Article No. jmbi.1999.3212 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (1999) 293, 1183–1193

JMB

De Novo Protein Design. II. Plasticity in Sequence Space

Patrice Koehl* and Michael Levitt*

Department of Structural Biology, Fairchild Building Stanford University, Stanford CA, 94305, USA

**Corresponding authors*

It is generally accepted that many different protein sequences have similar folded structures, and that there is a relatively high probability that a new sequence possesses a previously observed fold. An indirect consequence of this is that protein design should define the sequence space accessible to a given structure, rather than providing a single optimized sequence. We have recently developed a new approach for protein sequence design, which optimizes the complete sequence of a protein based on the knowledge of its backbone structure, its amino acid composition and a physical energy function including van der Waals interactions, electrostatics, and environment free energy. The specificity of the designed sequence for its template backbone is imposed by keeping the amino acid composition fixed. Here, we show that our procedure converges in sequence space, albeit not to the native sequence of the protein. We observe that while polar residues are well conserved in our designed sequences, non-polar amino acids at the surface of a protein are often replaced by polar residues. The designed sequences provide a multiple alignment of sequences that all adopt the same three-dimensional fold. This alignment is used to derive a profile matrix for chicken triose phosphate isomerase, TIM. The matrix is found to recognize significantly the native sequence for TIM, as well as closely related sequences. Possible application of this approach to protein fold recognition is discussed.

Keywords: protein design; random energy model; sequence space; Monte Carlo; fold recognition

Introduction

It has been hypothesized that the total number of different protein folds is finite, and roughly of the order of 1000 (Chothia, 1992; Orengo *et al*., 1994; Wang, 1998). Once examples of every fold are known, protein structure prediction would reduce to the inverse protein folding problem, which consists in identifying which sequences are compatible with a given fold (Drexler, 1981). This alternative view of structure determination is the basis of the new field of structural genomics, which aims to deliver structural information about most genome-derived protein sequences. While it is not feasible to determine experimentally the structure of every protein, useful models can be obtained by fold recognition and comparative modeling, provided there is a comprehensive library of folds (Kim, 1998). Structural genomics is currently focusing on the construction of such a library, and a figure of 10,000 to 100,000 representative proteins has been proposed (Sali, 1998). With such a library, the next step is to devise methods which will build a three-dimensional model for any unknown protein. All such methods rely on fold recognition (Bryant, 1996; Lemer *et al*., 1995; Miller *et al*., 1996; Mirny & Shakhnovich, 1998) and while performance is improving as demonstrated at CASP3 (see Koehl & Levitt, 1999a), there is room for improvement.

Intrinsic to fold recognition is finding sequences compatible with a given protein fold. This was first done by Ponder & Richards (1987), who assumed that residues in the interior of a protein determine its fold. They tested systematically combinations of side-chains fitting in the cores of small proteins, using criteria of steric overlaps, hydrogen bonding and packing density. The set of successful sequences thus selected defined the ''tertiary template'' of the structure. The number of residues that can be included in their combinatorial search is limited for computational reasons to less than ten.

Abbreviation used: NLM, non-linear mapping.

E-mail addresses of the corresponding authors: koehl@hyper.stanford.edu and michael.levitt@stanford.edu

0022-2836/99/451183–11 $30.00/0

Recently, explorations of sequence space have been developed as part of more general protein sequence design. These approaches optimize a single sequence for a given backbone either by Monte Carlo methods or by using the dead-end elimination theory. For example, successful designs of protein cores (Dahiyat & Mayo, 1997b) and of a small protein (Dahiyat & Mayo, 1997a) have been reported. Extension to larger proteins and to systematic search in sequence space is difficult due to excessive computer time requirements.

As a systematic search in sequence space is not yet feasible, current methods for fold recognition limit the search to available sequence databases. The focus is set on improving the ability to detect if a particular sequence can adopt a given structure. Two major approaches have been derived: (a) Each known protein structure is represented as a contact map containing position-dependent residue-residue contact preferences. Aligning a sequence using these contact maps requires a double dynamic programming algorithm (Taylor & Orengo, 1989). This approach has been successfully applied by Jones *et al*. (1992) for fold recognition and is referred to as the ''threading'' algorithm. Its drawback, however, is that it is computationally slow. Working with contact maps is however promising, as recently shown by Domany and colleagues (Mirny & Domany, 1996; Vendruscolo *et al*., 1999). (b) Protein structure information is reduced into one dimension, yielding so-called 3D-1D profiles (Bowie *et al*., 1991). In this approach, a position and structure-dependent scoring table is built, which contains as many rows as residues in the structure, and a column for each of the 20 amino acids. These profiles are analogous to profiles derived from multiple alignments and commonly used in sequence database searches (Gribskov *et al*., 1987). Aligning a sequence on such profile matrices uses the same very efficient dynamic programming methods developed for pairwise sequence comparisons (Smith & Waterman, 1981).

In previous applications of profile matrices, scores are usually calculated from the residue environment in the native protein structure (the so-called frozen approximation). One possible caveat of this approach is that it is based on the hypothesis that residue environments are conserved in proteins with similar folds. This hypothesis has been questioned by Russell & Barton (1994), who suggest that only core secondary structures are conserved, and more recently by Rodionov & Blundell (1998), who showed that only the environment of polar residues in the core of proteins is conserved. An attempt to circumvent the frozen approximation was proposed by Delarue & Koehl (1997), who computed a substitution or profile matrix based on the backbone of the protein. The matrix was derived by self-consistent optimization of the mean field generated in a chimeric protein, obtained by attaching to each CA atom of the known backbone structure multiple copies of the different side-chains, corresponding to all 20 natural amino acid residues. A knowledge-based potential was used to derive interaction energies between residues. The optimized structure-dependent profile matrix was shown to recover sequence information. This signal, however, was found to be weak compared to those obtained with an evolution derived substitution matrix.

We have recently defined a new procedure for the design of a full protein sequence (Koehl & Levitt, 1999b), based on an all-atom representation of the protein. In this procedure, the ''optimal'' sequence is defined as the sequence with the lowest energy when threaded on the template fold. Specificity, the compatibility of the designed sequence for the template fold and the incompatibility for competing folds, is ensured by maintaining the amino acid composition constant, in accordance with the random energy model (Shakhnovich & Gutin, 1993a,b; Pande *et al*., 1997). In the first paper of this series, we showed that this procedure is able to design specific sequences (Koehl & Levitt, 1999b). Here, we study how much sequence information can be derived from the protein backbone and show how this information can be used for fold recognition.

