideration for combination into a cluster. Since it is very unlikely for a set of properties to satisfy Eq. (B1), we avoid this problem (and still describe a cluster with one factor) by using a principal component analysis as an approximation, as described below.

The principal component model neglects the term **U** in Eq. (7), and the original physical properties **X** ($n \times 20$) are related to the average property **f** (1×20) and the

factor loading **a** ($n \times 1$) by Eq. (B2) rather than Eq. (7) (i.e., by using only one factor):

$$\mathbf{X} \approx \mathbf{af} \tag{B2}$$

The following procedure is used to obtain **f** of Eq. (B2). When NF = n in Eq. (7), then $\mathbf{U} = \mathbf{0}$, i.e., all properties are then described exactly by Eq. (7). Hence, under these circumstances, Eq. (7) may be rewritten as

$$\mathbf{X} = \mathbf{AF} \tag{B3}$$

where **A** is an $n \times n$ matrix and **F** is an $n \times 20$ matrix. Then, from Eq. (B3), the correlation coefficient matrix **S** is

$$\mathbf{S} = (1/20)\mathbf{X}\mathbf{X}^t = \mathbf{A}\mathbf{A}^t \tag{B4}$$

where the relation $(1/20)\mathbf{F}\mathbf{F}^t = \mathbf{I}$ of Eq. (8) was used. Let $\mathbf{\Lambda}$ and $\mathbf{\Omega}$ be an eigenvalue and an eigenvector matrix, respectively, of **S**. Then

$$\mathbf{\Omega}\mathbf{\Lambda}\mathbf{\Omega}^t = \mathbf{A}\mathbf{A}^t$$

or

$$\mathbf{A} = \mathbf{\Omega}\mathbf{\Lambda}^{1/2} \tag{B5}$$

Substituting Eq. (B5) into Eq. (B3) and multiplying $\mathbf{\Lambda}^{-1/2}\mathbf{\Omega}^t$ into both sides of this equation, we obtain the exact result (i.e., for NF = n)

$$\mathbf{F} = \mathbf{\Lambda}^{-1/2}\mathbf{\Omega}^t\mathbf{X} \tag{B6}$$

Now, we introduce an approximation. Suppose that the largest eigenvalue λ_1 of $\mathbf{\Lambda}$ is significantly larger than the other eigenvalues, and the row vectors of **A** are negligibly small except for **a**, which corresponds to λ_1 (i.e., $\mathbf{a} = \boldsymbol{\omega}_1\lambda_1^{1/2}$; $\boldsymbol{\omega}_1$ is the eigenvector corresponding to λ_1). In this maximum-eigenvalue approximation (equivalent to a one-factor model), **X** is approximated by Eq. (B2) in which the corresponding value of **f** is given by

$$\mathbf{f} = \lambda_1^{-1/2}\boldsymbol{\omega}_1^t\mathbf{X} \tag{B7}$$

Then, substituting the properties in a cluster that were given by the K-mean cluster analysis into **X**, we can obtain the corresponding average property **f**.

Finally, nine clusters of physical properties were converted into nine average properties **f**, i.e., nine characteristic properties, as given in Table III. Since each of the nine maximum eigenvalues λ_1 of each cluster accounts for an average of 92% of tr **S** ($= \sum_{i=1}^{n} \lambda_i$, where n is the number of properties in a given cluster), this demonstrates that the maximum-eigenvalue approximation and, hence, the approximation of Eq. (B2) are reasonably correct.

APPENDIX C. FACTOR ANALYSIS ALGORITHM

We present here some details of the factor analysis algorithm that enabled us to obtain the factor **F** of Eq. (7). For this purpose, we first obtain the factor loading **A**, then carry out an orthogonal transformation of the factor loading, and finally obtain the factor **F**.

C1. Estimation of Factor Loading (Jöreskog, 1967)

Of the various methods in the literature to solve Eq. (9), only the maximum-likelihood method (Jöreskog, 1967) avoids the problem of overfactorization. This method makes use of the concept of a population space.

