PROTEINS: Structure, Function, and Genetics 10:229–239 (1991)

Secondary Structure-Based Profiles: Use of Structure-Conserving Scoring Tables in Searching Protein Sequence Databases for Structural Similarities

Roland Lüthy, Andrew D. McLachlan, and David Eisenberg
Molecular Biology Institute and Department of Chemistry and Biochemistry, University of California–Los Angeles, Los Angeles, California 90024-1570

ABSTRACT **The profile method, for detecting distantly related proteins by sequence comparison, has been extended to incorporate secondary structure information from known X-ray structures. The sequence of a known structure is aligned to sequences of other members of a given folding class. From the known structure, the secondary structure (α-helix, β-strand or "other") is assigned to each position of the aligned sequences. As in the standard profile method,[1] a position-dependent scoring table, termed a profile, is calculated from the aligned sequences. However, rather than using the standard Dayhoff mutation table in calculating the profile, we use distinct amino acid mutation tables for residues in α-helices, β-strands or other secondary structures to calculate the profile. In addition, we also distinguish between internal and external residues. With this new *secondary structure-based profile* method, we created a profile for eight-stranded, antiparallel β barrels of the insecticyanin folding class. It is based on the sequences of retinol-binding protein, insecticyanin and β-lactoglobulin. Scanning the sequence database with this profile, it was possible to detect the sequence of avidin. The structure of streptavidin is known, and it appears to be distantly related to the antiparallel β barrels. Also detected is the sequence of complement component C8, which we therefore predict to be a member of this folding class.**

Key words: profile method, sequence comparison, secondary structure-based profile, protein sequence data bases

INTRODUCTION

The profile method[1] provides a mathematical way to represent information about families of related sequences or common sequence motifs. With this representation it is then possible to compare the families or motifs with a new sequence to learn if the new sequence belongs to the family or contains the motif. This method has been successful for new sequences related by homology to the family or motif.[2] Our goal is to extend the method so that it is also effective in establishing relationships in three-dimensional structure.

In this paper, as a first step toward this goal, we extend the profile method to incorporate information on the secondary structure and solvent-environment of residues.

In the original profile method, the comparison of sequences makes use of two kinds of data: the sequence alignment of known family members and the information about the mutability from one amino acid to another coming from the Dayhoff mutation table.[3] Mutation tables are used in creating a profile in the following way: The profile is a position-dependent scoring table having as many rows as there are sequence positions in the protein family, and 22 columns. The first 20 columns give the likelihood of each of the 20 amino acids occupying that position, and the final two columns give the penalties for opening and extending a gap at that position. The likelihoods in columns 1–20 are calculated from the sequence alignment and the mutation table. If, for example, the profile is created from a single probe sequence, the row of residue likelihoods is simply the row of the mutation table for the amino acid at that position in the probe sequence. If, on the other hand, the profile is created from several probe sequences, each row is an average of the rows from the mutation table for the different amino acids found at that position of the probe sequences. The actual relationship is given by Eq. (10) below.

The Dayhoff table, however, does not take into account that amino acids in certain positions are more easily mutated than others. For example, exposed residues are more variable than interior ones

Received May 25, 1990; revision accepted December 13, 1990.

Address reprint requests to David Eisenberg, Molecular Biology Institute, University of California at Los Angeles, Los Angeles, CA 90024-1570.

Andrew McLachlan's permanent address is: Medical Research Council, Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England.

© 1991 WILEY-LISS, INC.

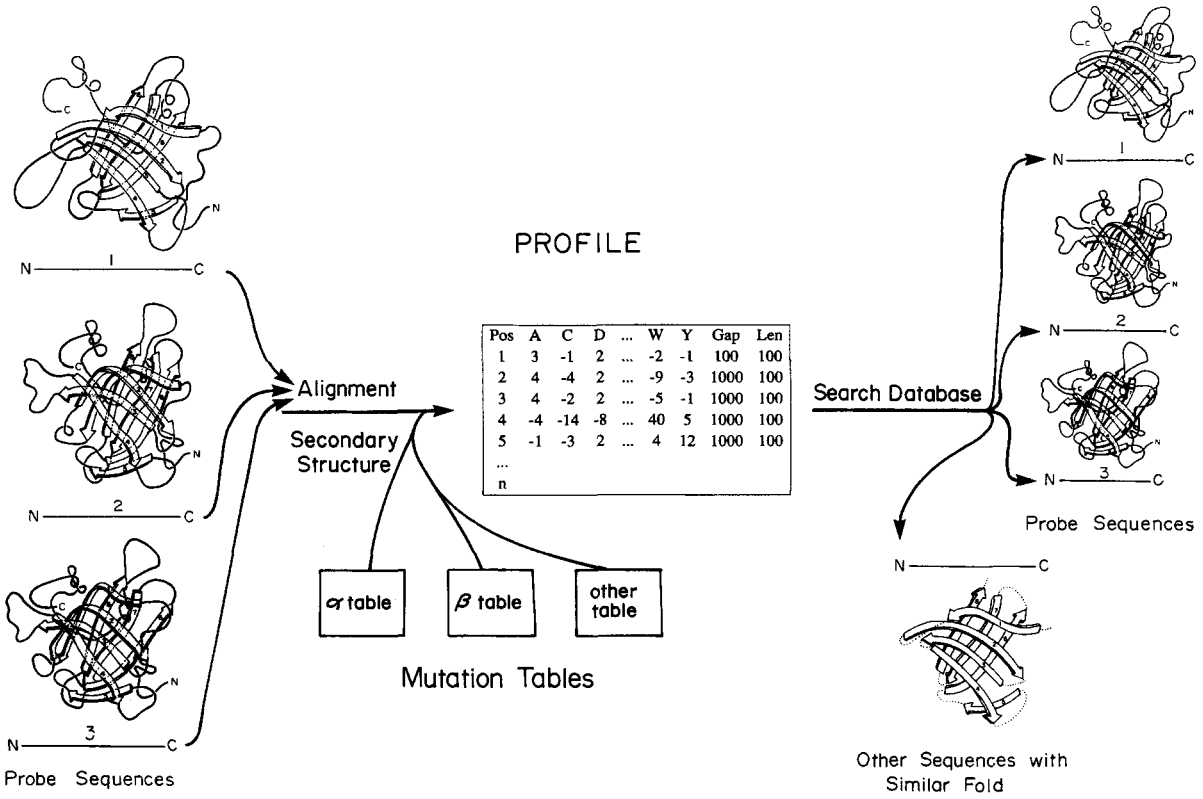

Fig. 1. Detection of a folding class, using secondary structure-based profiles. The profile is shown in the box at the center. It is a position-dependent scoring table describing sequences that fold as antiparallel β barrel proteins. It has 22 columns and *n* rows, where *n* is the number of residues of the fold. The first 20 columns for any given row give the likelihood of each of the 20 amino acids to occupy that position. The last two columns give the penalty for opening or extending a gap at that position. To create a secondary structure-based profile, one must align the probe sequences (shown at the left side under their corresponding structures) and one must know the three-dimensional structure of at least one of these, to assign the secondary structure at each position. The secondary structure at each position determines which mutation table is used to compute that row of the profile. The profile is then used to search the protein sequence database with a dynamic programming algorithm.[9] The highest scoring sequences are in general the probe sequences (shown to the right). Other high scoring sequences may have a closely related structure (such as that shown on the lower right). The protein drawings are adapted from Monaco et al.[21]

(see for example ref. 4). Also the frequencies of the amino acids vary in different structural positions. For example, hydrophobic residues tend to be in the interior of the protein and glycine and proline are rare in α-helices or β-strands. In the present approach of structure-based profiles, we include structural information in the profile by using different Dayhoff-like mutation tables for different structure types.

