Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 12098–12102, December 1992
Biochemistry

Sequence–structure matching in globular proteins: Application to supersecondary and tertiary structure determination

(inverse folding problem/protein structure/protein interaction pattern)

ADAM GODZIK AND JEFFREY SKOLNICK*

Department of Molecular Biology, The Scripps Research Institute, 10666 North Torrey Pines Road, La Jolla, CA 92037

Communicated by Marshall Fixman, August 17, 1992

ABSTRACT **A methodology designed to address the inverse globular protein-folding problem (the identification of which sequences are compatible with a given three-dimensional structure) is described. By using a library of protein fingerprints, defined by the side chain interaction pattern, it is possible to match each structure to its own sequence in an exhaustive data base search. It is shown that this is a permissive requirement for the validation of the methodology. To pass the more rigorous test of identifying proteins that are not close sequence homologs, but that have similar structure, the method has been extended to include insertions and deletions in the sequence, which is compared to the fingerprint. This allows for the identification of sequences having little or no sequence homology to the fingerprint. Examples include plastocyanin/azurin/pseudoazurin, the globin family, different families of proteases and cytochromes, including cytochromes *c'* and *b*-562, actinidin/papain, and lysozyme/α-lactalbumin. Turning to supersecondary structure prediction, we find that α/β/α fragments possess sufficient specificity to identify their own and related sequences. By threading a β-hairpin through a sequence, it is possible to predict the location of such hairpins and turns with remarkable fidelity. Thus, the method greatly extends existing techniques for the prediction of both global structural homology and local supersecondary structure.**

The native structure of a protein, as elucidated by x-ray crystallography or NMR spectroscopy, is a complex, three-dimensional arrangement of the amino acid chain. Despite the determination of almost 1000 protein structures (1), no one has been able to predict the protein's structure from its sequence alone. A simpler problem, whose solution is a necessary prerequisite for solving the full folding problem, has been formulated. Is it possible to find out which sequences are likely to fold into one of the known protein topologies (2)? Below, we present a method that not only elucidates which global structures are compatible with a given sequence (and vice versa) but also identifies elements of supersecondary structure.

The inverse protein-folding problem has been examined before by analyzing side chain packing in the presence of a rigid C^α backbone (3). This approach is computationally expensive and ignores backbone adjustments and the problem of false positives. In an elegant series of publications, the possibility of recognizing distant homologues on the basis of hydrophobicity, secondary structure preferences, and residue environment in the native structure was examined (4). While the latter is logically equivalent to the simplified version of the method presented here (the "frozen approximation"), the framework of the present approach facilitates extension to detect similarities between proteins having significantly changed residue environments. At the same time, the mutual equivalence of the topology fingerprint and the full three-dimensional structure make it possible to build protein models based on the predicted topology fingerprint. Finally, the specificity of interaction patterns without allowing for introduction of gaps was investigated (5, 6). The latter limitation, together with basing both approaches on C^α or C^β interactions, greatly limits its specificity.

THE MODEL

Every protein structure is analyzed as follows: each amino acid is classified as being buried or exposed to solvent by comparing the actual and the maximal possible exposed surface area for a given amino acid. Side chains having >70% of their surface area screened from the solvent are classified as buried. Next, all interacting pairs of residues are identified. Two side chains are said to interact if any pair of their heavy atoms is <5 Å apart. Triplets of interacting residues occur when all three amino acids satisfy the above criterion. Using this contact definition, there are no larger clusters of interacting residues.

