doi:10.1016/S0022-2836(02)00971-3 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 323, 551–562

JMB

Sequence Variations within Protein Families are Linearly Related to Structural Variations

Patrice Koehl* and Michael Levitt

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

**Corresponding author*

It is commonly believed that similarities between the sequences of two proteins infer similarities between their structures. Sequence alignments reliably recognize pairs of protein of similar structures provided that the percentage sequence identity between their two sequences is sufficiently high. This distinction, however, is statistically less reliable when the percentage sequence identity is lower than 30% and little is known then about the detailed relationship between the two measures of similarity. Here, we investigate the inverse correlation between structural similarity and sequence similarity on 12 protein structure families. We define the structure similarity between two proteins as the cRMS distance between their structures. The sequence similarity for a pair of proteins is measured as the mean distance between the sequences in the subsets of sequence space compatible with their structures. We obtain an approximation of the sequence space compatible with a protein by designing a collection of protein sequences both stable and specific to the structure of that protein. Using these measures of sequence and structure similarities, we find that structural changes within a protein family are linearly related to changes in sequence similarity.

Keywords: protein sequence; protein structure; sequence design; protein sequence alignment; protein structure alignment

Introduction

A protein sequence folds into a unique three-dimensional structure. Interestingly, this one-to-one correspondence is no longer valid when all proteins are considered. The size of the protein structure space is much smaller than the size of the protein sequence space: it is commonly assumed that there are 1000 different protein folds, covering 10,000 different protein sequence families.[1–4] The fact that the structure space is finite and much smaller than the sequence space has given rise to the hope that it is possible to have representative structures for all these sequences. Thus, it will be possible to generate a model for a new sequence without going into the expensive procedures of systematic experimental structure determination. The success of this approach is, however, strongly correlated to our ability to identify a proper structural template for a protein of interest. Techniques designed to solve this "fold recognition" problem rely on the assumption that similarities between the sequences of two proteins imply similarities between the structures of these proteins.

Although there are many individual exceptions,[5] it is believed that when two proteins share 50% or higher sequence identity, their backbones differ by less than 1 Å RMS deviation. The confidence level in this inference is much lower, however, in the so-called "twilight zone"[6] of 20–30% sequence identity. When the sequences of two evolutionary related proteins diverge to the twilight zone limit, their backbones can be expected to differ by 2 Å RMS deviation.[7] A recent study by Rost on more than a 10^6 sequence alignments has shown, however, that more than 95% of all sequence alignments found in the twilight zone correspond to proteins with different structures.[8] On the other hand, structure alignments have identified homologous protein pairs with less than 10% sequence identity.[5,9] Interestingly, the average sequence identity between pairs of proteins with similar structures is 8–10%.[10] This low level of sequence similarity is the so-called "midnight zone".[10] Fold recognition ultimately aims at identifying pairs of homologous proteins in this midnight zone. The

Abbreviations used: PDB, Protein Data Bank; SSA, sequence space annealing; REM, random energy model.

E-mail address of the corresponding author: koehl@csb.stanford.edu

0022-2836/02/$ - see front matter

Table 1. Protein families and PDB structures used

Name	Code	Structures
Alkane dehydrogenase	ADH	1ede (310), 1cqw (295)
Azurin/cyanins	AAA	1aaj (105), 1ag6 (99), 1aiz (129), 1kdi (102), 1paz (120), 2aza (129), 2plt (98), 3pcy (99), 7pcy (98)
Cystein protease	CPR	1aec (217), 1pe6 (212), 1ppo (216), 1yal (218)
Fatty acid binding protein	FAP	1crb (134), 1hmr (131), 1mdc (131), 1opbA (133)
Ferredoxin	FXN	1blu (80), 1fdn (55), 1fxd (58), 5fd1 (106)
Globins	GLB	1bvd (153), 1lht (153), 1myt (146), 1vxb (153), 2gdm (153), 2hbg (147), 3sdhA (145), 5mbn (153)
Glutathione transferase	GLU	1glpA (209), 1gne (232), 1hna (217), 3gstA (217)
Proline isomerase	ISM	1cynA (178), 1lopA (167), 2rmbA (165), 2rmcA (182)
SH3 domains	SH3	1awj (77), 1bb9 (83), 1bu1A (57), 1bymA (72), 1griA (61), 1ihvA (52), 1shfA (59), 1shg (57), 2abl (65), 2hsp (71)
Serine protease	SPR	1sgt (223), 2ptn (223), 2sga (181), 2tbs (222)
Protein G	PTG	1pgb (56), 2ptl (78)
Scorpion toxin	STX	1agt (38), 1chl (36), 1gps (47), 1ica (40), 1pnh (30), 1sis (34)
Thioredoxin	THX	1mek (120), 1thx (108), 1tof (112), 2tir (108), 3trx (105)
Triose phosphate isomerase	TPI	1btmA (251), 1htiA (248), 1timA (247), 2ypiA (247)

PDB structures in our database. The number of residues for each protein is given in parentheses.

success of sequence-based techniques designed to reach that goal depends on the ability to define a robust correlation between sequence information and structure information.