Results

Protein sequence optimization converges

Protein design *per se* is only concerned with finding one sequence specifically compatible for its template structure. In a broader sense, one needs to characterize the set of all protein sequences compatible with a given fold. Such an information would be of great practical value to the inverse protein folding problem (Pabo, 1983) as it would partition the sequence space with respect to the set of all possible protein folds.

We have chosen the small, highly stable protein 1CTF as a test molecule in our initial analysis. 1CTF, the C-terminal domain of the L7/L12 ribosomal protein of *Escherichia coli* is a $\alpha + \beta$ protein of 68 residues solved to 1.7 Å resolution by X-ray crystallography (Leijonmarck & Liljas, 1987). The 1CTF fold is unique in the present PDB database (Bernstein *et al*., 1977), though a structural similarity between one of its fragment and a fragment of the ovomucoid domain has been detected (Laurents *et al*., 1994). We carried out ten independent protein design optimizations over the whole sequence of 1CTF (Figure 1 and Table 1). All designed sequences are very similar to the native sequence of 1CTF, with a maximum of 44% sequence identity. In comparison, random permutation of the native sequence of 1CTF yields sequences which on average have 12(±5)% sequence identity to the native. THREADER (Jones *et al*., 1992) identifies the designed sequences as being specific to the native fold of 1CTF, with Z-score values greater than 10 (Table 1).

A multiple sequence alignment of all designed sequences and the native sequence results in an

```
     1         10        20        30        40        50        60
D1   DLKLKLVEAGGAKGSVIDKVRAIAGEDESAAFEKVVEAKKPLLTGKDEEVAKNLAEVLKAAGAAAEVA
D2   DKAVKVLEAGGNVESVISKLIELAGLPEDVAKEKAKKAAEAFLRGKDDGAAETVKKVLLEAGAAAEVA
D3   DRKVKLAEAGGAVESVIAKVIELEGLPEDAAKEKVKKAGEFLKTGVADDVALSLAKELKNAGAAAEVA
D4   DLVVKLLEAGGAVESVIEKFTAIAGENGDAAKEKVKEAAKPLLRGKDEDVAKSLKEVLKAAGAAAEVA
D5   DLKVKVEEAADARGSVIDGVIKLAGLPEDAAKSVVKAAKEALKTGEAEGEAKNVKELLLKAGLRPEVA
D6   DLKVKVEAAEGAVGAGIAKVIALEGLDESAAKEVVFKAKEALKTGVDEDAAKSLAKELKNFGAAAEVA
D7   DLKVKLEEAGGAIEAVIDKVRAFEGLSSDVAKELVKKAKEALKTGVAEDVAKNLAEGLKAAGAAAEVL
D8   DLAVKLAEAGGATDSVVAKFIEIEGKPEDAAKELVKKAGEKLKRGVAEDVAKSLAEVLKNAGAAPEVA
D9   DLKLKLVEAGGAKDSVIAKVIELAGRPEDEAKNVVKAVKEALFTGVDEGAAKSLKEELKAAGAAAEVA
D10  DRVLKLLAAEGAKGAVVEKFIEIEGKDTSAAKELVKKVAEALKVGEDDGNAKSLAEVLKAAGAAPEVA

CTF  EFDVILKAAGANKVAVIKAVRGATGLGLKEAKDLVESAPAALKEGVSKDDAEALKKALEEAGAEVEVK

Str   EEEEEEE    HHHHHHHHHHHH   HHHHHHHHH   EEEEEEE HHHHHHHHHHHHHH  EEEE
Acc  ebbbbbiibeeiieibbebbieibeiieeibiibbiebebibieibeeeebiebbeibiebibebibe
```

Figure 1. Multiple sequence alignment of ten designed sequences for 1CTF, illustrated by BOXSHADE (http://www.isrec.isb-sib.ch/software/BOX_form.html). At each position of the alignment, six or more identical residues appear on a black background, while six or more similar residue appear on a gray background (similarity is based on hydrophobicity). The positions of the secondary structures and the exposure of each residue as computed by DSSP (Kabsch & Sander, 1983) are given as Str and Acc, respectively. For secondary structure, H corresponds to helix and E to strands. For solvent exposure, residues with less than 30 % of accessible surface area (ASA) are considered buried (b), while residues with more an ASA greater than 50 % are considered exposed (e). An accessibility value between 30 % and 50 % is considered intermediate (i).

alignment without gaps (Figure 1). Interestingly, the designed sequences show higher levels of similarity among themselves than with the native sequence. This is described in detail in the legend to Figure 2. The average sequence identity of the $9 \times 10/2 = 45$ sequence pairs obtained from the ten designed sequences vary from 12.4 % for the ten initial random sequences to 58.9 % for the ten optimized sequences (Figure 2(a)). The averaged sequence identity of the optimized sequences to the native sequence is 36 %.

Evolution of these ten sequences cannot be directly visualized in the high dimensional sequence space. Therefore, we project sequence space onto a plane (Figure 2(b)), using non-linear mapping (NLM; see Materials and Methods). The trajectories corresponding to the ten optimizations show how all ten sequences converge in the same region, R, in sequence space, which does not include the native sequence for 1CTF. Moreover, optimization starting from the native sequence for 1CTF yields a sequence which is 34 % identical with the native sequence that lies in the same region of sequence space as all other designed sequences (Figure 2(b)).

Amino acid plasticity

Diversity is also observed amongst the designed sequences. This sequence variability can be used to characterize the tolerance of each type of amino acid for substitution. For this part of the work a set of ten different proteins was considered (Table 2) to allow better statistical sampling. Ten sequences were designed for each protein, yielding a set of ten multiple sequence alignments, each containing 11 sequences (one native, plus ten designed). For each type of amino acid, j, in the test proteins, substitutions between the native sequence and any of the designed sequence are recorded and used to compute the probability, $P_{\mathrm{des}}(j,j)$, that the same amino acid is recovered by the sequence design procedure at the same position. Our protein design strategy enforces specificity implicitly by maintain-

Table 1. Characteristics of ten designed sequences for 1CTF

Sequence	Sequence identity with 1CTF (%)	Sequence *P*-value[a]	THREADER *Z*-score
Native	100.00	NA	26.1
D1	32.35	−4.6	11.6
D2	35.29	−5.1	11.5
D3	39.71	−6.2	13.6
D4	33.82	−5.3	11.8
D5	32.35	−5.5	12.1
D6	38.24	−6.4	13.2
D7	44.12	−8.8	15.3
D8	35.29	−6.2	12.6
D9	36.76	−6.5	13.6
D10	33.82	−5.0	10.4
Average (std)	36.20 (3.7)	−6.0 (1.2)	12.6 (1.4)

[a] Log_{10} of the estimated significance P of the sequence match between the designed sequence and the native sequence, based on the model proposed by Levitt & Gerstein (1998).