The local profile method is first tested with two well-known protein families: the globin family and the immunoglobulin family. In a third test of the local profile method, we create a profile for a particular folding type: the eight-stranded antiparallel β barrel, typified by insecticyanin, which has been found in many proteins.[5-8] This presents a test of whether profile analysis using the new tables can detect this three-dimensional fold. The strategy of this approach is depicted in Figure 1.

METHODS

Tables

Mutation tables were calculated from accepted point mutations using the procedure developed by Dayhoff.[3] The accepted point mutations were collected from multiple sequence alignments. The entire procedure is described in the following.

Multiple alignments

For each family of sequences (Table I), an initial profile was made[2] from a single, representative sequence (usually the reference sequence from the PDB database[10]). With this profile the protein sequence database (combined NBRF release 20 and Swissprot release 8) was searched. The sequences scoring more than 10 standard deviations above the mean value were aligned with the profile and a multiple alignment was generated with the program

TABLE I. Protein Families Used for Building the Structure-Based Comparison Tables*

Protein family	PDB entry[10]	Number of sequences	Profile length	Strictly conserved residues
Hemoglobin α chain	3HHB	154	141	16
Myoglobin	1MBO	65	153	47
Ig variable light λ	3FAB	33	102	16
Ig variable heavy κ	1FBJ	19	119	34
Tryptophan synthase α	1WSY	7	284	28
Tryptophan synthase β	1WSY	7	385	123
Ribonuclease	5RSA	39	124	29
Cytochrome *c*	1CCR	85	111	23
Adenylate kinase	3ADK	11	156	40
Alcohol dehydrogenase	4ADH	19	374	121
Plastocyanin	1PCY	19	99	29
Azurin	2AZA	9	129	47
Thermolysin	3TLN	5	316	120
Lysozyme	1LZ1	16	130	33
Lactate dehydrogenase	4LDH	20	329	43
Ferredoxin	1FDX	14	54	10

*Sequences were taken from the NBRF and Swissprot protein sequence databases.

PROFILEMULT. From this alignment a new profile was calculated and more sequences with lower scores were aligned to this profile. This was repeated until all homologous sequences were aligned. After every cycle, the alignment was checked by inspection and corrected manually if necessary.

Structural information

The secondary structure information of the reference sequence was extracted from the PDB entry with the program DSSP.[11] Interior and surface residues were determined by calculating the accessible surface area[12] of each residue in the protein and comparing it with the reference area of the same amino acid in the tripeptide Gly-X-Gly.[13] Residues having an accessible surface area of more than 10% of their reference area were considered exposed to the solvent (outside); the others were considered to be interior (inside). The 10% limit was chosen because it correctly yields the helical inside/outside pattern for the helices of hemoglobin.

Accepted point mutations

The alignments and the structural information were used to count the frequencies of occurrence F_i of each residue type and the number of changes C_i experienced by each residue type. From these data, matrices of accepted point mutations A_{ik} (replacement of residue i with residue k) were generated. The indices i and k represent the 20 amino acids. Matrices were generated for six structural classes:

(accessible surface area > 10%) and (α-helix) = outside α
(accessible surface area > 10%) and (β-strand) = outside β
(accessible surface area > 10%) and (other secondary structure) = outside
(accessible surface area < 10%) and (α-helix) = inside α
(accessible surface area < 10%) and (β-strand) = inside β
(accessible surface area < 10%) and (other secondary structure) = inside

Alternatively matrices were generated for three secondary structure classes: α-helices, β-strands or other. Notice that for the point mutations A_{ik} and changes C_i, multiple occurrences of the substitution of amino acid i to k were counted only once at each position in the aligned sequences. All amino acids, which could be aligned with the reference sequence, were used for the amino acid frequencies F_i (Fig. 2). Following Dayhoff,[3] we calculate the relative mutability m_i and the normalized frequencies f_i for all six tables (Fig. 3) from the relationships:

$$m_i = \frac{100\, F_{\text{Alanine}}\, C_i}{C_{\text{Alanine}} F_i} \tag{1}$$

$$f_i = \frac{F_i}{\sum_a F_a} \tag{2}$$

in which the sum is over the 20 residue types.

Mutation probability matrices

The mutation probability matrix M_{ik} for a given evolutionary distance is the matrix with the diagonal elements,

$$M_{ii} = 1 - \lambda m_i \tag{3}$$

and the nondiagonal elements,

$$M_{ik} = \frac{\lambda m_k A_{ik}}{\sum_a A_{ak}} \tag{4}$$

Amino	Outside		Outside alpha		Outside beta		Inside		Inside alpha		Inside beta	
acid	frequency	changes	frequency	changes	frequency	changes	frequency	changes	frequency	changes	frequency	changes
A	1869	1098	3288	862	204	89	326	172	1611	262	311	95
C	525	125	102	102	35	6	299	38	280	64	298	35
D	2081	813	1334	563	132	59	226	125	366	57	92	19
E	1219	754	2074	705	287	113	151	102	252	116	88	24
F	990	231	530	116	140	56	557	73	752	92	314	43
G	3841	784	1090	448	119	42	645	107	741	102	161	39
H	1433	438	900	323	108	29	284	63	402	39	43	27
I	648	313	600	269	205	93	440	147	894	218	626	137
K	2533	739	3097	695	405	123	293	113	348	95	48	21
L	1290	494	2151	392	298	88	977	166	2078	239	360	88
M	271	186	385	189	29	37	171	53	468	141	106	34
N	1413	900	1067	556	140	80	184	152	60	50	63	35
P	2026	493	737	210	56	27	405	47	35	37	94	15
Q	998	641	1062	578	166	75	88	97	113	81	78	27
R	1069	514	779	424	120	59	113	87	185	57	105	26
S	2541	1151	1371	654	453	141	531	143	472	135	216	66
T	2007	770	998	529	407	155	481	136	360	146	327	85
V	1265	538	1200	381	472	151	540	176	1553	256	749	136
W	181	64	84	20	42	7	99	16	279	16	71	10
Y	1000	321	605	109	224	61	143	76	210	44	232	58

Fig. 2. Frequencies of occurrence F_i and number of changes C_i of amino acids in the protein families of Table I listed by structural classes.

in which λ is a constant that determines the evolutionary distance. This constant λ can be determined as follows. For an evolutionary distance of 1 PAM, meaning that 1 out of 100 amino acids has changed, the following relationship holds:

$$\sum_i f_i M_{ii} = 0.99. \quad (5)$$

Then replacing M_{ii} by Eq. (3) gives

$$\sum_i f_i(1 - \lambda m_i) = 0.99 \quad (6)$$

with $\Sigma_i f_i = 1$ and solving for λ, we find

$$\lambda = \frac{0.01}{\sum_i f_i m_i}. \quad (7)$$

Again following Dayhoff,[3] the matrices for 1 PAM were then multiplied 250 times by themselves to get the matrices for an evolutionary distance of 250 PAM

$$M_{ik} = (M^{250})_{ik} \quad (8)$$

and then weighted by the normalized frequencies (relatedness odds):

$$R_{ik} = M_{ik}/f_i \ . \quad (9)$$

The final structure-based comparison tables (Fig. 4) are $\ln(R_{ik})$.