To determine the empirical interaction free energies, a set of 56 nonrelated, high-accuracy protein structures (resolution better than 2.5 Å; *R*-factor <0.2) was chosen from the Brookhaven protein data base (1, 7). For each residue type, we compare the ratio of the buried positions to the expected number if the distribution were random; the negative natural logarithm of this ratio is the buried energy. The residue pair potential is constructed by comparing the number of observed pairs in contact to the expected number based on the product of the independent probabilities that the given pair is buried and in contact, corrected for residue size. Again, the negative natural logarithm of this ratio determines the pair energy. For the three-body terms, we calculate the number of occurrences, $n_{3,\text{actual}}$, of a three-body cluster in the data base and the expected number, $n_{3,\text{expected}}$, based on the product of the probabilities of the occurrence of the pairs. There are a total of 1330 distinct triplet combinations of residues that do not involve glycine (we assign glycine as the zero of energy). Obviously, for the majority of cases, the statistics are too poor to assess the validity of the three-body term. Thus, if $|n_{3,\text{actual}} - n_{3,\text{expected}}| < 10$, we set the three-body energy, E_3, equal to zero. There are 74 triplets of residues having nonzero E_3. For instance, the Cys, Cys, Cys cluster should be observed 377 times in the data base based on the buried/exposed and two-body interaction energy terms; instead there are only 90 such clusters. Thus, $E_3(\text{Cys, Cys, Cys}) = +1.43\ k_BT$. The full parameter set is available via anonymous ftp.†

The information about which residues are buried, together with the list of interacting pairs and triplets, forms the

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.

*To whom reprint requests should be addressed.

†The file Inverse is available via anonymous ftp from the pub directory of ftp.scripps.edu.

"topological fingerprint" of a protein. It can be used to calculate the energy of any protein structure:

$$E = \sum_i \Gamma_i^A E_1(A_i) + \sum_i \sum_{>j} C_{ij}^A E_2(A_i, A_j) + \sum_i \sum_{>j} \sum_{>k} C_{ij}^A C_{ik}^A C_{kj}^A E_3(A_i, A_j, A_k), \quad [1]$$

where i, j, and k are positions along the sequence in protein A; A_i, A_j, and A_k are the amino acids found at these positions; Γ_i^A is the buried/exposed classification of position i; and $C_{ij}^A = 1$ if there is a contact between positions i and j, and it is zero otherwise. Finally E_1, E_2, and E_3 are the one-, two-, and three-body contributions to the total energy.

The information about which residues along the sequence are buried or which groups of residues interact is taken to be independent of the particular amino acid occupying this position. The topological fingerprint provides the information that "position five is buried" and "residues at position five and ten interact." Thus, if one threads a different sequence (protein B) into the topology fingerprint, the energy is calculated according to Eq. 1, with Γ_i^A and C_{ij}^A obtained from protein A and the sequence from protein B. For proteins of different lengths, the structural fingerprint is slid along the sequence of the second protein, so that the numbering for the sequence is shifted in register and the lowest energy for each protein sequence is stored. To correct for amino acid composition, the zero of energy for each protein is taken as the energy for the randomized sequence.

RESULTS

Structural fingerprints of 86 proteins were used to scan the protein sequence data base Swiss-Prot (Release 15.0). In all cases, the fingerprint recognizes its own sequence as the one with the lowest energy. This is seen in Table 1, where the energy of the correct sequence, together with the energy and position of the highest scoring unrelated sequence, is shown. Energies from other, unrelated sequences are distributed according to a Gaussian distribution. Results for the set of 56 proteins used to derive the parameter set and the 30 additional proteins are qualitatively the same, so the results are definitely not an artifact of memorization of the data base.

To address the question of the origin of the high specificity of the topological fingerprint, the same experiment was performed using only buried/exposed and two-body interaction energies. The results for the buried/exposed energy are displayed in Table 1. Each energy contribution is in most cases sufficient to uniquely identify the correct sequence. The burial energy alone is wrong once, while the use of the two-body energy alone is incorrect in six cases. The energy difference between the best (correct) and the highest scoring unrelated sequence ranges from 270 kT to 8.7 kT, with the mean equal to 110 kT (for the burial energy alone, this difference decreases to 33 kT). Proteins that bind large ligands or prosthetic groups (which at present are ignored) have significantly smaller stability. Most failures of the two-body interaction scale can also be rationalized in this way. In most cases, the buried/exposed term is a sufficient indicator of the native sequence; other energy contributions enhance this specificity and correct for possible errors.