As in the statistics of political poll-taking, one can sample a small portion (the sample space) of the total population (the population space) to obtain information that characterizes the total population. In this sense, we may consider that the 20 amino acids form a sample space in a population space that may be assumed to contain an infinite number of amino acids; i.e., we are trying to apply a statistics of large numbers to a small set (namely 20) of amino acid residues. Then, **S** of Eq. (9) can be regarded as a sample correlation coefficient matrix because we use only 20 amino acids to evaluate **S**. Also, we can define the corresponding population correlation coefficient matrix $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} \equiv \mathbf{A}\mathbf{A}^t + \mathbf{V} \tag{C1}$$

Both **S** and $\boldsymbol{\Sigma}$ are matrices of the same size $(n \times n)$, formed from property matrices **X** and $\mathbf{X}^t$ [of Eq. (7)] that are $(n \times 20)$, $(20 \times n)$ and $(n \times \infty)$, $(\infty \times n)$, respectively.

We can estimate $\boldsymbol{\Sigma}$ by the maximum-likelihood principle (Kendall and Stuart, 1977), by assuming that the properties **X** form a multinormal distribution. Thus, to obtain **A** and **V**, and hence $\boldsymbol{\Sigma}$, we maximize the likelihood function L of the Wishart distribution (which is a product of two Gaussian distributions) (Morrison, 1976); i.e., $\mathbf{S}\,[=(1/20)\mathbf{X}\mathbf{X}^t]$, being a product of two Gaussian distributions, is a Wishart distribution. Omitting the part of L that is independent of $\boldsymbol{\Sigma}$, because it plays no role in the maximization, we have

$$\ln L = -\ln \det \boldsymbol{\Sigma} - \mathrm{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}) \tag{C2}$$

Changing the sign in Eq. (C2), we obtain the object function that is to be minimized with respect to **A** and **V**, namely

$$J(\mathbf{A}, \mathbf{V}) = \ln \det \boldsymbol{\Sigma} + \mathrm{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}) \tag{C3}$$

First, J is minimized with respect to **A** for a given **V**, and then it is minimized with respect to **V**.

For a given **V**, Jöreskog (1967) has shown that **A** that minimizes J can be approximated by

$$\mathbf{A} = \mathbf{V}^{1/2}\boldsymbol{\Omega}(\boldsymbol{\Lambda} - \mathbf{I})^{1/2} \tag{C4}$$

where $\boldsymbol{\Lambda}$ (NF×NF) and $\boldsymbol{\Omega}$ $(n \times \mathrm{NF})$ are matrices composed of the NF largest eigenvalues and of the corresponding eigenvectors, respectively, of $\mathbf{V}^{-1/2}\mathbf{S}\mathbf{V}^{-1/2}$.

We next must minimize J with respect to **V**. Substituting Eq. (C4) into Eq. (C3) and neglecting terms that are independent of **V**, it can be shown that J may be rewritten as a function of **V** by means of the equation

$$J(\mathbf{V}) = \sum_{i=\mathrm{NF}+1}^{n} \lambda_i + \sum_{i=1}^{\mathrm{NF}} \ln \lambda_i + \sum_{i=1}^{n} \ln v_i \tag{C5}$$

where λ_i is the ith eigenvalue of $\mathbf{V}^{-1/2}\mathbf{S}\mathbf{V}^{-1/2}$ and v_i is the ith diagonal element of

V. The value of $J(\mathbf{V})$ of Eq. (C5) was minimized by the Newton-Raphson method, the computation being carried out by the IMSL library subroutine OFCOMM.

One of the advantages of the maximum-likelihood method is that it enables us to rationalize the hypothesis that we can neglect $i > \mathrm{NF}$, i.e., that NF factors are sufficient to describe $\boldsymbol{\Sigma}$ of Eq. (C1). We can rationalize this hypothesis by showing that NF satisfies the criterion (Bartlett, 1954)

$$H(\mathrm{NF}) = [19 - (2n+5)/6 - 2\mathrm{NF}/3][J(\mathbf{A}, \mathbf{V}) - \ln \det \mathbf{S} - n] \tag{C6}$$

where $H(\mathrm{NF})$ is distributed essentially as a χ^2 distribution with a degree of freedom $d(\mathrm{NF})$,

$$d(\mathrm{NF}) = (1/2)[(n - \mathrm{NF})^2 - (n + \mathrm{NF})] \tag{C7}$$

We determined the smallest value of NF that satisfies a 0.05 significance level of the criterion expressed by $H(\mathrm{NF})$. As shown in Section 3, the value of NF is 4 for nine characteristic properties and 6 for 16 physical properties.