Profiles

The profiles were calculated as in Gribskov et al.[2]:

$$\text{Profile}(r,c) = \sum_{d=1}^{20} W_d(r)\text{Comp[s:r]}(\text{residue}_d,\text{residue}_c) \quad (10)$$

in which Comp[s:r] is the structure-based comparison table for the structure at position r, the sum is over the amino acid types, and W_d is a logarithmic weight (ref. 2, Eq. 3) which depends on the number of occurrences of each residue type d at position r. We used two methods to decide which table to use:

Method 1 was used when there was the amino acid sequence of a known three-dimensional structure in the alignment and the coordinates of the structure were available. From these coordinates the structural type of each position in the sequence was extracted, as explained above, and the corresponding table was selected for this position r. In Method 2 the alignment of sequences included an additional line with three possibilities in each position: α-helix, β-strand, or other (see Fig. 6). The corresponding table for position r was selected according to this line. The second method was used when the coordinates were not available or not complete, so that the solvent accessibility could not be calculated.

Gap penalties were also calculated as in Gribskov et al.,[2] but were multiplied by a factor of 10 if the position was in an α-helix or a β-strand. The rationale for this enhanced penalty is the greater likelihood of insertions and deletions into loops and turns than into elements of secondary structure. The value of 10 was arbitrarily chosen.

The profile was then aligned to every sequence in the database with PROFILE-SEARCH[2] and the scores of the alignments were normalized.[2] Then the mean and standard deviation of the normalized scores were calculated iteratively, omitting scores higher than 5 standard deviations above the mean value. Finally a Z-score was calculated:

$$Z\text{-score} = \frac{\text{normalized score} - \text{mean score}}{\text{standard deviation}}. \quad (11)$$

Comparison of Structure-Based Tables

The structure-based comparison tables (Fig. 4) were tested on two protein families: the globins and the immunoglobulin variable region, and then ap-

Amino acid	Outside		Outside alpha		Outside beta		Inside		Inside alpha		Inside beta	
	normalized frequency	relative mutability	normalized frequency	relative mutability	normalized frequency	relative mutability	normalized frequency	relative mutability	normalized frequency	relative mutability	normalized frequency	relative mutability
A	0.064	100.0	0.140	100.0	0.050	100.0	0.047	100.0	0.141	100.0	0.071	100.0
C	0.018	40.5	0.004	381.4	0.009	39.3	0.043	24.1	0.024	140.5	0.068	38.4
D	0.071	66.5	0.057	161.0	0.033	102.5	0.033	104.8	0.032	95.8	0.021	67.6
E	0.042	105.3	0.088	129.7	0.071	90.2	0.022	128.0	0.022	283.0	0.020	89.3
F	0.034	39.7	0.023	83.5	0.035	91.7	0.080	24.8	0.066	75.2	0.072	44.8
G	0.132	34.7	0.046	156.8	0.029	80.9	0.093	31.4	0.065	84.6	0.037	79.3
H	0.049	52.0	0.038	136.9	0.027	61.5	0.041	42.0	0.035	59.7	0.010	205.6
I	0.022	82.2	0.026	171.0	0.051	104.0	0.063	63.3	0.078	149.9	0.143	71.6
K	0.087	49.7	0.132	85.6	0.100	69.6	0.042	73.1	0.030	167.9	0.011	143.2
L	0.044	65.2	0.092	69.5	0.074	67.7	0.141	32.2	0.181	70.7	0.082	80.0
M	0.009	116.8	0.016	187.3	0.007	292.4	0.025	58.7	0.041	185.3	0.024	105.0
N	0.048	108.4	0.045	198.8	0.035	131.0	0.026	156.6	0.005	512.4	0.014	181.9
P	0.069	41.4	0.031	108.7	0.014	110.5	0.058	22.0	0.003	650.0	0.021	52.2
Q	0.034	109.3	0.045	207.6	0.041	103.6	0.013	208.9	0.010	440.8	0.018	113.3
R	0.037	81.8	0.033	207.6	0.030	112.7	0.016	145.9	0.016	189.5	0.024	81.1
S	0.087	77.1	0.058	182.0	0.112	71.3	0.076	51.0	0.041	175.9	0.049	100.0
T	0.069	65.3	0.043	202.2	0.101	87.3	0.069	53.6	0.031	249.4	0.075	85.1
V	0.043	72.4	0.051	121.1	0.117	73.3	0.078	61.8	0.136	101.4	0.171	59.4
W	0.006	60.2	0.004	90.8	0.010	38.2	0.014	30.6	0.024	35.3	0.016	46.1
Y	0.034	54.6	0.026	68.7	0.055	62.4	0.021	100.7	0.018	128.8	0.053	81.8

a

Amino acid	Other		Alpha helix		Beta strand	
	normalized frequency	relative mutability	normalized frequency	relative mutability	normalized frequency	relative mutability
A	0.061	100.0	0.140	100.0	0.061	100.0
C	0.023	34.2	0.011	189.4	0.040	34.5
D	0.064	70.3	0.049	159.0	0.027	97.5
E	0.038	108.0	0.067	153.8	0.045	102.3
F	0.043	34.0	0.037	70.7	0.054	61.0
G	0.124	34.3	0.052	130.9	0.033	81.0
H	0.047	50.4	0.037	121.2	0.018	103.8
I	0.030	73.1	0.043	142.1	0.099	77.5
K	0.078	52.1	0.099	99.9	0.054	89.0
L	0.063	50.3	0.121	65.0	0.078	74.9
M	0.012	93.5	0.024	168.6	0.016	147.2
N	0.044	113.9	0.032	234.4	0.024	158.6
P	0.067	38.4	0.022	139.5	0.018	78.4
Q	0.030	117.5	0.034	244.4	0.029	117.0
R	0.033	87.9	0.028	217.5	0.027	105.7
S	0.085	72.8	0.053	186.6	0.079	86.6
T	0.069	62.9	0.039	216.6	0.087	91.5
V	0.050	68.4	0.079	100.8	0.145	65.8
W	0.008	49.4	0.010	43.2	0.013	42.1
Y	0.032	60.0	0.023	81.8	0.054	73.0

b

Fig. 3. **(a)** Normalized frequencies f_i and relative mutability m_i of amino acids in six different structural classes. **(b)** Normalized frequencies and relative mutability of amino acids in three secondary structure classes.

