Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 1379–1383, February 1993
Biophysics

Three-dimensional profiles from residue-pair preferences: Identification of sequences with β/α-barrel fold

(inverse folding/protein sequence/protein structure/amino acid biosynthesis)

MATTHIAS WILMANNS AND DAVID EISENBERG*

Molecular Biology Institute and Department of Chemistry and Biochemistry, University of California at Los Angeles, Los Angeles, CA 90024-1570

Contributed by David Eisenberg, October 16, 1992

ABSTRACT **The three-dimensional profile method expresses the three-dimensional structure of a protein as a table, the profile, which represents the local environment of each residue. The score of an amino acid sequence, aligned with the three-dimensional profile, reflects its compatibility with the profiled structure. In the original implementation, each local environment was characterized by its polarity, the area buried of its side chain, and its secondary structure. Here we describe a modified three-dimensional profile algorithm that characterizes the local environment in terms of the statistical preferences of the profiled residue for neighbors of specific residue types, main-chain conformations, or secondary structure. Combined profiles of the original and the three new types were tested on β/α-barrel protein structures. The method identified the following enzymes of unknown three-dimensional structure as probable β/α-barrels, all of which catalyze reactions in the biosynthesis of aromatic amino acids: anthranilate phosphoribosyltransferase (trpD), glutamine amidotransferase (trpG), and phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase (hisA).**

Some two dozen enzymes are known to have the β/α-barrel fold: a central cylindrical, eight-stranded β-sheet surrounded by eight α-helices (1). A subset of these enzymes contains a common phosphate binding site formed by loops at the C termini of β-strands 7 and 8 (2). Yet, despite their resemblance in overall fold, these β/α-barrel proteins do not share overall sequence similarity. The sequences are in fact so diverse that new β/α-barrel folds were detected by x-ray structural studies rather than by sequence analysis. A notable exception was the successful prediction of the β/α-barrel fold of the α-subunit of tryptophan synthase (trpA) based on secondary structure prediction for a set of 10 aligned sequences (3). Numerous other entries in the amino acid sequence data base almost certainly belong to the β/α-barrel family, but which ones?

The goal of the three-dimensional (3D) profile method (4) is the assignment of amino acid sequences to known protein folds. In this method a table, the profile, is computed from the coordinates of a known 3D structure. This profile measures the compatibility of amino acid sequences with the 3D structure (5). High scoring sequences are likely to exhibit the fold of the profiled structure.

In the original implementation of 3D profiles, each residue position in the 3D structure was characterized by three environmental properties: its area buried, its secondary structure, and the polarity of its environment. On the basis of the values of these three properties, the position was assigned to 1 of 18 possible environmental classes. Each of these classes was characterized by a set of 20 3D–1D (1D, one dimensional) scores for the likelihood of finding each of the 20 amino acid residues in that class. A sequence was scored for compatibility with the profiled structure by finding the alignment that results in the highest total score for every aligned position. In 3D compatibility searches against a sequence data base, proteins of the same fold were detected by some 3D profiles even in the absence of detectable similarity with the sequence of the profiled structure.

However, 3D profiles prepared from β/α-barrel structures detected only sequences having significant similarity with the sequence of the profiled structure. Thus, identification of sequences having the β/α-barrel fold remains a challenge. We describe here a modified 3D profile algorithm based on preferences for residue pairs of the local environment of each profiled residue. We have tested these 3D profiles on known β/α-barrels and we identify sequences that show high compatibilities with the 3D profiles of these structures.

Our residue-pair preference algorithm moves 3D profiles toward expression of the free energy that a sequence would acquire if folded as the profiled structure. That is, the 3D–1D scores based on preferences for particular residue pairs in the local environment resemble in spirit the statistical estimates of short distance residue-pair interaction energies derived in earlier studies (6–9) and interaction energies computed by linear distance relationships for residue pairs (10, 11). The use of such residue–residue scores to evaluate the energy of a sequence folded as a given structure has been pioneered by Sippl and other researchers (9, 12–14).