C2. Orthogonal Transformation of the Factor Loadings

Equation (C1) is invariant in terms of an orthogonal transformation of **A**, namely

$$\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^t + \mathbf{V} = \mathbf{A}\mathbf{A}^t + \mathbf{V} \tag{C8}$$

where **B** $(n \times \mathrm{NF}) = \{b_{ij}\} = \mathbf{AT}$, and **T** $(\mathrm{NF} \times \mathrm{NF})$ is the orthogonal transformation matrix (i.e., $\mathbf{T}\mathbf{T}^t = \mathbf{I}$). The quantity b_{ij}^2 is a measure of how much variance of the ith property is accounted for by the jth factor. To obtain an interpretable solution for **A**, i.e., one that enables each factor to correlate as small a number of properties as possible, this matrix was rotated orthogonally by using the varimax criterion (Kaiser, 1958), which maximizes

$$n \sum_{j=1}^{\mathrm{NF}} \sum_{i=1}^{n} (b_{ij}/h_i)^4 - \sum_{j=1}^{\mathrm{NF}} \left[\sum_{i=1}^{n} (b_{ij}/h_i)^2 \right]^2 \tag{C9}$$

where $h_i = (\sum_{j=1}^{\mathrm{NF}} b_{ij}^2)^{1/2}$. This maximization allows a given jth factor to arise from a minimum number of properties (thereby providing a physical meaning to **A**) by maximizing the variance of the square of the normalized factor loadings $(b_{ij}/h_i)^2$ from the mean value $(1/n)\sum_{i=1}^{n} (b_{ij}/h_i)^2$ for all the factors. The IMSL library subroutine OFROTA was used for this calculation.

C3. Estimation of the Factor F

Since the factor loading matrix **A** in Eq. (7) has $n \times \mathrm{NF}$ elements $(\mathrm{NF} < n)$, there is no unique solution for the factor **F**. For the partial correlation factor analysis method, it is necessary to obtain an orthogonal solution, namely $\mathbf{F}_\mu \mathbf{F}_\nu^t = \mathbf{0}$, where $\mathbf{F}_\mu$ and $\mathbf{F}_\nu$ are two different factors. For this purpose, we chose the **F**'s so as to minimize $\mathbf{U}\mathbb{V}^{-1}\mathbb{U}$ under the constraint $(1/20)\mathbf{F}_\mu \mathbf{F}_\nu^t = \mathbf{I}$ [where **U** and **V** are given by Eqs. (7) and (8), respectively], by using the Anderson-Rubin (1956) method. The

optimum **F**'s obtained by the procedure are given by

$$\mathbf{F} = (\mathbf{A}^t\mathbf{V}^{-1}\mathbf{S}\mathbf{V}^{-1}\mathbf{A})^{-1/2}\mathbf{A}^t\mathbf{V}^{-1}\mathbf{X} \tag{C10}$$

Since $\mathbf{X}\mathbf{X}^t = 20\mathbf{S}$ and $\mathbf{V}^t = \mathbf{V}$ (**V** is a diagonal matrix), it follows that **F** of Eq. (C10) satisfies $\mathbf{F}\mathbf{F}^t = 20\mathbf{I}$; i.e., Eq. (C10) gives an orthogonal solution for **F**.

APPENDIX D. ESTIMATION OF MISSING VALUES

We employed two iterative procedures to estimate the missing values of a physical property for a given amino acid, but only for those properties for which the values for only one or two amino acids were missing. These estimates, however, could be made only after the factor analysis was carried out on the properties for which there were no missing values.

D1. Method 1

Equation (7) is rewritten for a physical property with missing values as

$$\mathbf{x} = \mathbf{a}\mathbf{F} + \mathbf{u} \tag{D1}$$

where **x** is a 1×20 row vector representing a physical property with one or two missing values, for which we make an initial guess of 0 (the average value of a standardized property); **F** is the $\mathrm{NF} \times 20$ factor matrix (which was first calculated by omitting those physical properties with missing values); **u** is a 1×20 row vector representing the residual part; and **a** is a $1 \times \mathrm{NF}$ row vector representing the factor loading for the property **x**, which, by multiplying both sides of Eq. (D1) into $\mathbf{F}^t$ and making use of Eq. (8), is found to be equal to $\mathbf{x}\mathbf{F}^t$. We estimate the missing value by using the equation