The ability to recognize correct sequences is not unique to the parameter set used here. The same calculations were performed by using several alternative sets of parameters including the two-body interaction scale derived from the same set of 56 proteins and other energy scales adopted from the literature (9, 10) with qualitatively identical results. This result should be compared to other approaches (5, 6), which in several cases have failed to recognize the correct sequence even among the sequences corresponding to the proteins in the structure library. Since they employ information about C^α or C^β positions with the corresponding set of effective interactions, it is possible that the side chain interaction patterns make the fingerprint so specific.

The very large energy separation between the energy of a protein sequence in its own fingerprint and the energy of the next, nonidentical sequence supports the notion of strong specificity of interaction pattern. It also illustrates the total lack of predictive power of this approach. For instance, when the energies of the plastocyanin sequence against fingerprints in the small structural data base are calculated, two other proteins with similar structures (pseudoazurin and azurin) have scores very close to random (Fig. 1). This problem is further illustrated by a detailed study of Table 1, where the energy and position of the best spurious match are shown. Only very close homologs are recognized, and in some cases, even formally identical sequences are missed, due to errors in the data base. Thus, the simple mixing and matching of sequences and structures, while useful for dismissing marginal interaction scales, are in and of themselves not much use for extending structure homology methods beyond one-dimensional sequence analysis techniques.

Two ways to extend the interaction fingerprint method to address these shortcomings are discussed below.

Introduction of Gaps. It is known that even closely related protein sequences differ not only by amino acid substitutions but also by relative shifts of sequence fragments, which can be described by introducing gaps/deletions into the alignment. Introducing gaps into Eq. 1 is complicated by the fact that the knowledge of the complete alignment is necessary to correctly include the two- and three-body energy terms. However, it is possible to introduce the following approximation:

$$E = \sum_i \Gamma_i^A E_1(B_i) + \sum_i \sum_{>j} C_{ij}^A E_2(B_i, A_j) + \sum_i \sum_{>j} \sum_{>k} C_{ij}^A C_{ik}^A C_{kj}^A E_3(B_i, A_j, A_k), \quad [2]$$

with the notation as before, but now B_i is the identity of an amino acid at position i in the sequence that is being investigated. We call this the frozen approximation, since the partners of residue type B_i contributing to the pair interactions (A_j) and triplets (A_j, A_k) are taken from the original protein, with the rationalization that the interaction environment might not drastically change between equivalent positions in related proteins. Now, the energy of every position does not depend on the alignment elsewhere in the sequence. Using standard dynamic programming algorithms, it is possible to find the optimal alignment between the sequence and the topological fingerprint by allowing for gaps/deletions, which are penalized by gap introduction/extension penalties.

Using this approximation, sequences of all proteins in our structural data base have been aligned with all fingerprints; the results are illustrated in the last three columns of Table 1. As before, in each case the match between correct sequence and structure always has the lowest energy. However, now the next best scores describe the structural similarity between proteins, even when the sequence similarity is weak or altogether absent. The algorithm correctly identifies similarities between copper binding proteins, the globins, different families of proteases and cytochromes, actinidin/papain, and lysozyme/α-lactalbumin. Furthermore, in some cases, even similar topological class is recognized (the copper binding protein and immunoglobulin folds; not shown) (11). The essential difference with the previous method that prohibits gaps is clearly illustrated in the *Inset* to Fig. 1, where both the azurin and pseudoazurin structures are clearly recognized by the plastocyanin sequence. The same algorithm as applied to a large sequence data base was able to identify all known