METHODS

Sequence Data Bases. An in-house data base PROT-99 with 26,648 nonidentical amino acid sequences has been created from the National Biomedical Resource Foundation (release 43.0) and GENPEPT (20/10/90) data bases (R. Lüthy, personal communication). From the PROT-99 data base, two 3D structure-oriented data bases have been extracted. The first data base, called KNOWN, contains 2833 sequences with known folds. The sequences were identified by using sequences of 128 known protein structures as probes in FASTA (15) and extracting all sequences that share at least 30% sequence similarity for a continuous segment of at least 50 residues or 25% of the length of the template sequence. The protein structures included all entries of the representative structure set (16) and a few recently determined structures. The second data base, called BARREL, contains 101 sequences known to have the β/α-barrel fold and is extracted from the data base KNOWN.

3D Structure Data Base for Profile Scoring Tables. The 3D–1D scoring tables for the residue-pair preference profiles are derived from a local data base that consists of 110 protein

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.

Abbreviations: 3D, three dimensional; 1D, one dimensional; TIM, triose phosphate isomerase; GOX, glycolate oxidase; FCB, flavocytochrome b_2; RuBisCO, ribulose 1,5-bisphosphate carboxylase/oxygenase.

*To whom reprint requests should be addressed.

crystal structures where 99 coordinate sets are taken from the Brookhaven data base (17) and the remaining ones were made available to us by courtesy of the investigators of the respective structures. Structures were included only if they were determined at 2.5 Å resolution or better, had an R value of $\leq$25%, and displayed good stereochemistry. In cases in which several structures of the same protein were available, only the best structure was used. Subunits of oligomeric proteins or noncovalently bonded complexes were separated into single-chain entities, and only single copies of identical molecules were used for evaluation of scoring tables. A detailed description of the local data base, the evaluation of scoring tables, and presentation of test results will be published elsewhere.

Residue-Pair Preference Profiles. In the residue-pair preference profiles, each score value, denoted $s_x^{i,j}$, represents the logarithmic likelihood ratio for the occurrence of residue pair (i, j) with pair property x in our 3D structure data base:

$$s_x^{i,j} = \ln(P_x^{i,j}/P^{i,j}),$$

where $P_x^{i,j}$ is the probability for the residue pair (i,j) with pair property x and $P^{i,j}$ is the probability for residue pair (i,j) with any pair property. Residue i is the profiled residue and residue j is a residue of the local environment. Starting at the N terminus of the 3D structure, each residue in turn is considered to be the profiled residue. The 3D–1D score for each profiled residue type a at position i, $R_x^{i(a)}$, is computed as the sum of scores $s_x^{i,j}$ for all pairs of the profiled residue and residues of its local environment, divided by the number of residue pairs found for residue position i, $m(i)$:

$$R_x^{i(a)} = \sum_{\substack{j=1[d(i,j)\leq d(\text{sphere});\\ j\neq i,i\pm 1]}}^{n} s_x^{i,j}/m(i).$$

The division by $m(i)$ adjusts for the variable number of residue pairs observed at different positions in the 3D structure: up to 50 pairs are found in the hydrophobic core and fewer pairs are found on the protein surface. This procedure is repeated by replacing the residue type of the profiled residue by the remaining 19 amino acid types while keeping the environmental residues unchanged. The resulting 20 numbers are the residue 3D–1D scores of this row i of the profile. The local environment is bounded by a sphere of 10–12 Å around the C^{β} position of the profiled residue, depending on which of the three profile types is used (Fig. 1). Virtual C^{β} positions are calculated for glycines. Pairs of the profiled residue with direct sequence neighbors ($i - 1$, $i + 1$) are excluded. The resulting 3D profile is expressed as a matrix, with n rows for an n-residue profiled structure and with 22 columns for the 3D–1D scores. The columns represent each of the 20 amino acids as well as penalty scores for opening and extending gaps. In the residue-pair preference profiles, the gap penalty values are currently set to a constant number for all residues but variable penalties can be used. Notice that row i of the profile encodes the pair interactions with the local environment of each of the 20 residue types at position i. Thus, the profile may be thought of as an expression of the energy of the 3D structure with each of the 20 residue types at each position.

Once a sequence of n residues has been aligned with the profile of a given 3D structure, the overall profile score S_x is given by:

$$S_x = \sum_{i=1}^{n} R_x^{i(a)},$$

in which $R_x^{i(a)}$ is the 3D–1D score of the residue type a at position i.