Chothia & Lesk[11] were the first to report that the extent of the structural changes observed between two proteins is directly related to the extent of the sequence changes. They proposed that the root-mean-square deviation in the position of the main-chain atoms of the two proteins, cRMS, is related to the fraction of mutated residues, H, according to an exponential law:

$$\text{cRMS} = 0.4 \exp(1.87H) \quad (1)$$

Recent studies have confirmed their findings with larger sets of proteins.[12–14] The asymptotic behavior of cRMS for small values of sequence identity makes it difficult to identify pairs of proteins with similar structures but low sequence similarities. Interestingly, Wood & Pearson have shown that if the cRMS value and sequence identity measures are replaced by their statistical significance, a linear relationship is found between protein sequence similarity and structure similarity.[15] Their data, however, suggest that higher-order effects may play a role at low sequence similarity. They also found that the slope that characterizes the linear fit differs significantly among protein families, with no obvious correlation to protein size or protein mutation rates.

All these studies focus on the same question: can differences in protein sequences explain differences in the corresponding protein structures? Here, we propose to study the inverse fundamental question: can we correlate the structural differences between two protein folds to differences between the subsets of sequence space that are compatible with these folds? To answer this question, we need a reasonable description of the sequence space $S(X)$ compatible with a protein fold X, as well as a measure of the distance between two sequence space subsets. We propose to explore the sequence space $S(X)$ by designing a large number of sequences compatible with and specific to the fold X. The distance between two subsets $S(X)$ and $S(Y)$ is then defined as the mean sequence identity between any pairs of sequences in the set $S(X) \times S(Y)$ (i.e. every sequence in $S(X)$ with every sequence is $S(Y)$). We first illustrate our approach by comparing two structurally similar proteins whose sequences show low levels of similarity. This initial test case is followed by a more systematic study of 14 diverse structural families of proteins, namely alkane dehalogenases (ADH), cyanins/azurins (AZN), cystein proteases (CPR), fatty acid binding proteins (FAP), ferredoxins (FXN), globins (GLB), glutathione transferases (GLT), proline isomerases (ISM), protein-G like domains (PTG), SH3 domains (SH3), serine proteases (SPR), scorpion toxins (STX), thioredoxins (THX) and triose phosphate isomerase (TPI) (Table 1). Finally, we discuss the consequences of this work to sequence-based fold recognition techniques.

Results

Two proteins with highly similar sequences almost always share the same fold (the exceptions arise when a protein can adopt more than one conformation). The reverse, however, is not always true: Rost[10] has shown that pairs of proteins with similar structures possess, on average, only 8–10% sequence identity. Is this observation biased by the fact that the native sequence of a protein is only one among all sequences compatible with the structure of that protein? More generally, how convergent are the subsets of sequence space compatible with two similar proteins?

Protein G and protein L share a significant subset of sequence space

Protein G, the B1 immunoglobulin-binding domain of streptococcal protein G (GB1) (56

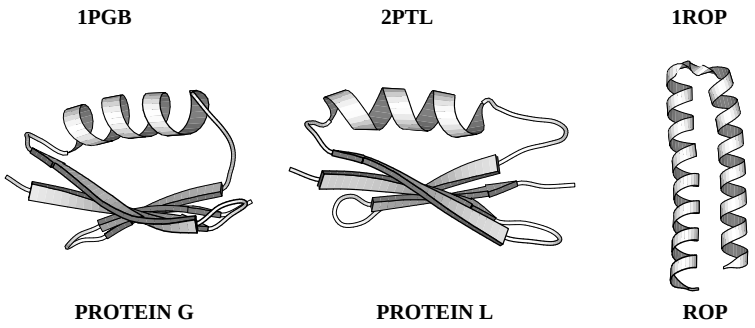

Figure 1. comparison of the folds of protein G, the B1 domain of streptococcal protein G (PDB code 1pgb), protein L, the B1 domain of *P. magnus* (PDB code 2ptl), and protein ROP, a transcription regulator of *E. coli* (PDB code 1rop). Only the last 61 residues of protein L are shown, since the first 17 residues have no equivalence in protein G. The drawings of the proteins were generated using MOLSCRIPT.[42]