Table 1. Compilation of results of matching sequences to structures

PDB code	Alignment with no gaps						Alignment with gaps, full energy		
	Full energy			Burial energy					
	Native structure	Best spurious match		Native structure	Best spurious match		Native structure	Second best score	
		Energy	Position		Energy	Position		Energy	Protein name
1bp2	−74.8	−43.2	10	−16.6	−5.2	9	−24.6	0.0	1alc
1alc	−89.2	−37.7	13	−16.9	−5.4	13	−26.8	−10.6	1lz1
1ccr	−100.2	−39.9	31	−33.1	−5.6	31	−36.5	−10.2	3c2c
1cse	−262.1	−33.3	9	−58.7	−5.8	3	−87.9	−3.9	3est
1eco	−125.2	−31.8	6	−38.8	−7.4	3	−55.3	−13.1	1mba
1fd2	−60.2	−44.9	5	−18.9	−6.3	5	−11.2	−7.4	2lhb
1fx1	−145.4	−47.7	2	−47.2	−5.3	2	−61.2	−19.5	4fxn
1gcr	−154.8	−45.6	20	−38.8	−5.2	8	−69.8	−7.7	1paz
1gd1	−265.5	−57.1	4	−48.5	−5.7	4	−91.3	−6.0	6ldh
1hmq	−86.0	−53.7	7	−19.0	−5.8	5	−33.0	−4.0	3rp2
1hne	−215.6	−115.2	3	−51.5	−4.3	3	−79.9	−39.6	3rp2
1l01	−129.4	−46.0	3	−34.6	−5.4	3	−51.3	−6.2	3tln
1lz1	−119.0	−52.7	10	−33.8	−4.5	9	−46.8	−22.7	1alc
1mba	−137.4	−44.8	6	−37.5	−7.2	4	−56.4	−18.0	1eco
1mbd	−155.4	−40.1	66	−44.7	−6.7	66	−58.9	−10.3	1mba
1p01	−197.8	−53.8	3	−48.4	−4.2	3	−72.8	−8.3	2sga
1paz	−113.7	−47.6	5	−23.4	−6.3	4	−45.5	−7.3	1pcy
1pcy	−110.9	−43.9	17	−28.9	−5.5	15	−45.1	−10.9	1paz
1sgt	−220.1	−77.2	2	−46.9	−5.0	2	−83.4	−35.3	4ptp
2act	−188.5	−65.8	3	−58.9	−5.0	4	−62.3	−30.6	9pap
2app	−303.4	−65.3	3	−73.8	−6.8	3	−106.2	−37.7	3apr
2aza	−143.4	−49.1	10	−32.5	−5.8	10	−52.7	−3.4	1pcy
2ca2	−226.9	−68.2	12	−55.1	−5.9	12	−83.8	−6.1	3grs
2ccy	−94.7	−42.3	2	−27.8	−5.3	2	−36.8	−3.4	1l01
2cdv	−68.3	−31.8	3	−13.7	−8.1	3	−18.9	−3.8	3rnt
2cna	−177.4	−56.8	3	−34.5	−4.7	3	−62.3	−5.2	2app
2cpp	−283.7	−119.3	2	−75.7	−10.0	2	−109.0	−8.7	1cse