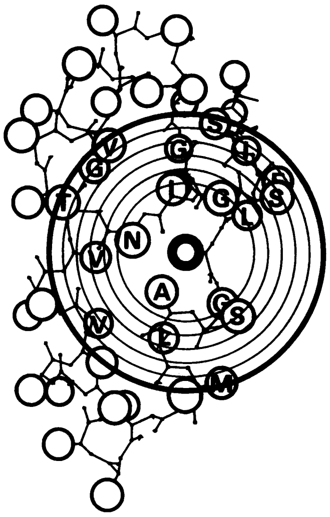

FIG. 1. Characterization of the local environment for type II profiles. In the protein fragment, only the main-chain and the C^{β} atoms highlighted by circles (radius, 1 Å) are displayed. Profiled residue (boldface) is in the center of the protein fragment. Type II profiles take into account residue types (one-letter code amino acid labels). Concentric circles around the profiled residue represent distance shells of the local environment.

Types of Profiles. Three new types of 3D profiles (II, III, and IV) are computed from a given three-dimensional structure and its sequence. Type I 3D profiles are generated with the original 3D profile method (4).

Type II profiles reflect the preference of the profiled residue of a given residue type (Ala, Cys, . . .) to interact with residues, each of a given type within the local environment at given distances. Type II 3D–1D scores are given in a 210 × 8 matrix. The 210 rows of this matrix represent all possible pairs of residue types (e.g., Ala-Ala, Ala-Cys), which are equivalent to the number of unique elements in a 20 × 20 symmetric matrix. The eight columns represent the distance shells of the local environment around the profiled residue. The outer radius of the first shell is 5 Å. Each of the remaining distance shells increases the radius of the sphere by 1 Å. The characterization of the local environment for type II profiles is illustrated in Fig. 1.

Type III profiles are based on the preference of the profiled residue of given residue type and main-chain conformation to interact with residues of given main-chain conformations within the local environment at given distances. The main-chain conformation of each residue is categorized by its dihedral angles ϕ and ψ and is classified as helical (α), extended (β), or neither α nor β (γ), based on a simplified version of the (ϕ, ψ) plot (18). The residue types of the environmental residues are not considered. Type III 3D–1D scores are given in an 80 × 9 matrix. The 80 rows of this matrix represent the 20 residue types of the profiled residue, interacting at four different distances (6, 8, 9.6, and 10.6 Å) with residues of the local environment. The distances have been chosen in such a way that about equal overall frequencies of residue pairs are found. The nine columns represent all possible pair combinations of (ϕ, ψ) classes of residue pairs (e.g., $\alpha\alpha$, $\alpha\beta$, . . .).

Type IV profiles reflect the preference of a profiled residue of a given residue type and secondary structural type to interact with residues of given secondary structural type within the local environment at a given distance. The secondary structure of each residue is categorized to be helical (H), β-sheet (E), or coil (C), including turns and bends, with the program DSSP (19). As in type III profiles, the residue types of the environmental residues are not considered. The format of the scoring table for type IV 3D–1D profile scores is identical to the scoring table for type III profile score. The nine columns represent all possible pair combinations of secondary structural classes of residue pairs (e.g., HH, HE, . . .). In contrast to type III profiles, type IV profiles depend on the structure of the 3D environment because they reflect the hydrogen bonding pattern to neighboring residues. The two methods differ most in that in method III only ≈10% of all residues fall in class γ, whereas in method IV about half of all residues are classified as C. Both type III and type IV profiles can be represented by a figure similar to Fig. 1 except that the residues of the local environment are represented by (ϕ, ψ) or secondary structure classes rather than by side-chain types.

3D Compatibility Search. Our procedure for applying combined 3D profiles to a given protein is to generate profiles of all four types I, II, III, and IV from its atomic coordinates. From these 4 individual profiles, combined profiles of all possible combinations (4 singlets, 6 doublets, 4 triplets, 1 quartet) are generated by summing the corresponding elements of the normalized profile matrices. The normalization of each profile is based on the sum of all absolute score values. All 15 profiles are used for 3D compatibility searches with the program PROFILESEARCH (20).