$$\mathbf{x}_1 = \mathbf{a}\mathbf{F} + \mathbf{u}_1 \tag{D2}$$

where **a** is now $\mathbf{x}\mathbf{F}^t$; the element of the residual part $\mathbf{u}_1$ for the missing value is approximated as 0 because we have no way to obtain its exact value, and $\mathbf{x}_1$ is a first approximation to **x**. The next improved estimate $\mathbf{x}_2$ is given by substituting $\mathbf{x}_1\mathbf{F}^t$ for **a**, i.e.,

$$\mathbf{x}_2 = \mathbf{x}_1\mathbf{F}^t\mathbf{F} + \mathbf{u}_2 \tag{D3}$$

where the element of $\mathbf{u}_2$ for the missing value is also approximated as 0. This procedure is iterated until the value of **a** converges.

When ten factors are used for **F** in this model, the iteration does not always converge, because too much freedom is allowed for the value of **a**; i.e., it is difficult to compute ten accurate components of **a** from 18 or 19 values. To avoid this problem, we used only the first four factors of Table V [these correspond primarily to: (1) α-helix-forming or bend-structure-forming preference, (2) bulkiness, (3) β-sheet-forming preference, and (4) hydrophobicity]. Hence, we can estimate the missing values with this method only for those physical properties that have a high correlation with these four factors.

D2. Method 2

In this second method, we also used only the first four factors of Table V. From Eq. (9), we find

$$\mathbf{C} = \mathbf{a}\mathbf{B}^t \tag{D4}$$

where **a** is the same as in Eq. (D1), **B** is a 9×4 matrix representing the factor loading matrix for the first four factors [$\mathbf{F}_4$ (4×20)] obtained from nine characteristic properties $\mathbf{X}_9$ (9×20) in Table III, i.e., $\mathbf{B} = \mathbf{X}_9\mathbf{F}_4^t$, and **C** is a 1×9 row vector that represents the correlation coefficients between **x** and the nine characteristic properties calculated by assuming the missing values to be 0 (i.e., assuming that the missing values do not affect the correlation coefficient). From Eq. (D4),

$$\mathbf{a}\mathbf{B}^t\mathbf{B} = \mathbf{C}\mathbf{B} \tag{D5}$$

where $\mathbf{B}^t\mathbf{B}$ is a 9×9 matrix. This leads directly to

$$\mathbf{a} = \mathbf{C}\mathbf{B}(\mathbf{B}^t\mathbf{B})^{-1} \tag{D6}$$

which provides a least-squares solution for **a**; i.e., the term $(\mathbf{C} - \mathbf{a}\mathbf{B}^t)^2$ from Eq. (D4) will take on a minimal value if Eq. (D6) is satisfied. Then **a** is substituted in Eq. (D2), and this procedure is iterated until the value of **a** converges.

Only 24 properties were missing one or two values, and some of these were estimated by both of the above methods. The average root mean square difference between the values obtained by the two methods is ±0.069 in the standardized form. This implies that these iterative procedures converge properly. The errors in the computed values of the missing properties can be estimated as follows. First, we tested the procedure on a property for which no values were missing, and pretended that one or two of the 20 values were missing. These one or two values were then calculated by the two methods described above, using only the first four factors. As the example used in this test, we "estimated" the pretended missing values of the molecular weight (no. 1 of Table I), and the results are shown in Table VII. The error was calculated by the procedure of footnote *b* of Table VII: where two values were "missing," a separate error is given for each estimated value. Since method 1 gave a *slightly* smaller error than method 2, method 1 was used in all subsequent computations. After this initial test, we then generated many *arbitrary* properties as linear combinations of the ten factors in Table V, with the weighting factors randomly chosen but satisfying an imposed assumed value of the variance VAR4.[6]

[6] The variance of a standardized property **x** may be written as

$$(1/20)\mathbf{x}\mathbf{x}^t = (\mathbf{a}\mathbf{F} + \mathbf{u})(\mathbf{a}\mathbf{F} + \mathbf{u})^t$$

$$= \mathbf{a}\mathbf{a}^t + v = \sum_{i=1}^{NF} a_i^2 + v$$

where we used the relation $(1/20)\mathbf{F}\mathbf{F}^t = \mathbf{I}$ of Eq. (8), and v is a diagonal element of the matrix **V** of Eq. (8) for the property **x**. The above equation shows that the variance is expressed by the sum of the squared factor loadings and the residual part. Hence, the variance accounted for by the first four factors (VAR4) is defined by

$$\text{VAR4} = \sum_{i=1}^{4} a_i^2$$

A similar definition holds for VAR10.