residues), is a highly stable and very regular protein, with 80% of the residues participating in secondary structure. Its structure has been solved both by NMR[16] (PDB code 1pgx) and by X-ray crystallography[17] (PDB code 1pgb). The fold of protein G is not unique, and other proteins were found to adopt a very similar structure. Among these proteins, protein L, the B1 immunoglobulin light chain binding domain of *Peptostreptococcus magnus* (74 residues, PDB code 2ptl) is unique in that its sequence is not similar to the sequence of protein G (sequence identity of 14%).[18] Sequence-based methods such as FASTA, BLAST and PSIBLAST do not detect the homology between protein G and protein L. These two proteins are, however, structurally very similar (Figure 1). The structural alignment of 1pgb and 2ptl covers 52 residues, with a cRMS over CA of 1.9 Å.

For each protein, we design a set of 100 sequences using the sequence space annealing (SSA) procedure described in Methods. We use the X-ray structure of protein G and the NMR structure of the last 61 residues of protein L as target conformations. The first 17 residues of protein L, which have no equivalent in protein G and do not interact with the rest of protein L, were omitted. The amino acid composition of the native sequence is used as input for sequence optimization. The two sets each of 100 optimized sequences represent the subsets of sequence space compatible with each protein. The distance in sequence space between these two subsets is

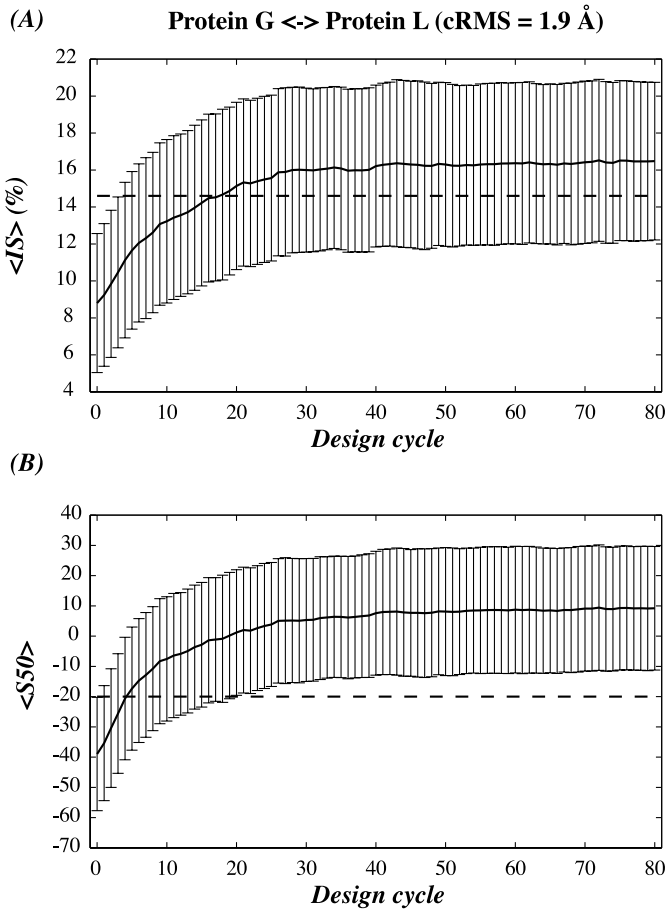

Figure 2. Comparing the subsets of sequence space compatible with protein G and protein L. The native sequences of protein G and protein L have no detectable similarities, while their structures are very similar (cRMS = 1.9 Å). Two subsets of 100 sequences are designed for these two proteins using the SSA procedure described in Methods. The sequences in each subset have the same amino acid composition as the corresponding native sequence. The mean sequence identity ⟨IS⟩ and the mean sequence similarity ⟨*S*50⟩ on the basis of the Blosum 50 matrix between these two subsets are plotted as a function of the cycle of the SSA design procedure. The vertical bars show the standard deviations of the corresponding distributions. Monotonous increases of both measures are observed, indicating a convergence in the sequence information contained in the two proteins. For comparison, the percentage sequence identity and similarity score between the native sequences of proteins G and L are shown as discontinuous lines.