RESULTS AND DISCUSSION

We combined the original and pair-preference 3D profile methods for the identification of sequences of β/α-barrel proteins. 3D profiles of types I, II, III, and IV were generated from 12 high-resolution coordinate sets of 9 known β/α-barrel structures. They include triose phosphate isomerase (TIM) from yeast (21) and from *Trypanosoma brucei brucei* (22); the α subunit of tryptophan synthase (trpA; ref. 23), indoleglycerol phosphate synthase (trpC; ref. 24), and phosphoribosylanthranilate isomerase (trpF; ref. 24); glycolate oxidase (GOX; ref. 25); flavocytochrome b_2 (FCB; ref. 26); and the large subunit of L_8S_8 ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCO) from spinach (27) and tobacco (28). The results described in the following are obtained from 3D compatibility searches with all 15 combinations of 3D profiles against all 14,439 sequences with ≥150 residues of our PROT-99 data base.

Identification of Sequences with Known β/α-Barrel Fold. We illustrate our method with the identification of trpA sequences with combined 3D profiles from the GOX structure. As in earlier work (5), compatibilities of sequences with the profile were expressed by the Z score, the number of standard deviations over the mean profile score, for sequences of similar length.

Table 1 shows that trpA sequences have reasonably high Z scores with most of the 15 different 3D profiles despite the lack of sequence similarity between trpA and GOX. All combined 3D profiles merged from at least three profile types identify one or two trpA sequences with Z scores of ≥6.0. Four of six profiles, combined from two types of 3D profiles, detect one trpA sequence with Z scores of ≥6.0. In contrast, the sequence pattern search program FASTA (15) does not detect any trpA sequences when using the GOX sequence from spinach (the only GOX sequence in the data base) as the search motif.

Our test results on combined 3D profiles from all seven β/α-barrel proteins are summarized in Table 2 by listing the highest scoring sequences with known β/α-barrel folds. Profiles of all 3D structures, except TIM from yeast, are capable of identifying trpA as a β/α-barrel (Z score, ≥6.0). TrpC and GOX are detected by one and two other β/α-barrel structures, respectively. TIM, trpF, and RuBisCO have not been clearly identified as β/α-barrel folds. This weaker detection may be related to the deviation of their structures from the ideal β/α-barrel fold. The central barrel of TIM is more elliptical than the other barrel structures (29). In trpF from *Escherichia coli* one of the eight helices is replaced by an extended peptide segment (24). RuBisCO contains an additional N-terminal domain of 160 residues and also has several long loops (27, 28).

Table 1. Tests on 3D profiles from GOX by identifying trpA sequences

Profile type(s)	trpA sequences with Z scores ≥6/4	Z score (rank) of highest scoring sequence		
		trpA	Unknown fold	Non-β/α-barrel
I	1/5	9.1 (4)	7.8 (5)	5.2 (19)
II	0/0		5.4 (2)	3.8 (38)
III	0/2	4.1 (35)	5.6 (4)	4.8 (11)
IV	0/6	5.3 (17)	7.5 (2)	4.1 (49)
I–II	0/5	5.3 (11)	5.8 (4)	5.5 (9)
I–III	1/3	6.2 (12)	7.2 (5)	6.7 (6)
I–IV	1/5	6.1 (3)	5.5 (5)	3.6 (69)
II–III	1/2	6.2 (5)	5.8 (6)	3.9 (55)
II–IV	0/3	4.8 (48)	8.2 (2)	4.2 (65)
III–IV	1/2	6.2 (2)	5.4 (3)	4.4 (12)
I–II–III	1/3	7.3 (4)	6.6 (5)	5.2 (12)
I–II–IV	2/5	6.5 (4)	5.9 (6)	4.2 (34)
I–III–IV	1/3	6.8 (4)	5.8 (5)	5.3 (8)
II–III–IV	1/3	7.1 (2)	5.1 (6)	3.7 (47)
I–II–III–IV	2/3	7.8 (4)	5.8 (6)	4.9 (9)

For each profile type, statistics are given for 3D-compatibility searches of 3D profiles from GOX for the highest scoring trpA sequences of the PROT-99 data base. In subsequent columns are statistics for sequences of unknown fold (not in the KNOWN data base) and sequences not folded as β/α-barrels (in the KNOWN data base and not in the BARREL data base). Ranks of these sequences, sorted with respect to their Z scores, are given in parentheses. Rows below the top four show results for combinations of 3D profile types. Lower threshold is Z score of 4.0. Gap opening and gap length penalties were, respectively, 7.0 and 0.05.