Table VII. An Example of the Estimation of the Missing Value(s) of a Property[a]

Amino acid with missing value	Original value	Estimated value	
		Method 1	Method 2
One missing value			
ALA	89.1	91.1 (0.02)[b]	90.2 (0.01)
ARG	174.2	171.1 (0.02)	173.7 (0.00)
ASN	132.1	131.7 (0.00)	132.3 (0.00)
ASP	133.1	126.1 (0.05)	125.5 (0.06)
CYS	121.2	110.1 (0.09)	110.3 (0.09)
GLN	146.2	143.2 (0.02)	143.6 (0.02)
GLU	147.1	141.3 (0.04)	139.5 (0.05)
GLY	75.1	77.2 (0.03)	78.5 (0.05)
HIS	155.2	152.2 (0.02)	151.8 (0.02)
ILE	131.2	136.8 (0.04)	137.8 (0.05)
LEU	131.2	141.8 (0.08)	140.9 (0.07)
LYS	146.2	163.8 (0.12)	164.2 (0.12)
MET	149.2	144.8 (0.03)	143.6 (0.04)
PHE	165.2	168.4 (0.02)	166.6 (0.01)
PRO	115.1	128.6 (0.12)	127.2 (0.11)
SER	105.1	102.7 (0.02)	103.7 (0.01)
THR	119.1	114.9 (0.04)	116.7 (0.02)
TRP	204.2	196.6 (0.04)	192.4 (0.06)
TYR	181.2	178.1 (0.02)	178.1 (0.02)
VAL	117.2	119.4 (0.02)	121.8 (0.04)
Two missing values			
{ ALA[c]	89.1	94.0 (0.05)	93.3 (0.05)
{ LEU	131.2	142.8 (0.09)	141.8 (0.08)
{ ARG	174.2	178.8 (0.03)	181.0 (0.04)
{ LYS	146.2	165.3 (0.13)	166.3 (0.14)
{ ASN	132.1	132.0 (0.00)	132.7 (0.00)
{ MET	149.2	144.8 (0.03)	143.6 (0.04)
{ ASP	133.1	126.1 (0.05)	125.5 (0.06)
{ PHE	165.2	168.3 (0.02)	166.8 (0.01)
{ CYS	121.2	111.4 (0.08)	111.5 (0.08)
{ PRO	115.1	127.2 (0.11)	125.3 (0.09)
{ GLN	146.2	143.1 (0.02)	143.5 (0.02)
{ SER	105.1	102.5 (0.02)	103.5 (0.02)
{ GLU	147.1	141.5 (0.04)	139.7 (0.05)
{ THR	119.1	115.2 (0.03)	117.1 (0.02)
{ GLY	75.1	78.7 (0.05)	80.5 (0.07)
{ TRP	204.2	195.9 (0.04)	191.7 (0.06)
{ HIS	155.2	152.1 (0.02)	151.7 (0.02)
{ TYR	181.2	177.9 (0.02)	178.0 (0.02)
{ ILE	131.2	138.9 (0.06)	141.1 (0.08)
{ VAL	117.2	122.6 (0.05)	125.4 (0.07)

[a] The property used here for illustration is the molecular weight (No. 1 of Table I).
[b] The error given in parentheses is defined as (original value − estimated value)/original value.
[c] The braces encompass the pairs of amino acids whose values are missing.

Table VIII. Errors in the Estimation of Missing Values

Variance accounted for by the four factors (VAR4)	RMS deviation of the estimated value from the real value	
	Property with one missing value	Property with two missing values
1.000	0.00	0.00
0.975	0.22	0.24
0.950	0.31	0.34
0.925	0.38	0.41
0.900	0.44	0.48
0.875	0.49	0.54
0.850	0.54	0.59
0.825	0.58	0.63
0.800	0.62	0.68

RMS deviations for the many arbitrary generated properties are calculated on the standardized scale. The poorer the representation of these properties by four factors (i.e., the lower the value in column 1), the greater is the error in the missing value. When two values are missing, the error is greater than when one value is missing.

