Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 4355–4358, July 1987
Biochemistry

Profile analysis: Detection of distantly related proteins

(amino acid/sequence comparison/protein structure/globin structure/immunoglobulin structure)

MICHAEL GRIBSKOV*, ANDREW D. MCLACHLAN†, AND DAVID EISENBERG*

*Molecular Biology Institute and Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90024; and †Medical Research Council, Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England, United Kingdom

Communicated by Paul Boyer, February 17, 1987 (received for review November 19, 1986)

ABSTRACT **Profile analysis is a method for detecting distantly related proteins by sequence comparison. The basis for comparison is not only the customary Dayhoff mutational-distance matrix but also the results of structural studies and information implicit in the alignments of the sequences of families of similar proteins. This information is expressed in a position-specific scoring table (profile), which is created from a group of sequences previously aligned by structural or sequence similarity. The similarity of any other sequence (target) to the group of aligned sequences (probe) can be tested by comparing the target to the profile using dynamic programming algorithms. The profile method differs in two major respects from methods of sequence comparison in common use: (*i*) Any number of known sequences can be used to construct the profile, allowing more information to be used in the testing of the target than is possible with pairwise alignment methods. (*ii*) The profile includes the penalties for insertion or deletion at each position, which allow one to include the probe secondary structure in the testing scheme. Tests with globin and immunoglobulin sequences show that profile analysis can distinguish all members of these families from all other sequences in a database containing 3800 protein sequences.**

Our ability to determine the three-dimensional structures of proteins has been outstripped by our capacity to determine amino acid sequences from DNA sequences. New ways of inferring structure from sequence are needed, and a promising method is sequence comparison (1–3): if a newly discovered sequence is sufficiently similar to the sequence of a protein of known structure, we can infer that the two proteins have similar structures (e.g., see refs. 4 and 5). One problem in making such an inference is deciding what degree of sequence similarity is necessary to infer structural similarity (6, 7). A different problem, which we address in this paper, is the detection of similar but distantly related proteins.

This problem is illustrated by the globin family (8). The globin polypeptide chains from organisms as diverse as humans, insects, and plants are folded in the same general three-dimensional pattern, yet there are only two positions within the some 150 residues of the chain that contain the same amino acid in all globins. That is, this "globin fold" is encoded in many different amino acid sequences, some differing from others in as many as 130 positions. Any single globin sequence represents just one realization of the globin fold. In attempting to decide whether an amino acid sequence encodes the globin fold, we need a pattern or "profile" that represents the fold. The profile described below represents the fold as a position-dependent scoring matrix, giving our best estimate of the likelihood that each amino acid can fit into the known fold.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Common methods for detection of similarity depend on pairwise alignment of sequences—for example, the dot matrix method (9, 10) or dynamic programming methods (11–14). Another class of methods are the rapid database searching methods (15, 16). All of these normally test every sequence in the database independently against a single probe sequence without using information implicit in the alignments of families of related sequences or including information available from structural studies. [An exception is the family comparison dot matrix method (9), which, however, does not allow for insertion or deletion.] Profile analysis brings in both structural and family information at the expense of a modest increase in computation time.