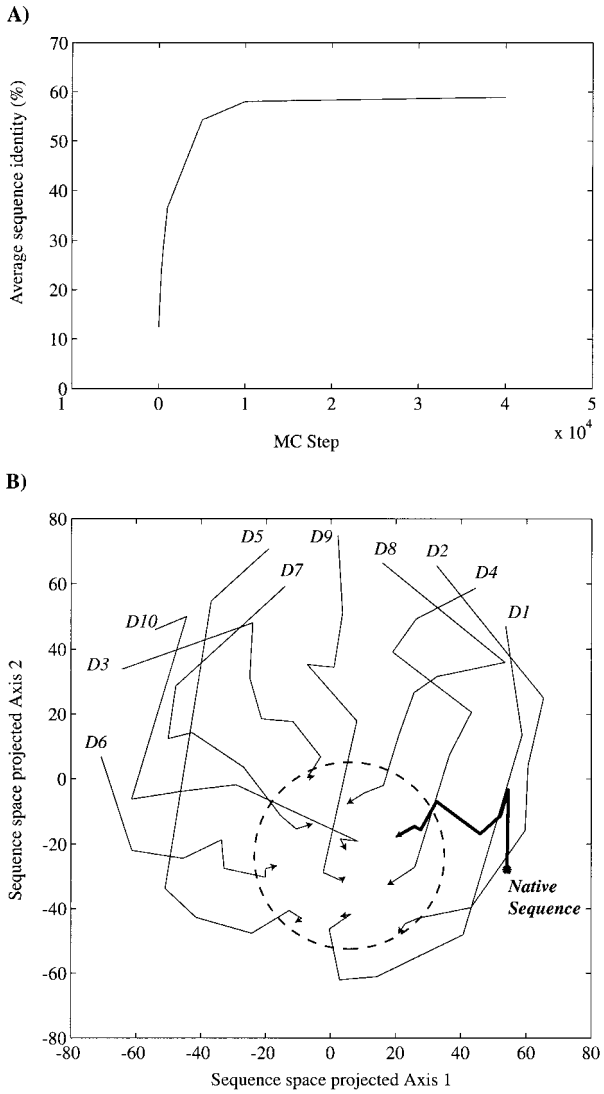

Figure 2. (a) Mean of the distribution of sequence identity among the ten designed sequences for 1CTF, *versus* the Monte Carlo step number. (b) Trajectory in sequence space of 11 Monte Carlo sequence design runs on 1CTF. The arrow indicates the end of the trajectory. Each sequence optimization started from a random sequence with the same amino acid composition than 1CTF, except for the trajectory marked with a star (*), which started from the native sequence. The designed sequences are given in Figure 1. The dotted circle is a visual representation of the zone of convergence. This two-dimensional representation of sequence space is created using the non-linear mapping described in Materials and Methods.

ing the amino acid composition fixed, in accord with the random energy model. To define a baseline, substitution probabilities, $P_{\mathrm{rand}}(j,j)$, were derived from ten random sequences with the native amino acid composition for each protein. A comparison of the P_{des} and P_{rand} values is shown in Figure 3(a). Based on our classification (see Figure 3), there are a total of 382 non-polar and 525 polar residues in the ten proteins of our data set. The probabilities to pick an hydrophobic or hydrophilic residue by chance are therefore 42 % and 58 %, respectively. These numbers are very close to the numerical values we found for the conservation of non-polar and polar residues in the random sequences (43 % and 59 %, respectively). The protein design procedure increases both figures significantly: 61 % of the hydrophobic residues remain non-polar, while 73 % of the polar residues remain hydrophilic.

At the level of individual amino acid residues, two residues stand out, namely Cys and Gly which are both highly conserved in the designed sequences (47 % and 60 %). Most of the cysteine residues considered here belong to 9wga, which contains 34 cysteine residues involved in 17 disulfide bridges. Though the potential energy function we used for design does not contain an explicit term for the formation of these S-S bridges, at least half of the cysteine residues were conserved in all ten designed sequences for 9wga. Interestingly, 5PTI behaved differently, and none of its three disulfide bridges were conserved in any of the designed sequences. This suggests that the design procedure could be improved by the addition of a term for disulfide bridge stabilization to the energy function used for sequence selection.

The 100 glycine residues in our ten proteins divide into two sets: glycine residues with positive ϕ and ψ psi values (37), and the others (63). The first set of glycine residues are strongly conserved with 76 % conservation in the designed sequences, *versus* 50 % for the other set. This result, which is expected for steric reasons as other amino acids cannot accommodate this combination of (ϕ,ψ) values, confirms that our potential energy function does impart structure information to the sequence.

The bias toward the conservation of a residue j introduced by the design procedure can be expressed as the log of the ratio of the observed probabilities in the designed sequences to that in the random sequences (log-odds energy):

$$E(j,j) = \log\left[\frac{P_{\mathrm{des}}(j,j)}{P_{\mathrm{rand}}(j,j)}\right] \qquad (1)$$

The $E(j,j)$ values (Figure 3(b)) show that the design procedure conserves hydrophobic residues more than polar residues. This is striking at the individual amino acid level: most non-polar, hydrophobic residues have a E-score of 1.5, while for polar, hydrophilic residues the average is 0.5, with two residues (Asn and Lys) having negative E values, i.e. the non-conservative substitution of these residues in the designed sequences is more frequent than chance.

All non-polar residues are not buried, and conversely, there is a reasonable number of polar residues that are buried (Miller *et al.*, 1987). We classified the residues into two categories, buried

Table 2. Our database of protein structures

PDB code	Residues	Method	Resol. (Å)	Protein	Reference
1CTF	68	X-ray	1.7	C-terminal domain of L7/L12 ribosomal protein	Leijonmarck & Liljas (1987)
2CI2-I	65	X-ray	2.0	Chymotrypsin inhibitor 2 (barley seed)	McPhalen & James (1987)
2HSP	71	NMR	N/A	SH3 domain of human phosphoric diester hydrolase	Kohda *et al.* (1993)
4ICB	76	X-ray	1.6	Bovine calbindin D9K	Svensson *et al.* (1992)
1LMB-3	92	X-ray	1.8	DNA binding protein (lambda)	Beamer & Pabo (1992)
5MBN	153	X-ray	2.0	Sperm whale myoglobin	Takano (1977)
7PCY	98	X-ray	1.8	Green alga plastocyanin	Collyer *et al.* (1990)
5PTI	58	X-ray	1.0	Bovine pancreatic trypsin inhibitor	Wlodawer *et al.* (1984)
1PGB	56	X-ray	1.9	B1 domain (protein G)	Gallagher *et al.* (1994)
9WGA	171	X-ray	1.8	Wheat germ agglutinin	Wright (1990)

(with an accessibility of less than 30 %) and accessible (the others), and computed all $P_{des}(j,j)$ for both groups (Figure 4).