plied to search for members of a third family, the eight-stranded antiparallel β barrels of the insecticyanin type. All tables were normalized to a sum of 0 and a square sum of 1, in order to compare the results between the tables. In total five different profiles were calculated, each using a different set of comparison tables:

1. Dayhoff table[3]
2. Six classes: outside α, outside β, outside, inside α, inside β, inside
3. Three classes: α, β, other
4. Two classes: inside, outside
5. Identity matrix, having unit diagonal and 0 elsewhere

For the globin family an alignment of seven globins from Lesk and Chothia[14] was used:

Hemoglobin α-chain	(Human [PDB entry 3 HHB])
Hemoglobin β-chain	(Rhesus macaque)
Myoglobin	(Human)
Globin, extracellular, small chain	(*Tylorrhynchus heterochaetus*)
Bacterial hemoglobin	(*Vitreoscilla* sp.)
Leghemoglobin II	(Yellow lupin)
Globin V	(Sea lamprey)

The immunoglobulin variable region alignment consisted of the following five sequences[15]:

Ig κ V	(mouse [PDB entry 1FBJ])
Ig λ-1 V	(mouse)
Ig heavy chain V-I	(human)
Ig heavy chain V-III	(human)
Ig κ V-I	(human)

a

Outside other

	W	F	Y	L	I	V	M	A	G	P	C	T	S	Q	N	E	D	H	K	R	
W	124	10	29	1	1	-1	9	-6	-12	-11	-3	-14	-7	-13	-6	-14	-10	-3	-19	3	W
		85	22	11	7	0	13	-13	-24	-24	-12	-13	-14	-15	-11	-18	-15	-6	-26	-14	F
W	129		64	2	-3	-2	4	-6	-16	-17	3	-9	-6	-9	-4	-10	-10	2	-16	-3	Y
F	11	67		45	19	13	24	1	-17	-10	-1	-4	-5	-2	-4	-4	-10	-6	-11	-4	L
Y	44	21	26		50	24	20	1	-16	-11	-13	0	-5	-3	-3	-3	-7	-10	-10	-4	I
L	-18	2	-2	37		39	14	3	-12	-8	-7	-1	-4	-2	-3	-2	-7	-9	-9	-3	V
I	-19	-4	0	9	31		40	1	-11	-11	-3	-2	-4	0	-1	-4	-7	-3	-8	0	M
V	-8	-9	-2	8	17	26		11	-5	1	-5	3	4	2	2	3	1	-4	-1	0	A
M	-20	2	1	13	18	12	61		39	-19	-15	-12	-6	-8	-4	-6	-7	-13	-15	-7	G
A	-20	-24	-4	-10	-1	0	-6	12		55	-28	-6	-3	-6	-8	-3	-9	-12	-9	-9	P
G	-23	-52	-16	-32	-20	-20	-32	5	52		109	-7	-10	-10	-7	-12	-13	-23	-16	-6	C
P	-47	-28	-20	-36	-28	-20	-25	-5	-11	83		30	4	1	2	1	-2	-5	-3	-1	T
C	0	-36	-7	-35	-17	-12	-17	1	-36	-30	93		17	2	3	2	2	-4	-1	1	S
T	-29	-37	-8	-18	-1	-2	-9	8	0	-14	1	35		16	3	5	2	2	5	6	Q
S	-29	-37	-6	-25	-11	-10	-18	9	7	-9	-11	10	36		11	2	4	2	-1	2	N
Q	-2	-20	0	-12	-4	-3	-5	6	3	-7	-5	8	8	7		15	6	-3	3	2	E
N	-6	-19	0	-10	-3	-3	-8	5	5	-11	-11	6	8	6	6		29	-5	-2	-2	D
E	-18	-30	-3	-14	-5	-4	-12	7	8	-7	-16	10	9	9	8	15		54	-8	-1	H
D	-16	-25	0	-16	-6	-4	-13	6	6	-7	-14	8	8	8	8	12	17		39	6	K
H	-21	-10	5	-18	-10	-6	-5	-2	-14	-23	-20	-5	-1	4	6	-3	5	65		32	R
K	-21	-30	-1	-15	-5	-4	-11	6	2	-5	-22	6	8	10	10	13	13	6	31		
R	2	-11	3	-10	-4	-3	-6	5	-2	-12	-10	6	6	6	5	8	8	5	10	9	R
	W	F	Y	L	I	V	M	A	G	P	C	T	S	Q	N	E	D	H	K	R	

Inside other

b

Outside Alpha

	W	F	Y	L	I	V	M	A	G	P	C	T	S	Q	N	E	D	H	K	R	
W	170	36	47	-3	-14	-9	-14	-20	-21	-29	16	-2	-19	-3	-18	-16	-12	7	-31	-8	W
		99	40	12	10	-11	13	-20	-21	-32	4	-16	-18	-13	-18	-24	-18	-9	-30	-18	F
W	130		105	-13	-14	-18	-19	-21	-14	-17	-7	-15	-16	-12	-10	-21	-14	-1	-25	-10	Y
F	4	58		54	13	-1	13	-18	-15	-24	-3	-6	-12	-13	-12	-16	-18	-12	-21	-7	L
Y	31	25	68		41	18	17	-5	-5	-8	4	3	-4	-4	-6	-10	-9	-4	-11	-3	I
L	-30	-4	-5	25		44	7	-5	-5	-6	5	3	-2	-5	-6	-6	-4	-8	-11	-3	V
I	-27	-4	-4	0	17		44	-3	-5	-10	4	2	-3	0	-5	-5	-7	-5	-7	0	M
V	-23	-9	-14	-6	5	21		24	1	-2	2	1	1	-1	-1	-2	-2	-5	-10	-2	A
M	-12	1	-2	2	3	3	17		29	0	4	3	5	2	4	1	2	0	-4	2	G
A	-50	-18	-15	-14	-2	-1	-2	22		67	-1	-3	1	1	-2	-1	1	-7	-8	-2	P
G	-57	-25	-17	-21	-6	-10	-6	5	53		9	4	5	2	1	-3	-1	0	-4	2	C
P	-33	-8	-3	-7	1	0	-1	5	6	4		16	4	2	3	0	1	1	-3	2	T
C	-45	-12	-8	-11	3	1	0	8	3	2	50		17	2	4	1	3	0	-3	2	S
T	-39	-13	-9	-6	2	2	1	6	4	4	8	10		13	2	3	3	2	-1	2	Q
S	-45	-14	-10	-10	1	-1	0	9	7	9	10	9	23		16	1	4	3	-2	2	N
Q	-15	-14	0	-7	0	-2	-1	2	5	2	0	3	4	6		24	6	-1	-4	-1	E
N	-34	-15	-4	-6	0	-1	0	4	7	2	4	3	4	4	4		23	1	-4	0	D
E	-40	-21	-8	-10	-2	-3	-3	4	8	4	-2	4	5	7	5	12		43	-6	0	H
D	-55	-26	-8	-20	-13	-15	-11	-1	0	5	-13	1	3	10	11	20	71		32	5	K
H	-41	-7	23	-10	-17	-27	-8	-30	-32	-14	-30	-16	-19	13	-3	-3	7	95		18	R
K	-23	-23	-13	-8	-1	-4	-3	1	2	3	-2	5	5	9	6	14	9	-9	33		
R	-35	-14	-3	-3	-1	-3	2	-2	0	2	-1	3	4	9	5	12	5	4	22	38	R
	W	F	Y	L	I	V	M	A	G	P	C	T	S	Q	N	E	D	H	K	R	