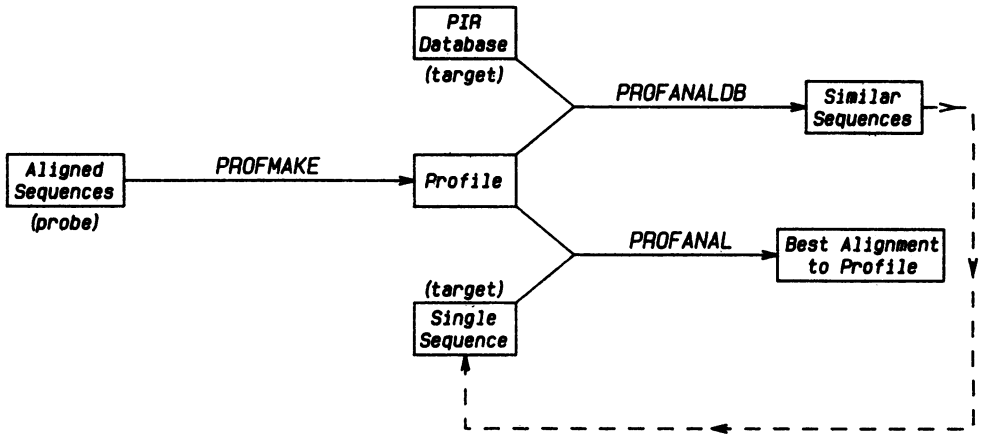
METHODS

Construction of the Profile (PROFMAKE). Profile analysis has two steps (Fig. 1*a*): (*i*) construction of the profile with the program PROFMAKE, and (*ii*) comparison of the profile with a database of sequences or a single sequence (program PROFANAL). The starting point for the creation of a profile is a sequence or group of sequences (the probe). This probe is usually a group of typical sequences of functionally related proteins that have been aligned by similarity in sequence or three-dimensional structure. Each sequence can be given a weight, which is useful when several of them are very similar. It is also possible to make a profile from a single sequence if additional information is used.

The profile is a sequence position-specific scoring matrix $M(p,a)$ composed of 21 columns and N rows (N = length of probe). The row p corresponds to a sequence position of the probe. The first 20 columns of each row specify the score for finding, at that position in the target, each of the 20 amino acid residues. An additional column contains a penalty for insertions or deletions at that position (Fig. 1*b*). In PROFMAKE, the profile is generated from the probe by using a comparison table derived from the mutational distance matrix (MDM78) of Dayhoff (17, 18). The value of the profile for amino acid a at position p is $M(p,a) = \sum_{b=1}^{20} W(p,b) \times Y(a,b)$, where $Y(a,b)$ is Dayhoff's matrix and $W(p,b)$ is a weight for the appearance of amino acid b at position p. This weight is determined as follows: Suppose that amino acid b appears $n(b,p)$ times in position p in the N_R probe sequences. Then a simple average weight is given by $W(b,p) = n(b,p)/N_R$. Another useful weighting (19) is to take W as proportional to $\log[n(b,p)/N_R]$, setting $n(b,p) = 1$ for any amino acid that never appears at p. This logarithmic weighting may be better when there are several closely related sequences, but the simple weighting is used for the examples given in this paper.

The profile specifies position-dependent penalties for insertions and deletions. Insertions and deletions in families of aligned homologous sequences tend to occur more frequently in regions between segments of regular secondary structure than within them. This is encoded in the 21st column of each row of the profile by a penalty for insertion/deletion of the corresponding probe residue. This penalty can be set high to