2cyp	−216.5	−67.2	2	−42.4	−6.7	2	−77.7	−8.4	1fx1
2fb4	−188.7	−62.0	27	−53.5	−5.8	11	−74.1	−8.2	1gcr
2lh2	−106.8	−37.2	4	−29.9	−6.8	4	−37.9	−6.6	1eco
2lhb	−100.4	−48.9	9	−29.5	−6.2	9	−31.9	−18.1	1mbd
2prk	−225.0	−69.6	2	−55.3	−6.2	2	−76.3	−24.6	1cse
2rhe	−88.4	−52.5	8	−19.2	−5.2	3	−25.9	−5.2	2fb4
2sga	−181.7	−53.5	2	−46.5	−4.1	2	−53.4	−14.8	2alp
2sns	−87.6	−43.3	3	−24.0	−5.9	3	−31.4	−5.5	1alc
2sod	−172.5	−42.1	13	−38.5	−4.7	7	−55.4	−3.8	2ca2
2wrp	−47.2	−38.5	2	−11.0	−12.4	1	−11.7	−6.9	2lh2
3apr	−333.4	−63.4	2	−78.1	−4.9	2	−120.1	−42.0	2app
3c2c	−81.0	−41.8	2	−16.8	−4.4	2	−27.8	−14.2	1ccr
3est	−199.2	−55.3	7	−40.5	−6.2	2	−76.1	−35.3	3tpi
3grs	−379.4	−131.3	2	−83.1	−5.6	5	−132.2	−6.6	1hmq
3rnt	−88.3	−52.4	5	−26.3	−5.6	7	−36.4	−3.5	1paz
3rp2	−189.8	−62.7	9	−45.3	−5.3	4	−68.4	−43.5	3est
3tln	−253.7	−82.0	4	−44.9	−5.2	4	−78.8	−7.0	1mbd
3tpi	−214.3	−50.3	9	−45.0	−4.9	7	−76.6	−34.8	3est
4fxn	−148.3	−53.1	3	−41.8	−5.3	4	−58.2	−11.0	1fx1
4hhB	−136.7	−48.9	195	−36.6	−6.7	186	−50.9	−6.8	4hhb
4hhb	−91.7	−33.9	208	−20.9	−5.5	157	−28.0	−11.6	4hhB
4i1b	−93.0	−52.5	2	−22.2	−4.8	3	−30.4	−8.2	3est
5cpa	−237.7	−74.4	4	−44.0	−7.8	4	−85.3	−6.8	3grs
5tnc	−129.9	−53.3	7	−37.5	−9.7	7	−50.9	−7.1	1mba
6ldh	−246.3	−57.9	7	−46.4	−7.3	3	−75.4	−6.3	4hhb
7rsa	−60.3	−39.6	39	−19.6	−9.3	28	−23.8	−3.9	2alp
8dfr	−146.3	−56.4	9	−38.1	−5.5	9	−55.5	−4.9	6tmn
9pap	−185.4	−48.3	5	−42.6	−7.0	2	−61.6	−29.5	2act
9wga	−181.6	−19.4	6	−43.0	−4.0	7	−62.8	−3.3	2ca2
1cd4	−139.5	−42.2	5	−39.8	−5.6	5	−44.9	−8.3	2ltn
1cla	−101.9	−67.7	5	−8.3	−6.8	5	−14.0	−6.0	5cpv
1cms	−240.8	−53.8	2	−59.6	−6.0	3	−64.8	−25.1	3apr
1csc	−314.9	−99.2	2	−62.8	−7.4	2	−71.4	−6.4	1eco
1fcb	−340.5	−135.8	2	−67.8	−9.0	3	−87.5	−10.1	3icd
1gp1	−141.1	−62.1	7	−24.4	−4.9	7	−33.9	−6.3	3dfr