Positions in the proteins of our data-set occupied by polar residues in the native sequences mainly remain polar in the corresponding designed sequences (73 % on average). The conservation is better for exposed polar amino acids (78 %). It is interesting to notice that polar amino acids are significantly conserved in the core (62 %). Polar residues in the core are usually small, and stabilized by hydrogen bonds with the backbone of the protein. Although there is no explicit term for hydrogen bonds in our potential energy function, these polar interactions seem to be well conserved in our designed sequences. It is likely that the consideration of surface area of contacts in the environment free energy part of the potential is responsible for this effect.

Similarly, positions occupied by non-polar amino acid residues in the native sequences mainly remain non-polar: 63 % on average, and this figure becomes 82 % and 33 % for buried and exposed positions, respectively. The environment free energy included in the design energy function is based on the accessible surface area of each atom, and penalizes large accessibility for non-polar atoms. This explains the high level of replacement observed for exposed positions that are occupied by non-polar amino acid residues in the native sequences, and by polar residues in the designed sequences.

The cores of the proteins in our data set contain, on average, 60 % of non-polar residues and 40 % of polar residues, while the cores of the model proteins based on the designed sequences are 64 % non-polar, and 36 % polar. The surface of the proteins show the opposite trend: it is 71 % polar and 29 % non-polar for the native sequences, and 75 % polar and 25 % non-polar for the designed sequences. These differences between native sequences and designed sequences may explain why the native sequence of a protein is not found in the zone of convergence in sequence space of our design procedure (Figure 2).

A structure-dependent substitution matrix

The preceding section was mainly focused on the derivation and application of $P_{des}(j,j)$, i.e. the probability that an amino acid of type j at position i in the native sequence is found to be identical at the same position in a designed sequence. The same procedure can be used in fact to generate a full substitution matrix, that gives, $P_{des}(j,k)$, that is the probability of amino acid type k is replaced by amino acid type j at position i. The objectives are different however: while we were interested above in the behavior of individual amino acids regardless of the protein, the substitution matrix should be protein specific. In fact, this matrix reflects how much sequence information can be extracted from a protein fold. We chose the TIM fold and designed 13 independent sequences on the backbone of the chicken triose phosphate isomerase (PDB code 1TIM-A, 247 residues with X-ray resolution 2.5 Å; Banner *et al.*, 1975). Each optimization was performed over 40,000 Monte Carlo cycles, requiring 48 hours of CPU time (4.4 seconds per cycle) on a DEC alpha processor at 533 MHz. Characteristics of the designed sequences are given in Table 3. Three independent profile matrices were consequently derived for TIM: (a) The sequence of TIM was used to scan the SWISS-PROT database (Bairoch & Apweiler, 1999; Bairoch & Böckman, 1991) (dated 98/05/05), and all triose phosphate isomerase sequences were selected, except those for which a structure has been determined and deposited in the PDB database. The PROFILEMAKE program (Gribskov *et al.*, 1990) as implemented in the UWGCG package (Devereux *et al.*, 1984) was used on these selected sequences to generate a profile from them, referred to as Prof_SEL. (b) The 13 designed sequences were used with PROFILEMAKE to generate the structure specific profile, Prof_DES. (c) A set of 13 random sequences with the amino acid composition of 1TIM-A were used by PROFILEMAKE to generate a ''random'' substitution table, Prof_INI (since the amino acid composition is fixed in our procedure, our ''baseline'' is a random sequence with correct amino acid composition).