Inside Alpha

4c

Outside Beta

	W	F	Y	L	I	V	M	A	G	P	C	T	S	Q	N	E	D	H	K	R	
W	150	32	11	-29	-25	-15	-21	-2	-33	-25	-34	-31	-13	-34	-30	-35	-40	-36	-14	-31	W
		47	24	6	7	-3	2	-4	-10	2	12	-8	-9	-9	-8	-12	-17	-3	-11	-9	F
W	115		56	-1	-10	-13	-4	-10	-14	3	-25	-16	-7	-14	-2	-13	-13	18	-14	-13	Y
F	3	58		42	16	8	12	-7	-21	3	-34	-12	-15	-13	-11	-13	-21	-9	-14	-8	L
Y	15	16	37		29	11	10	-2	-8	8	-26	-3	-10	-7	-8	-7	-10	-13	-9	-2	I
L	-8	-4	-12	25		24	7	-2	-14	1	-11	-1	-7	-8	-9	-8	-13	-19	-8	-4	V
I	-8	-10	-16	11	20		7	2	-2	6	-17	2	-2	-5	-3	-5	-7	-11	-5	2	M
V	-15	-13	-18	9	11	21		27	9	6	-22	3	5	-3	-2	-3	-4	-11	-6	-3	A
M	-15	-2	2	2	3	-5	35		59	-10	-33	-1	0	3	-4	4	3	-4	-1	6	G
A	-23	-12	0	-6	-8	-7	1	18		56	-22	-2	-4	1	5	-10	-2	-9	-12	-1	P
G	-32	-17	-9	-9	-9	-9	4	14	48		156	-7	6	-29	6	-30	-26	-33	-32	-28	C
P	-32	-9	-7	-3	-9	-9	-16	4	-14	95		20	4	2	-4	0	-1	-19	-1	2	T
C	-43	-40	-9	-29	-29	-28	-23	0	-10	-8	66		25	-6	0	-3	0	-11	-4	-4	S
T	-7	-17	1	-11	-7	-7	3	6	8	-1	-11	24		30	1	6	3	4	4	2	Q
S	-21	-16	3	-15	-12	-14	-2	12	16	-4	-1	12	28		18	0	9	14	-1	0	N
Q	-14	-16	9	-9	-14	-15	4	6	4	-17	7	4	10	42		25	10	-5	2	-1	E
N	-11	-6	13	-15	-14	-17	3	6	6	-13	5	8	13	18	18		38	12	-2	0	D
E	-19	-7	14	-21	-18	-24	5	8	13	-19	-1	9	13	15	22	59		84	-19	-7	H
D	-24	-24	7	-27	-22	-30	5	4	-4	-26	0	2	17	18	18	29	80		30	9	K
H	18	6	20	-15	-15	-19	3	2	-3	-15	-2	5	10	13	13	16	24	18		32	R
K	-11	-14	11	-11	-13	-12	6	4	-2	-16	2	7	11	22	16	10	30	14	35		
R	16	-10	7	-12	-14	-11	3	1	-8	-20	-7	10	3	24	10	-1	-1	11	22	60	R
	W	F	Y	L	I	V	M	A	G	P	C	T	S	Q	N	E	D	H	K	R	

Inside Beta

Fig. 4a–c. Legend appears on page 235.

d

Other

	W	F	Y	L	I	V	M	A	G	P	C	T	S	Q	N	E	D	H	K	R	
W	129	10	35	-5	-5	-3	4	-9	-15	-16	-1	-18	-11	-13	-7	-17	-13	-7	-21	2	W
F	10	83	22	10	5	-2	11	-16	-30	-26	-19	-19	-19	-18	-14	-22	-20	-9	-29	-15	F
Y	35	22	59	0	-2	-3	3	-5	-16	-16	0	-9	-5	-6	-2	-9	-7	3	-12	-1	Y
L	-5	10	0	48	19	14	24	-3	-22	-16	-10	-9	-10	-6	-7	-8	-14	-10	-15	-7	L
I	-5	5	-2	19	49	24	22	-1	-18	-15	-13	-1	-8	-4	-4	-5	-9	-12	-11	-6	I
V	-3	-2	-3	14	24	37	15	2	-14	-11	-8	-2	-6	-3	-4	-3	-7	-9	-9	-4	V
M	4	11	3	24	22	15	52	-1	-15	-15	-7	-4	-7	-2	-3	-7	-9	-4	-10	-2	M
A	-9	-16	-5	-3	-1	2	-1	11	-4	0	-5	3	4	3	2	3	2	-3	1	1	A
G	-15	-30	-16	-22	-18	-14	-15	-4	41	-17	-21	-10	-4	-6	-2	-4	-5	-12	-12	-6	G
P	-16	-26	-16	-16	-15	-11	-15	0	-17	58	-30	-7	-3	-6	-7	-3	-8	-12	-8	-8	P
C	-1	-19	0	-10	-13	-8	-7	-5	-21	-30	108	-6	-11	-11	-9	-15	-15	-23	-19	-9	C
T	-18	-19	-9	-9	-1	-2	-4	3	-10	-7	-6	31	5	2	3	2	-1	-5	-2	0	T
S	-11	-19	-5	-10	-8	-6	-7	4	-4	-3	-11	5	19	3	4	3	3	-3	1	2	S
Q	-13	-18	-6	-6	-4	-3	-2	3	-6	-6	-11	2	3	14	4	6	4	2	6	7	Q
N	-7	-14	-2	-7	-4	-4	-3	2	-2	-7	-9	3	4	4	10	3	5	2	2	3	N
E	-17	-22	-9	-8	-5	-3	-7	3	-4	-3	-15	2	3	6	3	15	8	-2	5	3	E
D	-13	-20	-7	-14	-9	-7	-9	2	-5	-8	-15	-1	3	4	5	8	27	-3	1	1	D
H	-7	-9	3	-10	-12	-9	-4	-3	-12	-12	-23	-5	-3	2	2	-2	-3	55	-6	0	H
K	-21	-29	-12	-15	-11	-9	-10	1	-12	-8	-19	-2	1	6	2	5	1	-6	38	7	K
R	2	-15	-1	-7	-6	-4	-2	1	-6	-8	-9	0	2	7	3	3	1	0	7	28	R
	W	F	Y	L	I	V	M	A	G	P	C	T	S	Q	N	E	D	H	K	R	