Table 1. (*continued*)

PDB code	Alignment with no gaps: Full energy: Native structure	Full energy: Best spurious match: Energy	Full energy: Best spurious match: Position	Burial energy: Native structure	Burial energy: Best spurious match: Energy	Burial energy: Best spurious match: Position	Alignment with gaps, full energy: Native structure	Second best score: Energy	Second best score: Protein name
1pfk	−289.8	−62.4	5	−70.4	−4.1	3	−77.0	−8.2	1fcb
1phh	−274.1	−64.4	4	−75.7	−5.1	4	−74.5	−5.8	8dfr
1rhd	−214.1	−60.3	3	−51.5	−7.4	3	−61.4	−7.3	2cyp
2gbp	−246.9	−61.9	2	−56.5	−7.4	2	−74.1	−7.0	1cse
2liv	−278.1	−78.6	5	−90.4	−7.0	5	−82.9	−8.1	3grs
2ltn	−112.7	−49.4	5	−19.8	−7.2	4	−20.8	−5.7	1paz
2rsp	−73.8	−45.4	2	−19.0	−7.2	2	−18.9	−5.5	9pap
2ypi	−200.0	−61.8	3	−50.7	−5.2	3	−53.3	−6.4	4tsl
3adk	−174.1	−38.5	7	−50.9	−4.1	6	−54.6	−8.6	5cpv
3b5c	−58.7	−32.7	5	−19.8	−10.4	9	−19.4	−5.3	1fcb
3bcl	−203.8	−29.6	2	−25.8	−8.9	2	−39.7	−9.3	2cyp
3blm	−148.7	−71.2	2	−38.8	−7.1	2	−40.8	−8.7	1fcb
3dfr	−114.3	−51.6	3	−32.7	−4.8	3	−31.9	−14.6	8dfr
3gap	−141.1	−58.9	4	−35.5	−8.2	4	−32.0	−6.3	1csc
3icd	−319.1	−67.7	2	−59.4	−5.8	2	−71.7	−7.6	5cpv
4dfr	−98.1	−44.3	5	−25.9	−9.3	3	−22.9	−6.7	8dfr
4er4	−294.9	−68.7	3	−73.3	−7.6	3	−87.1	−56.9	2app
4hvp	−86.8	−48.4	2	−21.4	−5.0	2	−23.1	−6.8	2sod
4mdh	−258.0	−57.8	4	−58.1	−5.2	4	−71.2	−8.0	3tpi
4tsl	−220.1	−65.3	4	−49.0	−8.3	2	−56.8	−5.8	2ca2
5cpv	−98.2	−43.2	25	−21.6	−5.9	14	−28.7	−5.2	4fxn
5xia	−311.7	−62.8	2	−57.5	−4.9	2	−80.5	−6.3	1cd4
7cat	−251.9	−85.4	6	−27.4	−8.7	5	−55.3	−5.7	4mdh
8adh	−309.0	−62.9	12	−74.5	−7.6	9	−66.1	−5.0	5tnc

Eighty-six proteins in the structural data base, identified with their PDB codes, were aligned with every sequence in the Swiss-Prot (Release 15) sequence data base with no gaps allowed. The energy of the protein's own sequence and the energy and position of the lowest energy spurious match are presented for the full and burial energy. After allowing for gaps in the alignment, the same calculation was performed, but only sequences in the structural data base were used (last three columns). Energy values are in kT.

proteins with the plastocyanin fold; this and several other examples are presented elsewhere (11).

It is possible to go beyond the static approximation by updating each residue environment according to the alignment $N_{A \to B}$ obtained in the first stage. Now, the partners appearing in Eq. 2 are replaced by the new partners (B_k) resulting from the first stage of the alignment. The energy in the next set of alignments is calculated according to

$$E = \sum_i \Gamma_i^A E_1(B_i) + \sum_i \sum_{>j} C_{ij}^A E_2(B_i, B_{N_{A->B}(j)}) + \sum_i \sum_{>j} \sum_{>k} C_{ij}^A C_{ik}^A C_{kj}^A E_3(B_i, B_{N_{A->B}(j)}, B_{N_{A->B}(k)}). \quad [3]$$

This procedure is iterated until the alignment converges. The results are illustrated in Fig. 2, where the sequence of cytochrome *b*-562 is checked against the library of fingerprints. There is one similar structure (12) [cytochrome *c'* (2ccy)] in the fingerprint library, but as shown in Fig. 2, it was missed by the alignment with the frozen approximation. In the second pass, which updates the residue environment, this similarity is recognized.

Fragment Matching. Here, fingerprints of smaller fragments of the structure are used: (*i*) an $\alpha/\beta/\alpha$ fragment from leucine binding protein (residues 121–168 from 21bp), (*ii*) a β-hairpin from proteinase B (residues 166–180 from 3sgb), and (*iii*) an α-hairpin from leghemoglobin (residues 105–146 from 1lh2).