Validation of Test Results. To assess the effectiveness of residue-pair preference profiles for identifying β/α-barrel sequences compatible with a known 3D β/α-barrel structure, we have analyzed the sequences with Z scores of ≥4.0 of 3D compatibility searches of all 180 β/α-barrel profiles (15 profiles of 12 coordinate sets) with respect to the KNOWN and BARREL sequence data bases. All such sequences in the BARREL data base are β/α-barrels and are classified as positives. All such sequences in the KNOWN data base but not in the BARREL data base are not folded as β/α-barrels and are classified as false positives. Fig. 2 compares the numbers of positive and false-positive sequences as functions of their Z scores. In general, a Z score of ≥7 for a sequence indicates a high probability that its 3D fold is similar to the profiled structure; a Z score of ≥5.5 indicates a moderate probability that the sequence has a fold similar to the profiled structure.

New β/α-Barrel Sequences. 3D compatibility searches provide the opportunity to identify sequences without known 3D structures as β/α-barrel folds. Here we consider proteins that are detected with Z scores of ≥6 by at least two different β/α-barrel proteins (Table 2). By this criterion, two other enzymes of the tryptophan biosynthesis pathway, anthranilate phosphoribosyltransferase (trpD) and glutamine amidotransferase (trpG), are identified by the profile of the trpC structure, and sequences of trpD are detected by the 3D profile from trpF. The β/α-fold for trpG has been independently predicted by the Garnier–Oshguthospe–Robson method from aligned sets of trpG sequences (T. Niermann and K. Kirschner, personal communication). We conclude that at least five enzymes of the tryptophan biosynthesis pathway (trpA, trpC, trpD, trpF, trpG), although not sharing

Table 2. *Z* scores of known and possible β/α-barrels

	No. of known (k) and possible (p) β/α-barrel sequences (PROT-99 data base)									
	TIM	trpA	trpC	trpF	GOX	FCB	RuBisCO	trpD	trpG	hisA
Codes for 3D	15	15	12	12	1	2	23	12	13	6
structures	k	k	k	k	k	k	k	p	p	p
TIM* (229/248)	79.6	4.5							6.2	5.7
TIM† (231/250)	79.3	6.3						4.2	4.9	
trpA (213/268)		69.7	4.4	5.5	6.5		4.8			8.6
trpC (213/254)	4.5	6.5	63.7		6.2		4.8	7.9	9.3	6.4
trpF (175/197)		6.7	4.3	50.1	4.0		4.7	6.3	4.5	5.2
GOX (209/359)	4.8	9.1	6.8	5.6	74.4	26.1				7.3
FCB (234/511)		7.5	5.7	4.1	34.0	63.6				6.3
RuBisCO‡ (249/442)	5.8	4.5					58.8			
RuBisCO§ (245/442)		6.7					56.2			

Z scores for compatibility of amino acid sequences for known (k) and possible (p) β/α-barrel proteins from 3D profiles of nine β/α-barrel structures. *Z* score of the highest scoring sequence of the protein is listed if it is $\geq$4.0. Gap opening and gap length penalties in 3D-compatibility searches were, respectively, 7.0 and 0.05. First number in parentheses indicates number of residues used to create profiles of limited β/α-barrels. Second number in parentheses represents overall length of the sequence of the respective structure.

*TIM from yeast.

†TIM from *T. brucei brucei*.

‡Rubisco from tobacco.

§Rubisco from spinach.

significant sequence similarity, are all folded as β/α-barrels (Fig. 3). The only enzyme of the trypophan biosynthesis pathway with unknown 3D structure that was not identified as a β/α-barrel is trpE (highest *Z* score, 5.1).

The biosynthetic pathways of the aromatic amino acids phenylalanine, tyrosine, and tryptophan branch from chorismate. It is thus intriguing that chorismate synthase (aroC) and the bifunctional enzymes from *E. coli* that catalyze the first two reactions of the phenylalanine and tryptophan pathways (pheA, tyrA) are all found with *Z* scores of $\geq$6 by profiles of one β/α-barrel (Fig. 3). Also isochorismate synthase (entC), which catalyzes the initial reaction of the enterobactin pathway (30), is found by profiles with *Z* scores $\geq$6 from two different β/α-barrels. These scores, when considered in light of Fig. 2, indicate only a moderate probability for β/α-barrel folds, and firm classification of these folds must await further computational and experimental developments.