Table IX. Estimated Missing Values

	Name of property having missing values	Estimated missing values[a]				Error[b]
5	Residue volume (Krigbaum & Komoriya, 1979)	0.0	(Gly)	42.0	(Pro)	8.0
7	Radius of gyration of side chain (Levitt, 1976)	0.58	(Gly)			0.09
8	Volume (Bigelow, 1967)	76.0	(Asn)	90.0	(Gln)	5.0
9	Residue volume (Goldsack & Chalifoux, 1973)	125.0	(Asn)	149.0	(Gln)	9.0
11	Polarizability parameter (Charton & Charton, 1982)	0.13	(Pro)			0.02
39	Average reduced distance of side chain (Rackovsky & Scheraga, 1977)	1.06	(Gly)[c]			0.08
41	Average reduced distance of side chain (Meirovitch *et al.*, 1980)	1.01	(Gly)[c]			0.07
42	Average orientation angle of side chain (Meirovitch *et al.*, 1980)	80.0	(Gly)[c]			7.4
47	Average interactions per side-chain atom (Warme & Morgan, 1978)	8.0	(Gly)[c]			0.8
100	Zimm-Bragg parameter s at 20°C (Sueki *et al.*, 1984)	0.92	(Cys)	0.66	(Pro)	0.10
125	Hydration potential (Wolfenden *et al.*, 1981)	−3.7	(Pro)			2.9
139	Conformation parameter of α-helix (Finkelstein & Ptitsyn, 1976)	0.71	(Pro)			0.10

[a] The units differ for each physical property (e.g., cm^3 for the volume, Å for distance, etc). The amino acid for which the value of the physical property was missing is given in parentheses.

[b] The error is measured in the same units as used for each physical property, as in Table I.

[c] Even though Gly has no side chain, these values (arising from the backbone and the solvent) express the hydrophobicity of Gly to some extent.

This provided 20 values of each property. Again, we pretended that one or two of the 20 values were missing, and calculated them by method 1. This procedure was repeated many times, i.e., for each arbitrary generated property, and the errors (the differences between the real and the estimated values) were averaged (yielding only one average error, even when two values were missing). We found that the average error is a function of the variance that can be accounted for by the first four factors (VAR4 in Table I). The relation between VAR4 and the average errors calculated by method 1 is shown in Table VIII. Then the errors given in Table I for the physical properties having missing values (properties preceded by a plus sign) were determined by calculating VAR4 and interpolating in Table VII; the error in Table I was then converted from the standard scale to the units of each particular physical quantity. The estimated missing values for 12 physical properties are shown in Table IX with their estimated errors. The missing values for the other 12 physical properties could not be estimated because some of them give rise to statistical problems [i.e., they are not distributed normally (designated by STP in Table I)] and the others have too low a value of VAR4 (the variance accounted for by the first four factors) to obtain a reliable estimate; i.e., a cutoff value of VAR4 of 0.8 (see Table VIII) was used to obtain a reliable estimate of the missing value(s).

ACKNOWLEDGMENTS

This work was supported by research grants from the National Institute of General Medical Sciences, National Institutes of Health (GM-14312), from the National Science Foundation (DMB84-01811 and INT82-10589), and from the Japan Society for the Promotion of Science (BAMR107/INT82-10589). These studies were supported in part by the National Foundation for Cancer Research. We thank Prof. Robert E. Bechhofer for helpful discussions about factor analysis.

REFERENCES

Anderson, T. W., and Rubin, H. (1956). In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 5, pp. 111-150.

Bartlett, M. S. (1954). *J. Roy. Stat. Soc. B* **16**, 296-298.

Beghin, F., and Dirkx, J. (1975). *Arch. Int. Physiol. Biochim.* **83**, 167-168.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, Jr., E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535-542.

Bigelow, C. C. (1967). *J. theor. Biol.* **16**, 187-211.

Browne, C. A., Bennett, H. P. J., and Solomon, S. (1982). *Anal. Biochem.* **124**, 201-208.

Bull, H. B., and Breese, K. (1974). *Arch. Biochem. Biophys.* **161**, 665-670.