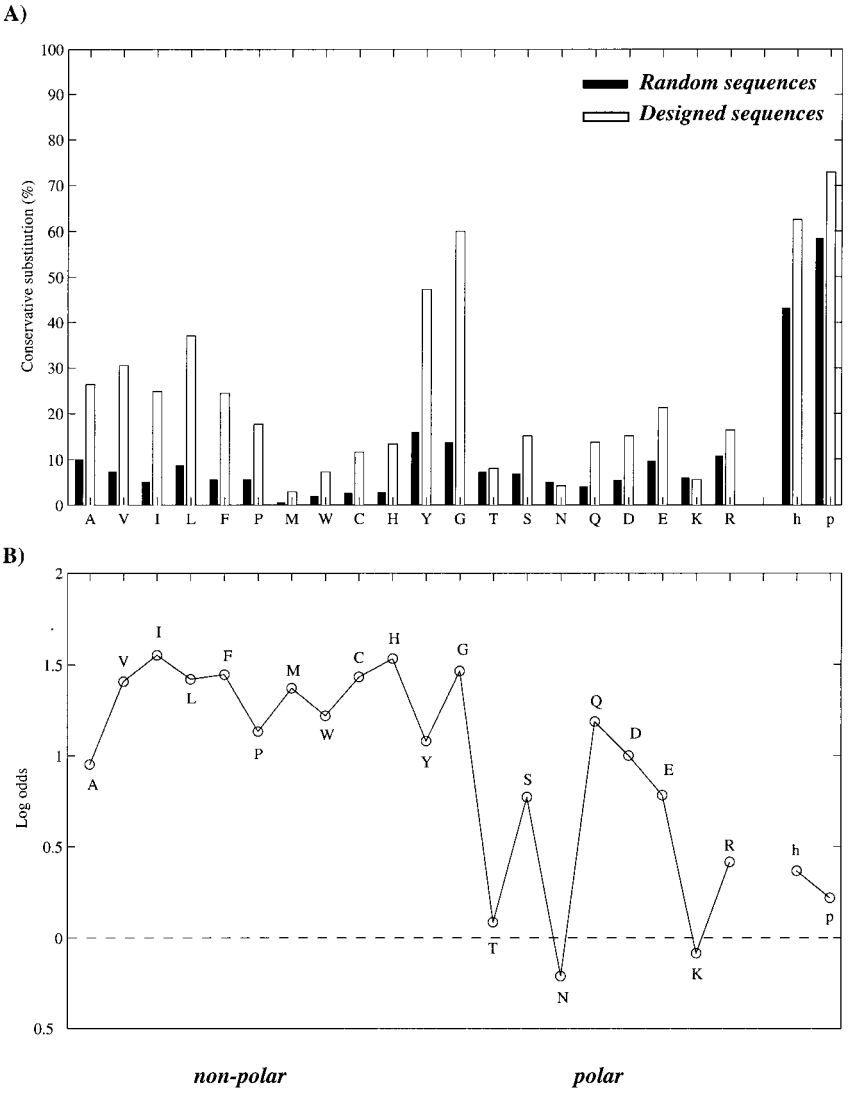

Figure 3. (a) Conservation of amino acid type between the native sequence of a protein, and random sequences with the same amino acid composition (filled bars), or computer-designed sequences (open bars), for each amino acid type. The letters h and p stand for hydrophobic residues (A, V, I, L, F, P, M, W, H, Y and C) and hydrophilic residues (G, S, T, N, Q, D, E, K, R), respectively. Conservation of an amino acid j is measured as the probability $P(j,j)$ of finding an amino acid j at any position in the sequence that is occupied by j in the native sequence. The probabilities were averaged over ten different proteins (Table 2). (b) specific conservation induced by the sequence design procedure, measured as the log-odds of the ratio of $P_{des}(j,j)$ and $P_{rand}(j,j)$, for each amino acid type j (see the text for details).

The PROFILESEARCH technique (Gribskov *et al.*, 1990) as implemented in the UWGCG package was used to scan the PDB sequence database (i.e. sequences of all proteins whose structure is included in the PDB databank) with all three profiles. A comparison of the results is shown in Figure 5.

There are 82 chains (including repeats) representing TIM sequences in the current release of the PDB. If we only keep one representative of chains with identical sequences, and eliminate the mutants, the 82 domains correspond in fact to nine TIMs from different species, which were all excluded from Prof_SEL. All 82 chains are well recognized by Prof_SEL, with a distinct separation from non-TIM sequences (Figure 5(a)).

Sequences designed for 1TIM are based on its amino acid composition and the backbone structure of the protein. No information on other TIM sequences of structures is included and the design proceeds with a physical energy function. Given this limitation, the profile matrix Prof_DES performs surprisingly well, as shown in Figure 5(a). The corresponding Z-scores for each chain correlate well with its cRMS to 1TIM-A (Figure 6). Interestingly, all chicken TIM chains obtained the highest Z-scores with both Prof_SEL and Prof_DES. It is worth noticing that each sequence

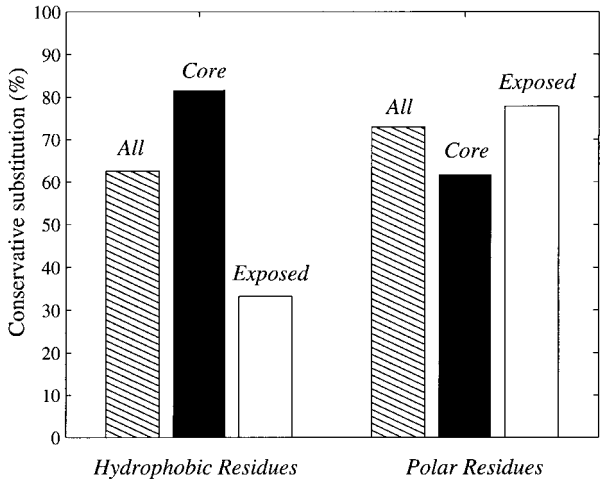

Figure 4. Conservation of amino acid type between the native sequence of a protein, and designed sequences with the same amino acid composition: all residues (filled bars), core residues (open bars) and exposed residues (hatched bars) (see the legend to Figure 3 for details).

included in the Prof_DES profile has a low sequence identity to 1TIM (see Table 3). A FASTA (Lipman & Pearson, 1985) search based on any of these sequences yields poor results, with no matches with an *E*-value lower than 0.1 (results not shown).

To test the influence of a fixed amino acid composition in our protein design procedure, the same analysis was performed using random sequences with the same amino acid composition than 1TIM. A scan of the PDB sequence database with the corresponding profile matrix, Prof_INI, does not identify any of the TIM sequence (Figure 5(b)). This clearly indicates that the ability of the profile matrix Prof_DES to identify TIM-like sequences is a result of our design procedure.

Discussion

A protein molecule is identified by both its amino acid sequence and its three-dimensional structure. A given protein sequence generally adopts a unique 3D structure of lowest free energy. The folding problem, which involves identification of this fold using the sequence, is a difficult and unsolved problem. The inverse folding problem, which involves identification of a sequence that corresponds to a particular fold, requires more caution, since it does not correspond to a physical process. We have recently proposed a new procedure to design protein sequences, based solely on the knowledge of the backbone of the template protein, and a physical potential. Though it uses an all-atom representation of the protein, this procedure is fast enough to allow design of several sequences for a given protein. Here, we focus on understanding how different these sequences are, as well as in potential applications of our design procedure to fold recognition.

We have shown that our design procedure does converge in sequence space, but not to the native sequence (Figure 2). There are two possible reasons for the fact that the zone of convergence does not include the native sequence: either our potential energy function introduces a bias, or nature has not selected the most stable sequence as the native sequence for the template fold we considered. These two reasons are not exclusive. The bias is apparent in the fact that the procedure tends to reduce the number hydrophobic residues at the protein surface to a lower level than that observed in the native sequence (Figure 4). While proteins do tend to establish a pattern of hydrophobic residues in the core, and hydrophilic residues at the surface (Miller *et al.*, 1987), there are exceptions. It was shown recently that proteins are tolerant to, and can even be stabilized by multiple polar-to-hydrophobic surface substitutions (Cordes &

Table 3. Characteristics of the 13 sequences designed for TIM

Sequence	Sequence identity with 1TIM (%)	Sequence *P*-value [a]	THREADER *Z*-score
Native	100	NA	19.8
D1	15.4	−1.7	5.0
D2	17.4	−4.7	9.7
D3	16.6	−0.24	7.2
D4	19.8	−5.12	9.4
D5	17.0	−0.06	7.1
D6	13.8	0.0	4.8
D7	16.6	−3.34	8.9
D8	16.6	−2.00	7.9
D9	16.6	0.0	5.1
D10	15.8	0.0	3.3
D11	16.6	−1.12	7.7
D12	17.8	−0.58	7.9
D13	15.8	0.0	5.0
Average (std)	16.6 (1.4)	−1.4 (1.8)	6.9 (2.0)

[a] Log_{10} of the estimated significance *P* of the sequence match between the designed sequence and the native sequence, based on the model proposed by Levitt & Gerstein (1998).