4 e

Alpha helix

	W	F	Y	L	I	V	M	A	G	P	C	T	S	Q	N	E	D	H	K	R	
W	162	19	36	-22	-25	-18	-11	-43	-46	-46	-13	-25	-41	-19	-38	-38	-35	-17	-43	-29	W
F		88	33	4	3	-9	8	-23	-26	-28	-4	-19	-21	-19	-22	-28	-24	-13	-34	-22	F
Y	133		97	-12	-11	-19	-12	-19	-14	-10	-8	-13	-14	-8	-7	-16	-12	4	-20	-8	Y
L	16	57		45	7	-4	8	-19	-19	-21	-8	-9	-15	-14	-14	-18	-21	-14	-21	-10	L
I	13	20	47		38	13	12	-6	-8	-7	5	1	-5	-6	-7	-10	-12	-9	-12	-6	I
V	-15	1	-7	34		40	6	-6	-9	-7	3	1	-4	-7	-8	-8	-9	-14	-13	-6	V
M	-12	-3	-15	13	28		40	-4	-6	-8	3	0	-4	-2	-5	-7	-9	-7	-8	-2	M
A	-17	-9	-15	8	11	23		23	2	1	4	2	3	1	1	0	0	-6	-6	-1	A
G	-20	-1	-3	8	8	2	25		37	3	4	4	6	3	6	3	4	-1	-1	3	G
P	-17	-9	-5	-7	-7	-5	1	23		59	-2	1	6	5	3	6	5	-4	-1	2	P
G	-33	-15	-13	-15	-11	-12	2	12	54		51	6	7	1	1	-5	-4	-5	-6	1	C
P	-28	-4	-1	-1	-1	-4	0	5	-13	78		14	5	3	4	2	2	0	0	2	T
C	-53	-31	-19	-34	-30	-27	-28	-5	-18	-12	97		17	4	5	4	5	0	1	4	S
T	-20	-13	-8	-12	-7	-4	2	4	3	-3	-18	21		11	5	7	6	4	4	5	Q
S	-19	-14	-3	-15	-13	-10	-3	7	8	-4	-10	8	26		13	5	7	5	3	5	N
Q	-26	-12	-4	-12	-13	-11	-3	0	4	-2	-11	4	2	31		23	10	1	3	4	E
N	-22	-9	4	-13	-12	-12	-2	1	0	2	-5	2	6	7	15		28	3	2	3	D
E	-29	-13	-3	-14	-13	-12	-3	0	6	-13	-21	4	5	11	8	31		52	-2	2	H
D	-32	-21	-5	-22	-17	-18	-5	-1	1	-7	-16	2	8	9	13	17	49		34	9	K
H	7	3	21	-13	-15	-18	-3	-4	-5	-10	-12	-4	3	8	13	5	17	55		19	R
K	-15	-14	-5	-13	-14	-10	-4	-3	0	-14	-23	4	4	12	7	10	8	-1	34		
R	-5	-11	-5	-10	-9	-7	2	-2	2	-6	-19	6	1	9	4	2	2	2	14	40	R
	W	F	Y	L	I	V	M	A	G	P	C	T	S	Q	N	E	D	H	K	R	

Beta strand

Fig. 4. Structure-based comparison tables for six structural classes (**a–c**) and three secondary structure classes (**d,e**). Similar amino acids are grouped together. Values were computed from data in Figure 3 and the matrix of accepted point mutations as described in Methods [Eqs. (1) to (9)] and multiplied by 100.

RESULTS AND DISCUSSION

Tables

Six structure-based comparison tables were calculated using three secondary structure classes (α-helix, β-strand and other) combined with two solvent accessibility classes (inside and outside). Figure 2 shows frequencies of occurrence F_i and number of changes C_i for all amino acids in these six structural types. These values were accumulated from the multiple alignments of the protein families in Table I. In Figure 3a the normalized frequencies and relative mutabilities for the six classes are given.

For comparison, three secondary structure-based tables were calculated, using data from only the secondary structure classes, and similarly, the data from only two solvent accessibility classes yielded a set of two tables. Figure 3b shows normalized frequencies and relative mutabilities for the three secondary structure classes.

The data from these tables and the accepted point mutations are recast into the structure-based comparison tables (Fig. 4) as described in Methods, Eqs. (1) to (9). One set (Fig. 4a–c) treats residues as belonging to six structural classes; the other (Fig. 4d and e) treats residues as belonging to three secondary structure classes. A major difference among the different tables is the mutability of cysteine: It seems much more conserved in β-strands than in α-helices. All changes of cysteine in the beta table have a negative value (see Fig. 4e) whereas the cysteine value on the diagonal is large and positive. The Dayhoff table[3] shows the same tendencies as the beta table, whereas the alpha table seems less stringent.

Amino	Outside			Inside			Overall		
acid	Other	α	β	Other	α	β	Inside	α	β
A	25	43	3	4	21	4	30	64	7
C	34	7	2	19	18	19	57	25	22
D	49	32	3	5	9	2	16	40	5
E	30	51	7	4	6	2	12	57	9
F	30	16	4	17	23	10	49	39	14
G	58	17	2	10	11	2	23	28	4
H	45	28	3	9	13	1	23	41	5
I	19	18	6	13	26	18	57	44	24
K	38	46	6	4	5	1	10	51	7
L	18	30	4	14	29	5	48	59	9
M	19	27	2	12	33	7	52	60	9
N	48	36	5	6	2	2	10	39	7
P	60	22	2	12	1	3	16	23	4
Q	40	42	7	4	5	3	11	47	10
R	45	33	5	5	8	4	17	41	9
S	46	25	8	10	8	4	22	33	12
T	44	22	9	11	8	7	26	30	16
V	22	21	8	9	27	13	49	48	21
W	24	11	6	13	37	9	59	48	15
Y	41	25	9	6	9	10	24	34	19

Fig. 5. Distribution of amino acids in the protein families of Table I among structural classes. On left: the six structural classes; on right: sum of all inside residues, α residues and β residues. Data are in percent of each amino acid. "Other" means neither α-helix nor β-strand according to the program DSSP.[11]

The distribution of amino acids between the different classes is summarized in Figure 5. They are consistent with previously published studies of secondary structure and internal/external preferences of amino acid side chains.[16–19] Consider as an example the distributions of glycine and proline: approximately 60% of all glycines and prolines are in solvent accessible, non- α-helix and non-β-strand regions. Also the relative mutability of these two residues, in the outside "other" class, where they mainly occur, is very low (34 for Gly, 41 for Pro). Other examples are the hydrophobic amino acids (Cys, Phe, Ile, Leu, Met, Val, Trp) which appear to be approximately 50 to 60% inside.

Comparing Results From Different Tables

The relative effectiveness of the structure-based comparison tables, of the Dayhoff table[3] and of the identity table for detecting related proteins by the profile method was tested. All these comparison tables are almost equally effective in recognizing closely related protein sequences by the profile method (Tables II and III). Even the table that uses only amino acid identities is effective, and in fact it seems to be the best for the immunoglobulin family. However the secondary structure-based comparison tables (Fig. 4d and e) are more effective than the others in detecting the more distantly related sequences of the same folding class, as is shown in the next section.