However, a stronger probability is received by sequences of an enzyme of the biosynthetic pathway of the heteroaromatic amino acid histidine, phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase (hisA). HisA sequences are detected by four 3D profiles of β/α-barrels (trpA, trpC, GOX, FCB; Table 2) with *Z* scores up to 8.4. Profiles of trpF and TIM from yeast find hisA sequences with *Z* scores of $\geq$5. Another enzyme of the histidine

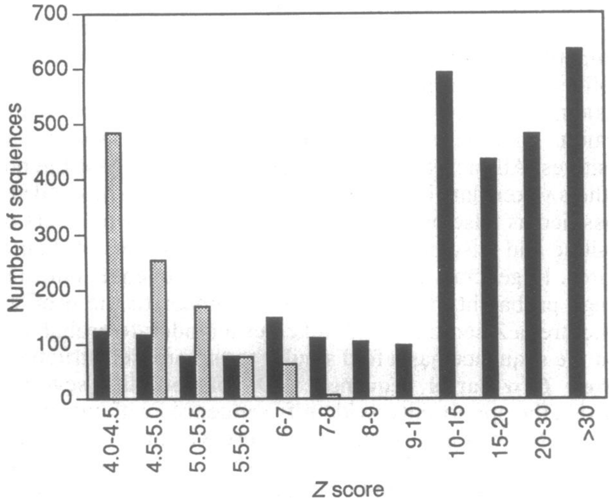

FIG. 2. Effectiveness of residue-pair preference profiles for detecting β/α-barrel sequences as tested on the KNOWN data base, in which the fold of each sequence is known. Numbers of positive (β/α-barrel) and false-positive (non-β/α-barrel) sequences are represented, respectively, by solid bars and shaded bars as a function of the *Z* score. Statistics are based on *Z* scores of 2.6×10^7 sequences (180 3D compatibility searches against the PROT-99 data base). Histogram shows ≈11% of all positive and 0.1% of all false-positive sequences. Remaining positive and false-positive sequences have *Z* scores of <4.

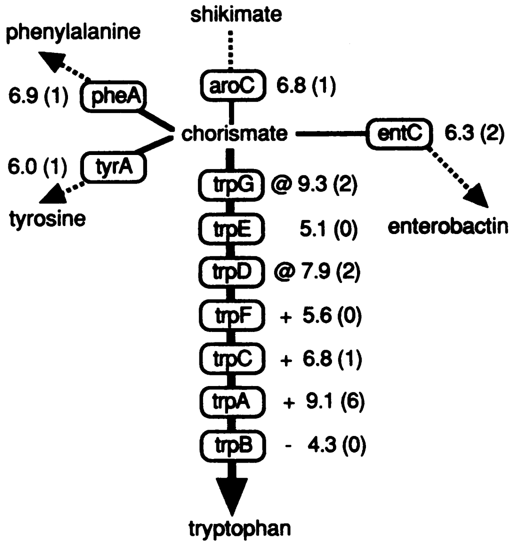

FIG. 3. Identification of β/α-barrel folds of sequences of enzymes involved in tryptophan biosynthesis pathway (thick line). Proteins are represented by their gene codes in boxes (see text). Number at each enzyme is the highest *Z* score of its sequence when scored against 3D profiles from known β/α-barrels. Proteins are marked with @ if their sequences were found with *Z* scores of $\geq$6 with at least two different β/α-barrels (numbers in parentheses). β/α-barrels and non-β/α-barrels with known 3D structures are indicated, respectively, with + and −. Chorismate synthase (aroC), isochorismate synthase (entC), and the bifunctional enzymes from *E. coli* catalyzing the initial reactions of the tyrosine and phenylalanine pathways (pheA, tyrA) have also been found with moderate *Z* scores in 3D compatibility searches of β/α-barrels and are therefore included in the figure (thin lines).

biosynthesis pathway, cyclase (hisF), shares weak sequence similarity with hisA (≈25%), especially at the putative phosphate binding site. Although neither of the two hisF sequences in our PROT-99 data base could be identified with high Z scores from β/α-barrel profiles, we assume that hisF has the same overall fold as hisA, which is probably a β/α-barrel (Table 2).