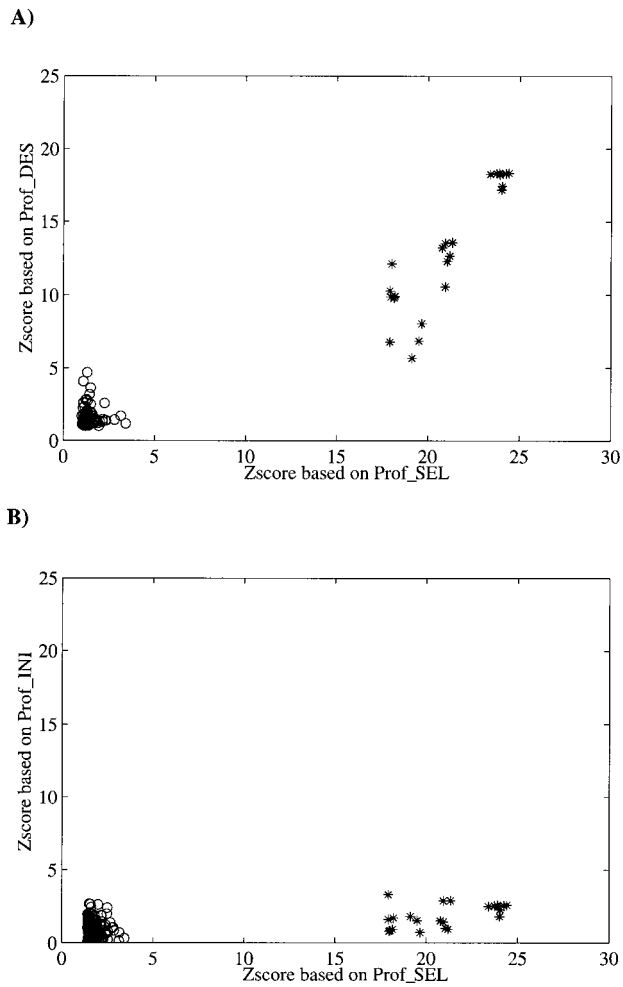

Figure 5. The sequences of all protein chains in the PDB databank were scanned using PROFILESEARCH (Gribskov *et al.*, 1990) with three different profiles specific to the sequence of chicken triose phosphate isomerase (PDB code 1TIM). The raw score of a sequence matched to a profile is converted to a Z-score, and only sequences with Z-scores greater than 1 are kept. See the text for the definition of the profiles. All multiple sequence alignments were generated by PILEUP, and the profiles were built using PROFILEMAKE (both programs are part of the UWGCG package). Chains corresponding to TIM sequences are identified with stars (*).

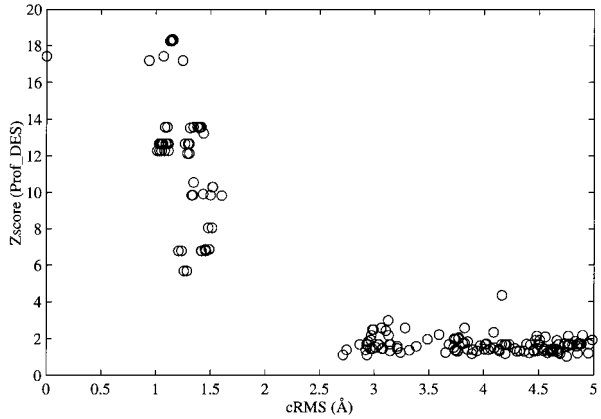

Figure 6. The sequences of all protein chains in the PDB databank are scanned using PROFILESEARCH with the profile matrix Prof_DES; the corresponding Z-scores are plotted *versus* the cRMS of the best structural alignment between the chain and the structure of 1TIM-A.

Sauer, 1999). Improvements of our energy function for protein design will have to include this fact. Interestingly, our procedure does maintain polar residues in the core (Figure 4), which gives specificity to designed sequences (see Koehl & Levitt, 1999b). The designed sequences obtained with our procedure show significant similarities to the native sequence. In particular, buried non-polar residues are well conserved. This is not consistent with the recent results by Rodionov & Blundell (1998) based on evolution-related multiple sequence alignments. The difference might not be significant we do observe conservation of polar residues in the core; the results are also highly dependent on the choice of the classification between polar and non-polar residues, which was not given in the study described by Rodionov & Blundell (1998).

Our approach to protein sequence optimization is designed to work on a full atom representation and allows direct optimization of the entire sequence. It remains fast: the average computing time for one Monte Carlo step during the optimization is one second for a 60 residue chain, and 4.4 seconds for a 250 residue chain (on a Compaq Alpha processor at 533 MHz). As a consequence, the procedure can be repeated several times, yielding a family of designed sequences. We have shown that the information contained in these designed sequences can be efficiently transferred to a profile or substitution matrix. This matrix was found to recognize to a significant level the native sequence corresponding to the protein fold used for design, as well as closely related sequences (i.e. sequences of the same protein in different species; Figure 5). It is worth noticing that this approach only includes the information about the backbone conformations of the protein and its amino acid composition.

One limitation of our designed profile as applied to threading is related to the use of too precise a structural template. Crystallographic studies have shown that both backbone and side-chain adjustments occur when residues within protein cores are mutated (for a review, see Baldwin & Matthews, 1994). We are currently working on including backbone flexibility, using multiple copies of the backbone (Koehl & Delarue, 1995).

Materials and Methods

Protein sequence design

A complete description of the protein design procedure is given in our previous work (Koehl & Levitt, 1999b). Briefly, the program starts from the backbone, B, of the template protein structure. A random sequence, S_0, is generated, based on the amino acid composition of the native sequence corresponding to the template fold. A full-atom model of the chimeric protein obtained by threading this sequence on the backbone B without gaps, is built using a self-consistent mean field (SCMF) approach to position the side-chains (Koehl & Delarue, 1994a). The energy E_0 of this model is computed including a Lennard-Jones potential for packing interactions, a Coulomb term for electrostatics interactions, and a surface area-dependent solvation term, which takes into account the overall environment of all atoms of the protein (Koehl & Delarue, 1994b). A new sequence S_1 is then considered by choosing two positions in S_0 at random, and exchanging the corresponding amino acid types. The energy, E_1, of the new model derived from sequence S_1 is calculated, and the move is accepted or rejected, using the classical Metropolis scheme (Metropolis *et al.*, 1953) (i.e. the move is accepted if a random number drawn from a uniform distribution between 0 and 1 is lower than $\exp[(E_0 - E_1)/kT]$). All simulations were carried out over 30,000 cycles of this procedure. This optimization scheme derives a stable sequence for the template fold. Specificity is assumed by keeping the amino acid composition fixed, according to the random energy model (Shakhnovich & Gutin, 1993a,b; see Koehl & Levitt, 1999b, for a discussion on the validity of using REM).