"Insecticyanin Class" Antiparallel β Profile

It is desirable to test the secondary structure-based profile method on a set of sequences less closely related to each other than the homologous families of Tables II and III. To do this we chose a family of proteins that bind small hydrophobic ligands. These proteins belong to the antiparallel β folding class.[20] Can a profile based on aligned sequences of members from this class detect other proteins with this fold? Does the use of the structure-based comparison tables improve the sensitivity of this profile? We created a profile for this class based on an alignment of three proteins that have this fold: retinol-binding protein,[5] β-lactoglobulin,[21] and insecticyanin.[7] These proteins have quite different sequences but essentially the same three-dimensional structure. The profile was prepared from the sequence alignment and the assignment of the β-strands of Banaszak,[22] shown in Figure 6. Using this alignment three profiles were calculated as described in method 2 above: one using the Dayhoff table, one with the identity table, and one using the three-class, secondary structure-based comparison table. For the latter profile only the β and the other tables were effectively used, since there are no α-helices in these structures. For the two profiles using the six class table and the inside/outside table, the profiles were calculated with method 1 using the three-dimensional structure of insecticyanin[7] as the reference structure. The protein database was searched with all five profiles. The results are summarized in Table IV. All three sets of structure based tables find more members of the antiparallel-β family than the Dayhoff table and the identity table. The three structure based tables also find fewer false positives.

The proteins that were used to build the profile receive lower scores when the α/β/other table was used, than with the Dayhoff table[3] or the identity table. In contrast, the sequences that are more distantly related to those included in the profile, but which show significant sequence homology and therefore are thought to have the same structure, generally receive higher Z-scores with the α/β/other table. This suggests that the α/β/other table is more sensitive in detecting distantly related sequences. The proteins that are known to be related to those in the profile belong to a family of small hydrophobic molecule binding proteins. The established members of this family are[23–25] retinol-binding protein, β-lactoglobulin, insecticyanin, retinal retinol-binding protein, placental protein, apolipoprotein D, minor major urinary protein, α_1-microglobulin, olfactory protein, α_2-microglobulin, major urinary protein, androgen-dependent epididymal protein, odorant-binding protein, and aphrodisin. Complement component C8 γ chain, which scores highly with both tables, but higher with the α/β/other ta-

TABLE II. Effectiveness of Different Comparison Tables in Detecting Globin Sequences by the Profile Method*

	ABoIO	ABo	IO	DAY	IDENT	Total
Hemoglobin α	169	169	169	169	169	169
Hemoglobin β	154	154	154	154	154	154
Myoglobin	69	69	69	69	69	69
Total hemoglobin	380	380	380	380	379	381
Other globins	23	27	23	29	28	31
Leghemoglobin	12	12	12	12	12	12
False positives	1	0	5	1	2	

*The entries of a column give the number of homologous sequences scoring more than 5 standard deviations above the mean. The comparison tables are ABoIO: six class table (outside α, outside β, outside, inside α, inside β, inside); ABo: three class table α, β, other; IO: two classes inside, outside; DAY: Dayhoff table[3]; IDENT: identity matrix.

TABLE III. Effectiveness of Different Comparison Tables in Detecting Immunoglobulin Variable Region Sequences by the Profile Method*

	ABoIO	ABo	IO	DAY	IDENT	Total
Ig κ chain V	168	168	168	168	168	169
Ig λ chain V	39	39	39	38	39	39
Ig heavy chain V	108	118	153	117	151	164
T-cell receptor α	11	16	21	25	36	47
T-cell receptor β	7	10	12	10	17	38
T-cell receptor γ	1	5	6	5	12	22
False positives	13	13	3	3	11	

*Entries and tables as in Table II.

```
        1                                                  50
Retin   SGTWYAMAKKDPEGLFLQD.NIVAEFSVDETGQMSATAKGRVRLLNNWDV
Insec   AGAWHEIAKLPLENENQGK.CTIAEYKYD...GKKASVYNSFVSN.....
Lacto   AGTWYSLAMAASDISLLDAQSAPLRVYVEELKPTPEGDLEILLQKWENGE
STRUC   oBBBBBBBBBoooooooooooBBBBBBBBBoooBBBBBBBBBBoooooooo

        51                                                 100
Retin   CADMVGTFTDTED.......PAKFKMKYWGVASFLQKGNDDHWIVDTDYD
Insec   GVKEYMEGDLEIAPDAKYTKQGKYVMTFKFGQRV...VNLVPWVLATDYK
Lacto   CAQKKIIAEKTKI.......PAVFKIDAL........NENKVLVLDTDYK
STRUC   BBBBBBBBBBoooooooooooBBBBBBoooooooooooBBBBBBBBBBoo

        101                     128
Retin   TYAVQYSCRLLNLDGTCADSYSFVFSR
Insec   NYAINYNCDYHPDKKA.HSIHAWILSK
Lacto   KYLLFCMENSAEPE...QSLACQCLVR
STRUC   BBBBBBBBBBooooooooBBBBBBBBBB
```

Fig. 6. Alignment of three antiparallel β barrel protein sequences. This alignment, based on Banaszak,[22] was used to calculate the antiparallel β barrel profiles. Protein names are the codes from the NBRF protein database: Retin, retinol binding protein (NBRF code: VAHU); Insec, insecticyanin (NBRF code: CUWOI); Lacto, β-lactoglobulin (NBRF code: LGBO). STRUC is the secondary structure assignment, in this case either B for β-strand or o for other.

ble, has also been suggested to belong to this group of proteins.[26] Our method tends to confirm this suggestion.

The known sequences not identified by the profile using the Dayhoff table were one member of the α_1-microglobulin family, three members of the α_2-microglobulins, aphrodisin, odorant-binding protein, androgen-dependent epididymal protein, and avidin. Aphrodisin, avidin and one of the α_2-microglobulins scored above 3.5 only when the α/β/other table was used.

Some of the high scoring sequences with the α/β/other table score very low with the Dayhoff table. Some of those are known members of the ligand-binding family: for example the odorant-binding protein from rat, α_2-microglobulin from mouse, and androgen-dependent epididymal protein. Other high scoring sequences are potential β barrel proteins. One of these sequences is avidin. The structure of streptavidin, a homologous protein, has been recently determined[27,28] and it was shown that it consists of an eight-stranded, antiparallel β barrel. Another good candidate is the macrophage inflammatory protein 1. Figure 7 shows the alignments for avidin and macrophage inflammatory protein 1 with the "insecticyanin class" antiparallel β barrel profile in order to predict the β strands.