Burgess, A. W., Ponnuswamy, P. K., and Scheraga, H. A. (1974). *Isr. J. Chem.* **12**, 239-286.

Charton, M. (1981). *J. Theor. Biol.* **91**, 115-123.

Charton, M., and Charton, B. I. (1982). *J. Theor. Biol.* **99**, 629-644.

Chothia, C. (1976). *J. Mol. Biol.* **105**, 1-14.

Chou, P. Y., and Fasman, G. D. (1978). *Adv. Enzymol.* **47**, 45-148.

Cohn, E. J., and Edsall, J. T. (1943). *Protein, Amino Acids, and Peptides*, Reinhold, New York.

Crawford, J. L., Lipscomb, W. N., and Schellman, C. G. (1973). *Proc. Natl. Acad. Sci. USA* **70**, 538-542.

Dawson, D. M. (1972). In: *The Biochemical Genetics of Man* (Brock, D. J. H., and Mayo, O., eds.), Academic Press, New York, pp. 1-38.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978*a*). In: *Atlas of Protein Sequence and Structure*, Vol. 5, Suppl. 3 (Dayoff, M. O., ed.), National Biomedical Research Foundation, Washington, D. C., pp. 345-352.

Dayhoff, M. O., Hunt, L. T., and Hurst-Calderone, S. (1978*b*). In: *Atlas of Protein Sequence and Structure*, Vol. 5, Suppl. 3 (Dayhoff, M. O., ed.), National Biomedical Research Foundation, Washington, D.C., p. 363.

Fasman, G. D., ed. (1976). *Handbook of Biochemistry and Molecular Biology*, Vol. 1, *Proteins*, 3rd ed., CRC Press, Cleveland, Ohio.

Finkelstein, A. V., and Ptitsyn, O. B. (1976). *Biopolymers* **16**, 469-495.

Garel, J. P., Filliol, D., and Mandel, P. (1973). *J. Chromatogr.* **78**, 381-391.

Goldsack, D. E., and Chalifoux, R. C. (1973). *J. Theor. Biol.* **39**, 645-651.

Grantham, R. (1974). *Science* **185**, 862-864.

Harman, H. H. (1976). *Modern Factor Analysis*, 3rd ed., University of Chicago Press, Chicago, Illinois.

Hartigan, J. A. (1975). *Clustering Algorithms*, Wiley, New York.

Hopfinger, A. J. (1977). *Intermolecular Interactions and Biomolecular Organizations*, Wiley, New York.

Hopp, T. P., and Woods, K. R. (1981). *Proc. Natl. Acad. Sci. USA* **78**, 3824-3828.

Hutchens, J. O. (1970). In: *Handbook of Biochemistry*, 2nd ed. (Sober, H. A., ed.), Chemical Rubber Co., Cleveland, Ohio, pp. B60-B61.

IMSL (1982). *IMSL Library Reference Manual*, 9th ed., Institute for Mathematical and Statistical Subroutine Library, Houston, Texas.

Isogai, Y., Némethy, G., Rackovsky, S., Leach, S. J., and Scheraga, H. A. (1980). *Biopolymers* **19**, 1183-1210.

Janin, J. (1979). *Nature* **277**, 491-492.

Janin, J., Wodak, S., Levitt, M., and Maigret, B. (1978). *J. Mol. Biol.* **125**, 357-386.

Jones, D. D. (1975). *J. Theor. Biol.* **50**, 167-183.

Jöreskog, K. G. (1967). *Psychometrika* **32**, 443-482.

Jukes, T. H., Holmquist, R., and Moise, H. (1975). *Science* **189**, 50-51.

Jungck, J. R. (1978). *J. Mol. Evol.* **11**, 211-224.

Kaiser, H. F. (1958). *Psychometrika* **23**, 187-200.

Kanehisa, M. I., and Tsong, T. Y. (1980). *Biopolymers* **19**, 1617-1628.

Kendall, M., and Stuart, A. (1977). *The Advanced Theory of Statistics*, 4th ed., Macmillan, New York.

Kidera, A., Konishi, Y., Ooi, T., and Scheraga, H. A. (1985). *J. Protein Chem.*, submitted.

Krigbaum, W. R., and Komoriya, A. (1979). *Biochim. Biophys. Acta* **576**, 204-228.