Schematic representation of the sequence space

A set of N sequences can be represented by N points $\{A_i\}$ in a high dimensional sequence space. This space cannot be visualized for N greater than 4. Sammon non-linear mapping techniques (Kruskal, 1964; Sammon, 1969; Agrafiotis, 1997) provide a means to project these N points on a two-dimensional plane. They work by generating a set of N coplanar points $\{P_i\}$, such that the set of inter-point distances are a reasonable approximation to the inter-sequence distances in sequence space.

The Euclidian distance between two points P_i and P_j in the plane is given by:

$$\mathrm{De}(P_i,\ P_j) = \left(\sum_{k=1}^{2}(P_{ik} - P_{jk})^2\right)^{1/2} \tag{2}$$

where P_{ik} is the kth coordinate of point P_i.

Let us define the "distance" between two sequences A_i and A_j as:

$$\mathrm{Ds}(A_i,\ A_j) = 100 - I(A_i,\ A_j) \tag{3}$$

where $I(A_i,A_j)$ is the level of sequence identity between A_i and A_j:

$$I(A_i,\ A_j) = \sum_{k=1}^{L}\delta(\mathrm{aa}(A_i,\ k) - \mathrm{aa}(A_j,\ k)) \tag{4}$$

where L is the length of each sequence, $\mathrm{aa}(A_i,k)$ the amino acid type (from 1 to 20) at position k in sequence A_i, and d a step function ($\delta(x) = 1,\ x = 0$ and $\delta(x) = 0$ otherwise).

The mapping error X between the high-dimensional sequence space, and its 2D representation is given by:

$$X(P) = \frac{1}{\sum_{i=2}^{N}\sum_{j=1}^{i-1}\mathrm{Ds}(A_i, A_j)} \times \sum_{i=2}^{N}\sum_{j=1}^{i-1}(\mathrm{De}(P_i, P_j) - \mathrm{Ds}(A_i, A_j))^2 \tag{5}$$

In non-linear mapping, the position of the points in the low dimension space are derived by a gradient descent procedure which minimizes the value of X. The procedure is iterative, and the positions of the $(m+1)$ configuration of $\{P_i\}$ is computed from the mth configuration using:

$$P_{ik}(m+1) = P_{ik}(m+1) - \alpha g_{ik}(m) \tag{6}$$

where:

$$g_{ik}(m) = \frac{\dfrac{\partial X}{\partial P_{ik}}}{\left|\dfrac{\partial^2 X}{\partial P_{ik}^2}\right|} \tag{7}$$

and α is a factor, which we set to 0.5.

Structural alignments

We have used STRUCTAL (Subbiah *et al.*, 1993) for protein structure superposition. This method starts with an arbitrary equivalence of the residues of the two proteins. This equivalence is used to perform a classical superposition of the two structures, from which a structural alignment matrix SA is calculated. The best structural alignment is then obtained by standard global dynamic programming on SA. Since the alignment may depend on the initial residue equivalence, the procedure is repeated for five different initial sets of correspondence, and the optimal alignment is taken as that with the highest score.

Acknowledgments

This work was supported by grants to M.L. from the Department of Energy (DE-FG03-95ER62135) and National Institutes of Health (GM45415). It was carried out while P.K. was on leave of absence from the CNRS institute, Strasbourg, France, partially funded by a long term fellowship from the Union Internationale Contre le Cancer, Geneva, Switzerland.

References

Agrafiotis, D. K. (1997). A new method for analyzing protein-sequence relationships based on Sammon maps. *Protein Sci.* **6**, 287-293.

Bairoch, A. & Apweiler, R. (1999). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucl. Acid Res.* **27**, 49-54.

Bairoch, A. & Böckman, B. (1991). The SWISS-PROT sequence databank. *Nucl. Acids Res.* **19**, 2247-2249.

Baldwin, E. P. & Matthews, B. W. (1994). Core-packing constraints, hydrophobicity and protein design. *Curr. Opin. Biotechnol.* **5**, 396-402.

Banner, D. W., Bloomer, A. C., Petsko, G. A., Phillips, D. C., Pogson, C. I., Wilson, I. A., Corran, P. H., Furth, A. J., Milman, J. D., Offord, R. E., Priddle, J. D. & Waley, S. G. (1975). Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5 Å resolution using amino-acid sequence data. *Nature,* **255**, 609-614.

Beamer, L. J. & Pabo, C. O. (1992). Refined 1.8 Å crystal-structure of the lambda-repressor operator complex. *J. Mol. Biol.* **227**, 177-196.

Bernstein, F. C., Koetzle, T. F., Williams, G., Meyer, D. J., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein DataBank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.

Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science,* **253**, 164-170.

Bryant, S. H. (1996). Evaluation of threading specificity and accuracy. *Proteins: Struct. Funct. Genet.* **26**, 172-185.

Chothia, C. (1992). One thousand fold families for the molecular biologist? *Nature,* **357**, 543-544.

Collyer, C. A., Guss, J. M., Sugimura, Y., Yoshizaki, F. & Freeman, H. C. (1990). Crystal structure of plastocyanin from a green-alga; enteromorpha-prolifera. *J. Mol. Biol.* **211**, 617-632.

Cordes, M. & Sauer, R. (1999). Tolerance of a protein to multiple polar-to-hydrophobic surface substitutions. *Protein Sci.* **8**, 318-325.

Dahiyat, B. I. & Mayo, S. L. (1997a). *De novo* protein design: fully automated sequence selection. *Science,* **278**, 82-87.

Dahiyat, B. I. & Mayo, S. L. (1997b). Probing the role of packing specificity in protein design. *Proc. Natl Acad. Sci. USA,* **94**, 10172-10177.

Delarue, M. & Koehl, P. (1997). The inverse protein folding problem: self consistent mean field optimisation of a structure specific mutation matrix. In *Proceedings of the Pacific Symposium on Biocomputing* (Altman, R. B., Dunker, A. K., Hunter, L. & Klein, T., eds), World Scientific, Singapore.