Concluding Remarks. Profiles based on residue-pair preferences, when used in concert with the original 3D profile method, enhance the assignment of sequences to known 3D structures. With these combined profiles, three sequences of unknown 3D structure score as probable β/α-barrel folds. None of them is similar to any known β/α-barrel by standard sequence comparison methods. Our results demonstrate that 3D protein folds of enzymes involved in biosynthesis of aromatic amino acids, if related by divergent evolution, have diverged more slowly than their sequences over evolutionary time. These results suggest that identification of protein folds by 3D profile analysis can be a tool in exploration of protein evolution.

We thank Drs. James Bowie, Roland Lüthy, Kam Zhang, Gilbert Privé, Michael Gribskov, Manfred Sippl, and Ron Bauerle for helpful discussions and National Institutes of Health for support. M.W. was in part supported by a research fellowship of the Deutsche Forschungsgemeinschaft.

1. Farber, G. K. & Petsko, G. A. (1990) *Trends Biochem. Sci.* **15**, 228–234.
2. Wilmanns, M., Hyde, C. C., Davies, R. D., Kirschner, K. & Jansonius, J. N. (1991) *Biochemistry* **30**, 9161–9169.
3. Crawford, I. R., Niermann, T. & Kirschner, K. (1987) *Proteins* **1**, 118–129.
4. Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991) *Science* **253**, 164–170.
5. Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 4355–4359.
6. Tanaka, S. & Scheraga, H. A. (1976) *Macromolecules* **9**, 945–950.
7. Miyazawa, S. & Jernigan, R. L. (1985) *Macromolecules* **18**, 534–552.
8. Gregoret, L. M. & Cohen, F. E. (1990) *J. Mol. Biol.* **211**, 959–974.
9. Hinds, D. A. & Levitt, M. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 2536–2540.
10. Sippl, M. J. (1990) *J. Mol. Biol.* **213**, 859–883.
11. Casari, G. & Sippl, M. J. (1992) *J. Mol. Biol.* **224**, 725–732.
12. Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M. J. (1990) *J. Mol. Biol.* **216**, 167–180.
13. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992) *Nature (London)* **358**, 86–89.
14. Godzik, A., Kolinski, A. & Skolnick, J. (1992) *J. Mol. Biol.* **227**, 227–238.
15. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
16. Hobohm, U., Scharf, M., Schneider, T. & Sander, C. (1992) *Protein Sci.* **1**, 409–417.
17. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Brice, E. F., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **196**, 199–216.
18. Rooman, M. J., Kocher, J.-P. A. & Wodak, S. J. (1991) *J. Mol. Biol.* **221**, 961–979.
19. Kabsch, W. & Sander, C. (1983) *Biopolymers* **22**, 2577–2637.
20. Gribskov, M., Lüthy, R. & Eisenberg, D. (1990) *Methods Enzymol.* **183**, 146–159.
21. Lolis, E. & Petsko, G. A. (1990) *Biochemistry* **29**, 6619–6625.
22. Wierenga, R. K., Noble, M. E. M., Vriend, G., Nauche, S. & Hol, W. G. J. (1991) *J. Mol. Biol.* **220**, 995–1015.
23. Hyde, C. C., Ahmed, S. A., Padlan, E. A., Miles, E. W. & Davies, D. R. (1988) *J. Biol. Chem.* **263**, 17857–17871.
24. Wilmanns, M., Priestle, J. P., Niermann, T. & Jansonius, J. N. (1992) *J. Mol. Biol.* **223**, 477–507.
25. Lindqvist, Y. (1989) *J. Mol. Biol.* **209**, 151–166.
26. Xia, Z. X. & Mathews, F. S. (1990) *J. Mol. Biol.* **212**, 837–863.
27. Knight, S., Andersson, I. & Brändén, C. I. (1990) *J. Mol. Biol.* **215**, 113–160.
28. Curmi, P. M. G., Cascio, D., Sweet, R. M., Eisenberg, D. & Schreuder, H. (1992) *J. Biol. Chem.* **267**, 16980–16989.
29. Lasters, I., Wodak, S. J., Alard, P. & van Cutsem, E. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 3338–3342.
30. Ozenberger, B. A., Brickman, T. J. & McIntosh, M. A. (1989) *J. Bacteriol.* **171**, 775–783.