a



b

POS	PROBE	CONSENSUS		A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	+/-
1	E G V L	V	V	3	-2	3	4	0	4	-1	3	-1	4	4	1	1	1	-2	1	2	6	-6	-2	9
2	L L S P	L	L	2	-2	-2	-1	3	0	-1	3	-1	6	5	-1	3	0	-1	3	1	4	1	-1	9
3	V V V V	V	V	2	2	-2	-2	2	2	-3	11	-2	8	6	-2	1	-2	-2	0	2	15	-9	-1	9
4	K E A T	A	A	6	-2	5	6	-5	4	1	0	5	-2	0	3	3	3	1	3	6	0	-6	-4	9
5	A P L P	P	P	6	-1	0	1	-2	2	0	1	0	2	2	0	8	2	0	2	2	3	-5	-4	9
6	G G G G	G	G	7	1	7	5	-6	15	-1	-3	0	-4	-3	4	3	2	-3	6	4	2	-11	-7	9
7	S S Q E	D	D	4	-1	7	7	-6	7	2	-2	2	-3	-2	4	3	6	1	6	2	-1	-6	-5	9
8	S S T P	S	S	4	4	2	2	-4	4	-1	0	2	-3	-2	2	7	0	1	10	6	0	-2	-4	9
9	V L V A	V	V	5	0	-1	-1	3	1	-2	7	-2	7	6	-1	1	-1	-3	0	2	10	-5	-1	9
10	K R R S	R	R	0	-1	1	1	-5	0	2	-2	8	-3	1	3	3	3	10	5	1	-2	7	-5	9
11	M L I I	I	I	0	-2	-3	-2	7	-3	-3	11	-1	11	10	-2	-2	-1	-2	-2	1	9	-3	1	9
12	S S T S	S	S	4	6	2	2	-3	5	-1	0	2	-3	-2	3	4	-1	1	12	6	0	0	-4	9
13	C C C C	C	C	3	15	-5	-5	-1	2	-1	3	-5	-8	-6	-3	1	-6	-3	7	3	3	-13	10	9
14	K S Q R	K	K	1	-2	3	3	-6	1	3	-2	7	-3	0	3	3	5	7	4	1	-2	2	-5	9
15	A A G S	A	A	10	3	4	3	-5	8	-1	-1	1	-2	-1	3	4	1	-2	7	4	2	-6	-4	9
16	T S D S	S	S	4	3	5	4	-5	6	0	0	2	-3	-2	4	3	1	1	9	6	0	-3	-4	9
17	G G S Q	G	G	5	1	6	5	-6	9	1	-2	1	-3	-2	4	3	4	0	6	3	0	-6	-6	9
18	Y F L S	F	F	-1	2	-4	-3	9	-3	0	4	-3	6	3	-1	-3	-3	-3	1	-1	2	7	7	9
19	T T R L	T	T	1	-2	0	1	0	0	0	2	2	2	3	1	1	1	3	1	7	2	1	-2	9
20	F F . L	F	F	-2	-3	-6	-4	10	-4	-1	6	-4	9	6	-3	-4	-4	-3	-2	-1	3	7	8	4
21	S S . D	S	S	3	2	5	4	-4	5	0	-1	2	-3	-2	4	3	1	1	8	2	-1	-2	-3	4
22	S . . S	S	S	2	3	1	1	-2	3	-1	0	1	-2	-1	2	2	0	1	8	2	0	1	-2	4
23	. . . G	G	G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
24	. . . D	D	D	1	-1	4	3	-2	2	1	0	1	-1	-1	2	1	2	0	1	1	0	-3	-1	4
25	. . . G	G	G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
26	. A G N	A	A	6	0	4	3	-4	6	1	-1	1	-2	-1	5	2	2	-1	3	3	1	-5	-3	4
27	Y N Y T	Y	Y	0	5	0	-1	5	-1	2	1	-1	0	-1	4	-3	-2	-2	0	3	0	3	6	4
28	E D D Y	D	D	2	-2	9	8	-3	3	4	-1	1	-3	-2	5	-1	4	-1	1	1	-1	-6	0	9
29	L M A L	L	L	3	-5	-3	-1	6	-1	-2	6	-1	10	10	-2	0	0	-2	-1	0	6	-1	0	9
30	Y N A W	N	N	4	1	3	2	0	2	3	-1	1	-1	-1	8	0	1	-1	2	1	-1	-1	2	9
.	.													.										
.	.													.										
.	.													.										
48	S G N S	S	S	4	3	5	3	-4	7	0	-2	2	-4	-3	6	3	1	0	10	3	0	-2	-4	9
49	S S N Y	S	S	2	5	2	1	1	2	1	0	1	-2	-2	5	1	-1	0	8	1	-1	3	1	9

FIG. 1. The concept of a profile. (*a*) A flow diagram of profile analysis. (*b*) A 49-residue sample profile for the immunoglobulin variable-region domain, generated from the four-probe sequences shown at the left (see Fig. 2*b* for details). The profile is shown in the box. The rightmost column of the profile gives the penalty for insertion/deletion (+/−). Positions 31–47 of the profile are omitted from the figure for clarity. Notice that where gaps appear in some of the probe sequences, the insertion/deletion penalty is lower than elsewhere.

prevent insertions inside known regular secondary structural elements and set low to allow insertions in regions where insertions are observed in the probe. By setting the insertion/deletion penalty to zero at a given position, an insertion or deletion of any size is permitted. The penalty applied, PEN, for creating a gap during the match of profile to target is given by PEN = PEN′ [OPEN + EXN × L] in which PEN′ is the penalty given in column 21 of the profile, L is the number of residue positions in the gap, and OPEN and EXN are the penalties for gap opening and gap extension supplied interactively by the user of PROFANAL. In the examples of Fig. 2, OPEN was taken as 5 and EXN as 0.1. The value of the penalty at position p in the profile, PEN′(p) is computed in relation to the highest score YMAX in our modified version (18) of the Dayhoff matrix Y and the length of the longest gap LMAX(p) in the probe that includes position p as follows: PEN′ = YMAX/[OPEN + EXN + LMAX(p)].