Kyte, J., and Doolittle, R. F. (1982). *J. Mol. Biol.* **157**, 105-132.

Lawley, D. N., and Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method*, 2nd ed., Butterworths, London.

Levitt, M. (1976). *J. Mol. Biol.* **104**, 59-107.

Levitt, M. (1978). *Biochemistry* **17**, 4277-4285.

Lewis, P. N., Momany, F. A., and Scheraga, H. A. (1971). *Proc. Natl. Acad. Sci. USA* **68**, 2293-2297.

Lifson, S., and Sander, C. (1979). *Nature* **282**, 109-111.

Manavalan, P., and Ponnuswamy, P. K. (1978). *Nature* **275**, 673-674.

Maxfield, F. R., and Scheraga, H. A. (1976). *Biochemistry* **15**, 5138-5153.

Meek, J. L., and Rossetti, Z. L. (1981). *J. Chromatogr.* **211**, 15-28.

Meirovitch, H., Rackovsky, S., and Scheraga, H. A. (1980). *Macromolecules* **13**, 1398-1405.

Morrison, D. F. (1976). *Multivariate Statistical Methods*, McGraw-Hill, New York.

Nagano, K. (1973). *J. Mol. Biol.* **75**, 401-420.

Némethy, G., and Scheraga, H. A. (1977). *Q. Rev. Biophys.* **10**, 239-352.

Nishikawa, K., and Ooi, T. (1980). *Int. J. Peptide Protein Res.* **16**, 19-32.

Oobatake, M., and Ooi, T. (1977). *J. Theor. Biol.* **67**, 567-584.

Pliska, V., Schmidt, M., and Fauchére, J. L. (1981). *J. Chromatogr.* **216**, 79-92.

Ponnuswamy, P. K., Prabhakaran, M., and Manavalan, P. (1980). *Biochim. Biophys. Acta* **623**, 301-316.

Prabhakaran, M., and Ponnuswamy, P. K. (1982). *Macromolecules* **15**, 314-320.

Ptitsyn, O. B., and Finkelstein, A. V. (1983). *Biopolymers* **22**, 15-25.

Rackovsky, S., and Scheraga, H. A. (1977). *Proc. Natl. Acad. Sci. USA* **74**, 5248-5251.

Rackovsky, S., and Scheraga, H. A. (1982). *Macromolecules* **15**, 1340-1346.

Robson, B., and Osguthorpe, D. J. (1979). *J. Mol. Biol.* **132**, 19-51.
Robson, B., and Suzuki, E. (1976). *J. Mol. Biol.* **107**, 327-356.
Rose, G. D., and Roy, S. (1980). *Proc. Natl. Acad. Sci. USA* **77**, 4643-4647.
Simon, Z. (1976). *Quantum Biochemistry and Specific Interactions*, Abacus Press, Tunbridge Wells, Kent, England.
Sneath, P. H. A. (1966). *J. Theor. Biol.* **12**, 157-195.
Späth, H. (1980). *Cluster Analysis Algorithms for Data Reduction and Classification of Objects*, Halsted Press, New York.
Sueki, M., Lee, S., Powers, S. P., Denton, J. B., Konishi, Y., and Scheraga, H. A. (1984). *Macromolecules* **17**, 148-155.
Tanaka, S., and Scheraga, H. A. (1977). *Macromolecules* **10**, 9-20.
Vásquez, M., Némethy, G., and Scheraga, H. A. (1983). *Macromolecules* **16**, 1043-1049.
Von Heijne, G., and Blomberg, C. (1979). *Eur. J. Biochem.* **97**, 175-181.
Warme, P. K., and Morgan, R. S. (1978). *J. Mol. Biol.* **118**, 289-304.
Weber, A. L., and Lacey, Jr., J. C. (1978). *J. Mol. Evol.* **11**, 199-210.
Wertz, D. H., and Scheraga, H. A. (1978). *Macromolecules* **11**, 9-15.
Woese, C. R. (1973). *Naturwissenschaften* **60**, 447-459.
Wolfenden, R., Andersson, L., Cullis, P. M., and Southgate, C. C. B. (1981). *Biochemistry* **20**, 849-855.
Zimmerman, J. M., Eliezer, N., and Simha, R. (1968). *J. Theor. Biol.* **21**, 170-201.