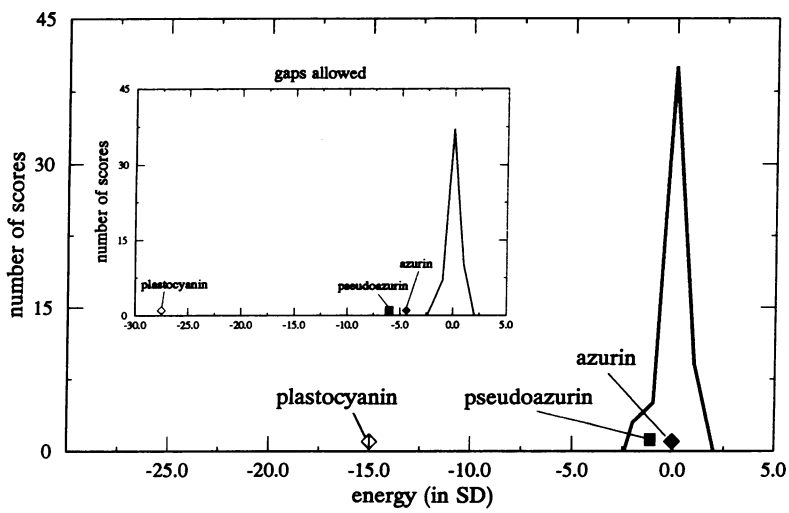

FIG. 1. Distribution of energies of the plastocyanin (1pcy) sequence in different topological fingerprints without and with (*Inset*) allowing for gap introduction. Energies of the plastocyanin sequence in its own (◇), azurin (◆), and pseudoazurin (■) structures are shown on each plot. Energy units are in standard deviations (SD) of the Gaussian distribution.

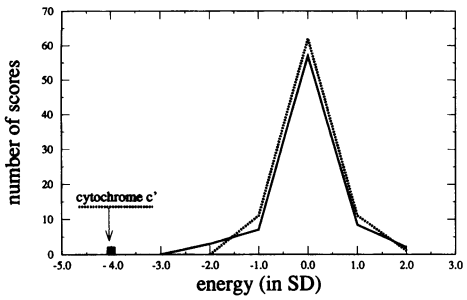

FIG. 2. Distribution of energies of cytochrome *b*-562 sequence in different topological fingerprints with a frozen residue environment (——) and after updating (- - -). Energy units are in standard deviations (SD) of the Gaussian distribution.

In each case, the structural fingerprint of the fragment was built by using the buried/exposed pattern of the fragment in the protein of interest as well as that portion of the contact map associated with those residues in the fragment. These fingerprints were used to scan the sequence data base of nonhomologous proteins with well-refined structures. The energies for all sequence fragments were stored and rank ordered according to their energy.

The $\alpha/\beta/\alpha$ fragment had all the characteristics of the whole protein; it easily recognized its own sequence and sequences of similar fragments from other highly homologous proteins. After this group, separated by large energy gaps, were random scores. This result is consistent with experiments that indicate that $\alpha/\beta/\alpha$ fragments can be circularly permuted or added to the sequences of existing α/β barrels and retain their identity as autonomous folding units (13).

The α-hairpin had a marked preference for helical hairpins, with its own sequence on the top of the list. There were, however, both α/β- and β-hairpins with energies very close to the lowest energy.

The most interesting results have been obtained with the β-hairpin. β-Hairpins of different types constitute about 80% of the best scores, and in the top 25 scores, only two fragments did not have a β-hairpin structure (both were long fragments of β stands, with two glycine residues or Gly-Xaa-Pro in the middle). A detailed analysis of the energy profile along a single sequence proved more interesting. As illustrated in Fig. 3 for plastocyanin, all local minima in the energy profile can be identified, within shifts of ±1 residue, with either turns or bends. Similar results were obtained for other β proteins. In all cases, the template had no sequence similarity with hairpins identified by the fingerprint.