CONCLUSION

When used with the profile method, a variety of amino acid comparison tables are successful in detecting the members from homologous protein families. However in detecting members from a broader folding class, the "insecticyanin class" antiparallel β barrels, secondary structure-based comparison tables proved superior to a single Dayhoff table. The families tested comprised of the globin and immunoglobulin families and a family of proteins which bind small hydrophobic molecules. In the latter case the α/β/other table was more sensitive in detecting distantly related sequences than the usual Dayhoff

TABLE IV. Effectiveness of the "Insecticyanin Class" Antiparallel β Profile for Identifying Sequences of This Class*

	ABoIO	ABo	IO	DAY	IDENT
Known antiparallel β barrels (correct positives)	37	40	35	32	30
Known other structure (false positives)	7	10	4	13	7
Unknown structure	23	25	19	22	25

*The entries of a column give the number of sequences scoring more than 3.5 standard deviations above mean. The columns refer to the tables defined in the caption of Table II.

```
Avidin  8 GKWTNDLGSNMTIGAVNSRGEFTGTYTTAVTATSNEIKESPLHGTENTIN
   MIP  4 ..................................STTALAVLLCTMTLC
 STRUC    BBBBBBBBBoooooooooBBBBBBBBBooBBBBBBBBBBoooooooB

Avidin 58 KRTQPTFGFTVNWKF......SESTTVFTGQCF.IDRN.....GKEVLKT
   MIP 19 NQVFSAPYGADTPTACCFSYSWKIPRQFIVEYFETSSLCSQPGVIFLTKR
 STRUC    BBBBBBBBBooooooooooooooooBBBBBBooooooooooooooooBBBB

Avidin 96 MWLLRSSVNDIGDDWKATRV
   MIP 69 NWQICADSKETWVQEYITDLEL
 STRUC    BBBBBBBooBBBBBBBBBBooooooooBBBBBBBBBB
```

Fig. 7. Alignment of avidin and macrophage inflammatory protein with the β strands of the antiparallel β barrel profile. The sequences of avidin and macrophage inflammatory protein (MIP) were aligned with the profile for antiparallel β barrels. STRUC is the secondary structure assignment from Figure 6, aligned to the sequences in order to predict the β strands.

table. This profile also detected avidin, a sequence which has no significant amino acid sequence similarity, but a similar three-dimensional fold as the proteins used to build the profile. In general we expect the secondary structure-based profiles to be more sensitive in detecting distantly related proteins, since secondary structures are more strongly conserved than the primary sequences of proteins.

ACKNOWLEDGMENTS

We thank Dr. Jim Conway for his help in building the mutation tables, Prof. L. Banaszak for an alignment of "insecticyanin class" antiparallel β barrel proteins, Prof. H. Holden for the coordinates for insecticyanin, and C. Raynar for preparation of Figure 1.

This work was supported by NIH Grants GM-31299 and GM-39558.

REFERENCES

1. Gribskov, M., McLachlan, A.D., Eisenberg, D. Profile analysis detection of distantly related proteins. Proc. Natl. Acad. Sci. U.S.A. 84:4355, 1987.
2. Gribskov, M., Lüthy, R., Eisenberg, D. Profile analysis. Methods Enzymol. 183:146–159, 1990.
3. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C. A model of evolutionary change in proteins. In: "Atlas of Protein Sequence and Structure." Dayhoff, M.O. (ed.), Vol. 5, Suppl. 3. Washington, D.C.: National Biomedical Research Foundation, 1978, 345.
4. Rees, D., DeAntonio, L., Eisenberg, D. Hydrophobic organization of membrane proteins. Science 245(4917):510–513, 1989.
5. Newcomer, M.E., Jones, T.A., Aqvist, J., Sundelin, J., Eriksson, U., Rask, L., Peterson, P.A. The three-dimensional structure of retinol binding protein. EMBO J. 3(7): 1451–1454, 1984.
6. Huber, R., Schneider, M., Mayr, I., Muller, R., Deutzmann, R., Suter, F., Zuber, H., Falk, H., Kayser, H. Molecular structure of the bilin binding protein (BBP) from *Pieris brassicae* after refinement at 2.0 Å resolution. J. Mol. Biol. 198(3):499–513, 1987.
7. Holden, H.M., Rypniewski, W.R., Law, J.H., Rayment, I. The molecular structure of insecticyanin from the tobacco hornworm *Manduca sexta* L. at 2.6 Å resolution. EMBO J. 6(6):1565–1570, 1987.
8. Jones, T.A., Bergfors, T., Sedzik, J., Unge, T. The three-dimensional structure of P2 myelin protein. EMBO J. 7(6): 1597–1604, 1988.
9. Smith, T.F., Waterman, M.S. Comparison of biosequences. Adv. Appl. Math. 2:482–489, 1981.
10. Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. The Protein Data Bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
11. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
12. Richmond, T.J., Richards, F.M. Packing of α-helices: Geometrical constraints and contact areas. J. Mol. Biol. 119: 537, 1978.
13. Eisenberg, D., Wesson, M., Yamashita, M. Interpretation of protein folding and binding with atomic solvation parameters. Chem. Scripta 29A:217–221, 1989.
14. Lesk, A.M., Chothia, C. How different amino acids sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. J. Mol. Biol. 136:225, 1980.
15. Kabat, E.A., Wu, T.T., Reid-Miller, M., Perry, H.M., Gottesman, K.S. "Sequences of Proteins of Immunological Interest," 4th ed. U.S. Department of Health and Human Services, 1987.
16. Lee, B.K., Richards, F.M. The interpretation of protein structures: Estimation of static accessibility. J. Mol. Biol. 55:379–400, 1971.
17. Janin, J. Surface and inside volumes in globular proteins. Nature (London) 277:491–492, 1979.
18. Chothia, C. The nature of the accessible and buried surfaces in proteins. J. Mol. Biol. 105(1):1–12, 1976.
19. Rose, G.D., Gerzlowitz, A.R., Lesser, G.J., Lee, R.H., Zehfus, M.H. Hydrophobicity of amino acids in globular proteins. Science 229:834–838, 1985.
20. Richardson, J. The anatomy and taxonomy of protein structure. Adv. Protein Chem. 34:167–334, 1981.
21. Monaco, H.L., Zanotti, G., Spadon, P., Bolognesi, M., Sawyer, L., Eliopoulos, E.E. Crystal structure of the trigonal form of bovine beta-lactoglobulin and of its complex with retinol at 2.5 Å resolution. J. Mol. Biol. 197(4):695–706, 1987.
22. Banaszak, L.J. Personal communication, 1990.
23. Pevsner, J., Reed, R.R., Feinstein, P.G., Snyder, S.H. Molecular cloning of odorant-binding protein: Member of a ligand carrier family. Science 241:336–339, 1990.
24. Godovac-Zimmermann, J. The structural motif of β-lactoglobulin and retinol-binding protein: A basic framework for binding and transport of small hydrophobic molecules? Trends Biochem. Sci. 13:64–66, 1988.
25. Cowan, S.W., Newcomer, M.E., Jones, T.A. Crystallographic refinement of human serum retinol binding protein at 2 Å resolution. Proteins 8(1):44–61, 1990.
26. Hunt, L.T., Elzanowski, A., Barker, W.C. The homology of complement factor C8 gamma chain and α_1-microglobulin. Biochem. Biophys. Res. Commun. 149:282–288, 1987.

27. Hendrickson, W.A., Pahler, A., Smith, J.L., Satow, Y., Merritt, E.A., Phizackerley, R.P. Crystal structure of core streptavidin determined from multiwavelength anomalous diffraction of synchrotron radiation. Proc. Natl. Acad. Sci. U.S.A. 86(7):2190–2194, 1989.

28. Weber, P.C., Ohlendorf, D.H., Wendoloski, J.J., Salemme, F.R. Structural origins of high-affinity biotin binding to streptavidin. Science 243(4887):85–88, 1989.