A consensus sequence $C(p)$ is generated at each position of the profile to aid the display of alignments of target sequences with the profile. The consensus residue c is the amino acid at

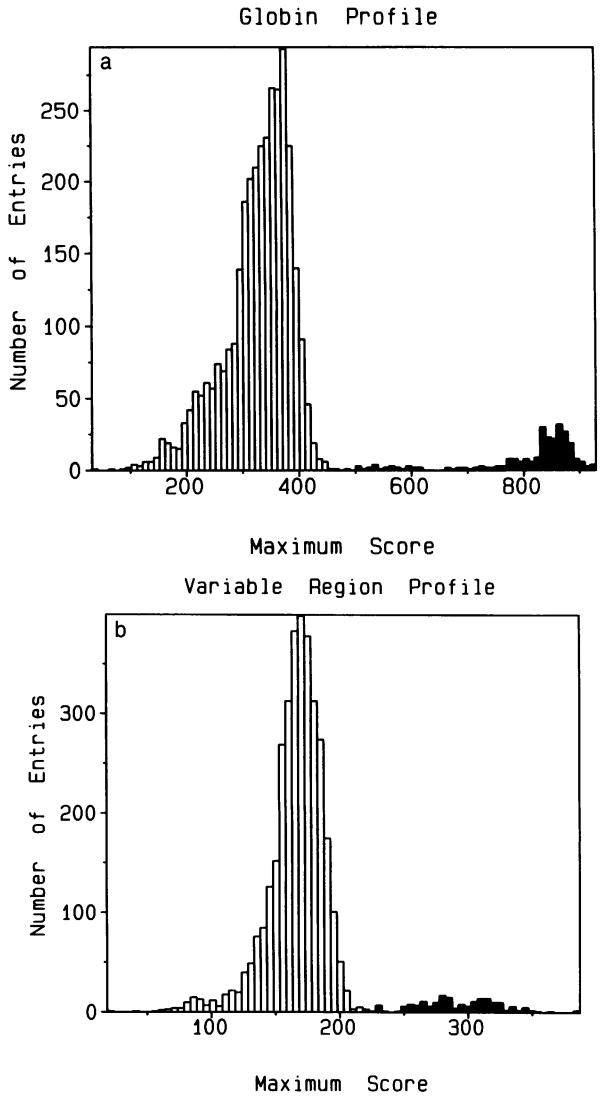


FIG. 2. Profile analysis of globin and immunoglobulin sequences. (*a*) Globins: Distribution of scores for the comparison of a profile generated from human α-hemoglobin, rhesus monkey β-hemoglobin, human myoglobin, lamprey cyanohemoglobin, and soybean leghemoglobin, to the PIR database. The profile is 124 residues long, with an average of 5.6 gapped positions per sequence. Scores for nonglobin proteins ranged from 32.4 to 455.8, with a mean of 326.9 and SD of 60. Scores for globin sequences (shaded) ranged from 469.5 to 928.7. (*b*) Variable region immunoglobulin domain: Distribution of scores for the comparison of a profile generated from 49-residue (including an average of 3.5 gaps per sequence) segments of two immunoglobulin heavy-chain (PIR entries G1MSAA and M3HUWE), one λ-chain (L3HUSH), and one κ-chain (K2HUCM) variable regions. The segments used to generate the profile correspond to a region beginning 5 residues before β strand B and ending 15 residues after β strand C as defined by Taylor (20). Scores for nonimmunoglobulin sequences (T-cell receptor and immunoglobulin constant regions excluded) ranged from 21.7 to 220.7, with a mean of 165.6 and SD of 21.7. Scores for immunoglobulin variable regions (shaded) ranged from 221.9 to 386.9. Immunoglobulin constant regions scored from 172.1 to 206.8.

p that has the highest score $M(p,c)$. It is the amino acid most mutationally similar to all the aligned residues of the probe sequences at its position, rather than merely the most common residue present in the collection of sequences.

Detection of Similar Sequences and Database Searching (PROFANAL). We use a modified dynamic programming algorithm to compare sequences to the profile. Dynamic programming algorithms are designed to generate the best alignment given replacement scores and penalties for insertions and deletions (11–14). The major modification to these algorithms for use with profiles lies in the scoring system. In the unmodified algorithm, the score at a given position in the alignment score matrix is based on the comparison of the amino acid residues at the corresponding positions in the two sequences. In profile analysis, the score is read from the column of the profile corresponding to the amino acid residue in the target sequence and the row corresponding to the position in the probe.

For the comparison of a single sequence to a profile, we generally use the local homology method of Smith and Waterman (12). In contrast, for dot matrix plots and database searches, we use the forward/backward matrix method (18). To simplify inspection of results, a traceback of the alignment is not generated; only the highest score or, for dot plots, scores above a specified threshold, are retained.