Devereux, J., Haeberli, P. & Smithies, O. (1984). A comprehensive set of sequence analysis programs for the vax. *Nucl. Acids Res.* **12**, 387-395.

Drexler, K. E. (1981). Molecular engineering: an approach to the development of general capabilities for molecular manipulation. *Proc. Natl Acad. Sci. USA,* **78**, 5275-5278.

Gallagher, T., Alexander, P., Bryan, P. & Gilliland, G. L. (1994). Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. *Biochemistry,* **33**, 4721-4729.

Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA,* **84**, 4355-4358.

Gribskov, M., Lüthy, R. & Eisenberg, M. (1990). Profile analysis. *Methods Enzymol.* **183**, 146-159.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature,* **358**, 86-89.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers,* **22**, 2577-2637.

Kim, S. H. (1998). Shining a light on structural genomics. *Nature Struct. Biol.* **5**, 643-645.

Koehl, P. & Delarue, M. (1994a). Application of a self consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.* **239**, 249-275.

Koehl, P. & Delarue, M. (1994b). Polar and non-polar atomic environment in the protein core: implications for folding and binding. *Proteins: Struct. Funct. Genet.* **20**, 264-278.

Koehl, P. & Delarue, M. (1995). A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. *Nature Struct. Biol.* **2**, 163-170.

Koehl, P. & Levitt, M. (1999a). A brighter future for protein structure prediction. *Nature Struct. Biol.* **6**, 108-111.

Koehl, P. & Levitt, M. (1999b). *De novo* protein design. I. In search of stability and specificity. *J. Mol. Biol.* **293**, 1161-1182.

Kohda, D., Hatanaka, H., Odaka, M., Mandiyan, V., Ullrich, A., Schlessinger, J. & Inagaki, F. (1993). Solution structure of the Sh3 domain of phospholipase C-gamma. *Cell,* **72**, 953-960.

Kruskal, J. (1964). Multi-dimensional scaling by optimizing goodness of fit to a non-metric hypothesis. *Psychometrica,* **29**, 1-27.

Laurents, D. V., Subbiah, S. & Levitt, M. (1994). Different protein sequences can give rise to highly similar folds through different stabilizing interactions. *Protein. Sci.* **3**, 1938-1944.

Leijonmarck, M. & Liljas, A. (1987). Structure of the C-terminal domain of the ribosomal protein-L7 protein-L12 from *Escherichia coli* at 1.7 Å. *J. Mol. Biol.* **195**, 555-580.

Lemer, C. M. R., Rooman, M. J. & Wodak, S. J. (1995). Protein structure prediction by threading methods: evaluation of current techniques. *Proteins: Struct. Funct. Genet.* **23**, 337-355.

Levitt, M. & Gerstein, M. (1998). A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl Acad. Sci. USA,* **95**, 5913-5920.

Lipman, D. J. & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science,* **227**, 1435-1441.

McPhalen, C. A. & James, M. N. G. (1987). Crystal and molecular structure of the serine proteinase-inhibitor Ci-2 from barley seeds. *Biochemistry,* **26**, 261-269.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087-1092.

Miller, R. T., Jones, D. T. & Thornton, J. M. (1996). Protein fold recognition by sequence threading: tools and assessment techniques. *FASEB J.* **10**, 171-178.

Miller, S., Janin, J., Lesk, A. & Chothia, C. (1987). Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641-656.

Mirny, L. & Domany, E. (1996). Protein fold recognition and dynamics in the space of contact maps. *Proteins: Struct. Funct. Genet.* **26**, 391-410.

Mirny, L. A. & Shakhnovich, E. I. (1998). Protein-structure prediction by threading: why it works and why it does not. *J. Mol. Biol.* **283**, 507-526.

Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature,* **372**, 631-634.

Pabo, C. (1983). Designing proteins and peptides. *Nature,* **301**, 200.

Pande, V. S., Grosberg, A. Y. & Tanaka, T. (1997). Statistical mechanics of simple-models of protein-folding and design. *Biophys. J.* **73**, 3192-3210.

Ponder, J. W. & Richards, F. M. (1987). Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775-791.

Rodionov, M. A. & Blundell, T. L. (1998). Sequence and structure conservation in a protein core. *Proteins: Struct. Funct. Genet.* **33**, 358-366.

Sali, A. (1998). 100,000 protein structures for the biologist. *Nature Struct. Biol.* **5**, 1029-1032.

Sammon, J. (1969). A non-linear mapping for data structure analysis. *IEEE Trans. Comp. ser. C,* **18**, 401-409.

Shakhnovich, E. I. & Gutin, A. M. (1993a). Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl Acad. Sci. USA,* **90**, 7195-7199.

Shakhnovich, E. I. & Gutin, A. M. (1993b). A new approach to the design of stable proteins. *Protein Eng.* **6**, 793-800.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-197.

Subbiah, S., Laurents, D. V. & Levitt, M. (1993). Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Curr. Biol.* **3**, 141-148.

Svensson, L. A., Thulin, E. & Forsen, S. (1992). Proline *cis-trans* isomers in calbindin D9k observed by X-ray crystallography. *J. Mol. Biol.* **223**, 601-606.

Takano, T. (1977). Structure of myoglobin refined at 2.0 Å resolution. II. Structure of deoxymyoglobin from sperm whale. *J. Mol. Biol.* **110**, 569-584.

Taylor, W. R. & Orengo, C. A. (1989). Protein structure alignment. *J. Mol. Biol.* **208**, 1-22.

Vendruscolo, M., Najmanovich, R. & Domany, E. (1999). Protein folding in contact map space. *Phys. Rev. Letters,* **82**, 656-659.

Wang, Z. X. (1998). A reestimation for the total numbers of protein folds and superfamilies. *Protein Eng.* **11**, 621-626.

Wlodawer, A., Walter, J., Huber, R. & Sjolin, L. (1984). Structure of bovine pancreatic trypsin-inhibitor: results of joint neutron and X-ray refinement of crystal form-Ii. *J. Mol. Biol.* **180**, 301-329.

Wright, C. S. (1990). 2.2 Å-resolution structure analysis of two refined *N*-acetylneuraminyl-lactose: wheat germ-agglutinin isolectin complexes. *J. Mol. Biol.* **215**, 635-651.

Edited by F. E. Cohen

(Received 15 July 1999; received in revised form 8 September 1999; accepted 9 September 1999)