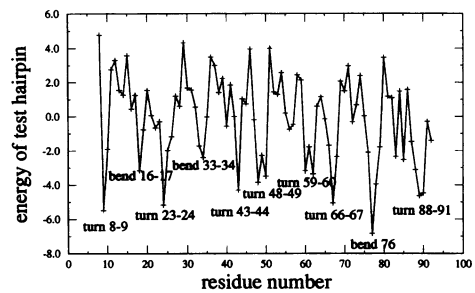

FIG. 3. Energy of fitting sequence fragments of plastocyanin into the structure of a β-hairpin fragment from proteinase B. The 10 lowest energy fragments are labeled by the closest structural fragment from the plastocyanin (1pcy) structure.

CONCLUSIONS

It is possible to use a protein's structure, as described by its topology fingerprint, which contains information about buried/exposed positions and pairs and triplets of interacting residues, to uniquely identify the correct sequence that folds to it. This is done by estimating the energy of all sequences in the sequence data base in the topology (defined by a structural fingerprint) of the query structure and choosing the sequence with the lowest energy. The buried/exposed pattern is most specific; two- and three-body interactions enhance the specificity. Taken together, these terms add up to more than a 100 *kT* mean energy difference between the correct and closest spurious match in searches using the large sequence data base. Surprisingly, this result depends very little on the interaction scale used, with several parameterizations producing quantitatively identical results.

However, without allowing for gaps/insertions into the fingerprint, only very closely related sequences can be identified. Gaps can be allowed by the further approximation of freezing the identity of residues interacting with the residues of the analyzed sequence to that of the query structure. This procedure easily matches the structural fingerprint with correct sequences, even when large sequence data bases are searched; moreover, proteins having similar topology, but no sequence similarity, can be identified.

After initial alignment, the interactions can be thawed to reflect the actual protein environment, thereby allowing for identification of more remote sequence similarities.

On the other hand, shorter structural fragments such as α- or β-hairpins can be successfully used to identify sequence fragments that are likely to adopt identical or very similar structure, but are too small to correctly identify their own sequence. As shown in the plastocyanin example, all turns and bends can be correctly predicted. Similar results can be obtained for helical proteins probed with helical hairpins.

Larger fragments, such as the $\alpha/\beta/\alpha$ fragment from the leucine binding protein, already exhibit some characteristics of the whole protein, by being able to identify their own sequence.

Thus, we conclude that it is possible to build a library of entire protein and protein fragment fingerprints to cover all known folds and types and combinations of secondary and supersecondary structure elements. New protein sequences can be studied by identifying structures ranging from supersecondary elements to domains to the entire fold that the new protein is likely to adopt.

This research was supported in part by Grant GM-37408 of the Division of General Medical Sciences, National Institutes of Health.

1. Bernstein, F. C. (1977) *J. Mol. Biol.* **112,** 535–542.
2. Pabo, C. (1983) *Nature (London)* **301,** 200.
3. Ponder, J. W. & Richards, F. M. (1987) *J. Mol. Biol.* **193,** 775–791.
4. Bowie, J. U., Luethy, R. & Eisenberg, D. (1991) *Science* **253,** 164–170.
5. Crippen, G. M. (1991) *Biochemistry* **30,** 4232–4237.
6. Casari, G. & Sippl, M. J. (1992) *J. Mol. Biol.* **224,** 725–732.
7. Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992) *Prot. Science* **1,** 409–417.
8. Luger, K., Szadkowski, H. & Kirschner, K. (1990) *J. Prot. Engineer.* **3,** 249–258.
9. Miyazawa, S. & Jernigan, R. L. (1985) *Macromolecules* **18,** 534–552.
10. Hinds, D. A. & Levitt, M. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 2536–2540.
11. Godzik, A., Skolnick, J. & Kolinski, A. (1992) *J. Mol. Biol.* **277,** 227–238.
12. Weber, P. C., Salemme, F. R., Matthews, F. S. & Bethge, P. H. (1981) *J. Biol. Chem.* **256,** 7702–7704.