Equipment. The experiments described below were performed on a Digital Equipment (Maynard, MA) VAX 11/780 under the VAX VMS operating system (version 4.3). The computer has 16 Mbytes of memory and a floating point accelerator. Most sequences were obtained from the National Biomedical Research Foundation PIR database.‡ For a profile of 45 residues, the PIR database of 3800 sequences can be searched in ≈20 central processing unit min.

RESULTS

Globins. We have tested the ability of profile analysis to detect distantly related proteins by searching the PIR database with profiles for two protein families. Fig. 2*a* shows results for the globin family. The profile was made from five representative globin sequences aligned by structural criteria (21): human α-hemoglobin, rhesus monkey β-hemoglobin, human myoglobin, lamprey cyanohemoglobin, and soybean leghemoglobin. The scores for globin sequences are in every case greater than scores for nonglobin sequences. Even globins distantly related to the group used to make the profile—e.g., erythrocruorin and other monomeric globins of plants and invertebrates—are distinguished from nonglobin sequences. All scores for globin sequences, including the newly sequenced bacterial hemoglobin (22), are at least 2.4 SD above the mean of the nonglobin sequences.

For comparison, we searched the database with the Lipman–Pearson FASTP (15, 16) algorithm using human α-hemoglobin as a probe. The FASTP program selected 244 of the 271 globins in the database for score optimization. Even after score optimization, the leghemoglobins could not be clearly distinguished from nonglobin sequences.

Immunoglobulins. A second test of profile analysis was performed for immunoglobulin variable regions. The profile (Fig. 1*b*) was made from a 45-residue segment of four immunoglobulin variable-region sequences, two from heavy-chain sequences and two from light-chain sequences. This profile detected all the variable regions in the PIR database (Fig. 2*b*). The only sequence entries with scores greater than the lowest variable-region sequence are those of the α and β subunits of the T-cell receptor. These proteins are thought to be homologous to immunoglobulins (23), which is consistent with their scores. The FASTP program, when used with a 45-residue heavy-chain probe, detected heavy-chain variable regions but also picked out some other variable-region sequences.

‡Protein Identification Resource (1986) Protein Sequence Database (Natl. Biomed. Res. Found., Washington, DC), Release 8.0.

DISCUSSION

Selectivity of Profile Analysis. An ideal method for detecting homologous proteins would separate a database of sequences into two groups with no overlap in scores between them: the homologous proteins and all other proteins. Fig. 2 suggests that profile analysis is powerful by this criterion. This selectivity comes from (*i*) the information implicit in aligned sequences, encoded in the flexible scoring system of the profile, and (*ii*) the ability of dynamic programming methods to position gaps, as guided by the penalties in the profile. The essence of the profile is that both the gap penalty and amino acid preference are position dependent. The position-dependent gap penalty introduces structural information, such as the known locations of secondary structure elements. The position-dependent amino acid preference introduces information about the character of the allowed side chains in each position.

Comparison with Other Methods. The profile method is useful for learning whether a protein sequence belongs to a known family of sequences. The method differs from both rapid database methods and standard dynamic programming methods in that these methods are designed for pairwise, rather than family, comparisons. Dynamic programming methods have been applied to align three sequences (24) but may be hard to apply for large numbers of sequences. With dynamic programming methods, information from a family of proteins can be included by comparing the members of the family by twos or threes and then synthesizing an overall alignment from the individual alignments. This tedious process is replaced in profile analysis by the position-specific scoring table.

The profile method shares characteristics of template methods. Template (20, 25) or fingerprint (27) methods fit a sequence to a rigid pattern of amino acid residues with no gaps allowed. This rigidity can be softened by breaking the template into segments separated by variable-length regions where any residue is allowed (functionally equivalent to gaps). The size of these regions is determined either by fitting each segment independently and checking that the order and spacing of the segments is reasonable (20), or by making a different template for every possible allowed spacing (27).

A template can be considered a special case of a profile in which any amino acid occurring in the probe sequences is given a score of 1.0, and in which the insertion/deletion penalty is set high in regions corresponding to segments (to prevent gaps), and low in the regions between segments. In contrast, profile analysis assigns positive scores even to target amino acid residues that are not observed in the probe and permits gaps within segments if a much better alignment can be obtained. Profile analysis thus includes template and fingerprint methods as special cases.

Extensions of the Method. Any set of properties that can be represented as similarity or difference scores for pairs of amino acids can be used to construct profiles. The scoring system used in the examples shown here is based on observed frequencies of replacement in homologous proteins. Other properties such as hydrophobicity, α or β structural preference (28), or side-chain volume can be used as scoring tables.

A possible eventual use for the profile method is to infer information on three-dimensional structure from sequence. Creation of a set of profiles for a variety of protein families will offer a library of structural motifs. Comparison of any newly discovered sequence with the library may yield information on structural motifs within the protein.

Copies of this program may be obtained from the authors at the University of California at Los Angeles. Programs are available in a format compatible with the University of Wisconsin Genetics Computer Group (UWGCG) software package or in an independent implementation. Program development was aided by the UWGCG procedure library (26).

We thank Drs. A. M. Lesk and C. Chothia for discussions of their template method for comparison. This work was supported by grants from the National Science Foundation (PCM 82-07520), National Institutes of Health (GM 31299), the University of California Biotechnology Research and Education Program, the Simon Guggenheim Foundation to D.E., and the National Cancer Society (PF-2649) and Lita Annenberg Hazen to M.G.

1. Fitch, W. M. (1966) *J. Mol. Biol.* **16,** 9–16.
2. McLachlan, A. D. (1972) *J. Mol. Biol.* **62,** 409–424.
3. Doolittle, R. F. (1981) *Science* **214,** 149–159.
4. Blundell, T. L., Sibanda, L. & Pearl, L. (1983) *Nature (London)* **304,** 273–275.
5. Sweet, R. M. (1986) *Biopolymers* **25,** 1565–1577.
6. Kabsch, W. & Sander, C. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 1075–1078.
7. Sweet, R. M. & Eisenberg, D. (1983) *J. Mol. Biol.* **171,** 479–488.
8. Dickerson, R. E. & Geis, I. (1983) *Hemoglobin* (Benjamin/Cummings, Menlo Park, CA).
9. McLachlan, A. D. (1983) *J. Mol. Biol.* **169,** 15–30.
10. Maizel, J. V., Jr., & Lenk, R. P. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 7665–7669.
11. Needleman, S. B. & Wunsch, C. D. (1970) *J. Mol. Biol.* **48,** 443–453.
12. Smith, T. F. & Waterman, M. S. (1981) *Adv. Appl. Math.* **2,** 482–489.
13. Sellers, P. H. (1974) *SIAM J. Appl. Math.* **26,** 787–793.
14. Boswell, D. R. & McLachlan, A. D. (1984) *Nucleic Acids Res.* **12,** 457–464.
15. Wilbur, W. J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 726–730.
16. Lipman, D. J. & Pearson, W. R. (1985) *Science* **227,** 1435–1442.
17. Dayhoff, M. O. (1979) in *Atlas of Protein Sequence and Structure*, eds. Schwartz, R. M. & Dayhoff, M. O. (Natl. Biomed. Res. Found., Washington, DC), Vol. 5, Suppl. 3, pp. 353–358.
18. Gribskov, M. & Burgess, R. R. (1986) *Nucleic Acids Res.* **14,** 6745–6763.
19. Staden, R. (1984) *Nucleic Acids Res.* **12,** 505–519.
20. Taylor, W. R. (1986) *J. Mol. Biol.* **188,** 233–258.
21. Lesk, A. M. & Chothia, C. (1980) *J. Mol. Biol.* **136,** 225–270.
22. Wakabayashi, S., Matsubara, H. & Webster, D. A. (1986) *Nature (London)* **322,** 481–483.
23. Yanagi, Y., Yoshikai, Y., Leggett, K., Clark, S. P., Aleksander, I. & Mak, T. W. (1984) *Nature (London)* **308,** 145–149.
24. Murata, M., Richardson, J. S. & Sussman, J. L. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 3073–3077.
25. Taylor, W. R. & Thornton, J. M. (1984) *J. Mol. Biol.* **173,** 487–514.
26. Devereux, J., Haeberli, P. & Smithies, O. (1984) *Nucleic Acids Res.* **12,** 387–395.
27. Wierenga, R. K., Terpstra, P. & Hol, W. G. J. (1986) *J. Mol. Biol.* **187,** 101–107.
28. Chou, P. Y. & Fasman, G. D. (1978) *Annu. Rev. Biochem.* **47,** 251–276.