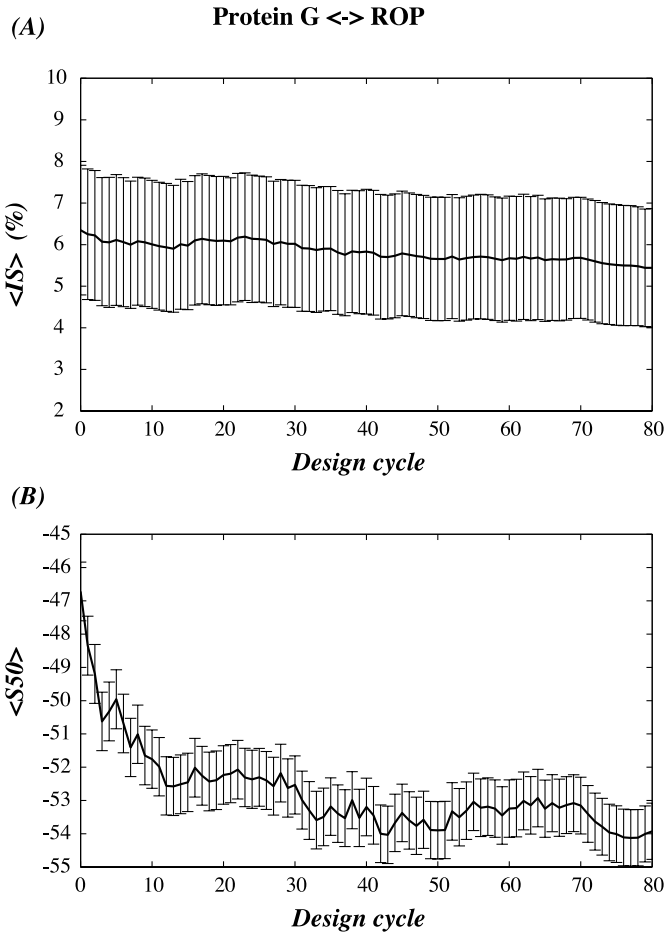

Figure 3. Comparing the subsets of sequence space compatible with protein G and protein ROP. Protein G and protein L have no detectable sequence or structure similarities. Two subsets of 100 sequences are designed for these two proteins using the SSA procedure described in Methods. The mean sequence identity ⟨IS⟩ and the mean sequence similarity ⟨*S*50⟩ based on the Blosum 50 matrix between these two subsets are plotted as a function of the cycle of the SSA design procedure. No convergence is sequence space is observed.

defined either as the mean sequence identity, ⟨IS⟩, or the mean sequence similarity on the basis of the BLOSUM50 substitution matrix, ⟨*S*50⟩, computed over all pairs of sequences *A* and *B* designed for protein G and protein L, respectively. Percentage sequence identity and sequence similarities are computed on the basis of the structural alignment between protein G and protein L, as described in Methods. Both ⟨IS⟩ and ⟨*S*50⟩ are plotted against the SSA cycle number in Figure 2. The initial sequences chosen for protein G are 8.8% identical with those chosen for protein L, on average, with a mean similarity score ⟨*S*50⟩ of −39. Both the mean sequence identity and the mean similarity score increase monotonically during the SSA optimization. After convergence, sequences designed for protein G are 16.5% identical with sequences designed for protein L, on average, with a mean similarity score of 9. The two most similar sequences designed for protein G and protein L, respectively, have a sequence identity of 33%, and a similarity score of 89. These figures are outside the twilight zone, and would warrant weak detection of sequence homology, using any sequence comparison method. The two distance measures in sequence space (i.e. sequence identity, IS, and sequence similarity, *S*50) show that there is convergence between the sequence information derived from protein G, and the sequence information derived from protein L.

The specificity for protein G of the sequences designed for protein G was tested using THREADER2.[19] Since THREADER2 is a fold recognition program that uses scoring functions on the basis of statistics of interactions in known native protein structures rather than physical forces, it can be considered as a reasonable independent assessor of our sequence design procedure. The highest scoring folds for all 100 sequences designed for protein G correspond to the protein G folds included in the THREADER2 library (1pgb, 2igg and 1igd), with significant *Z*-scores higher than 3 (THREADER2 *Z*-score are defined such that significant matches have large positive *Z*-scores). Interestingly, 34 of these sequences also identify protein L as a compatible fold with a *Z*-score higher than 2. In comparison, the native sequence of protein G does not show any specificity to protein L, as measured by THREADER2 (*Z*-score − 10).

In parallel, the specificity for protein L of the sequences designed for protein L was also tested using THREADER2. For 80 out of these 100 sequences, the protein L fold has the highest score. Interestingly again, 67 of these sequences also identify protein G as a compatible fold, with a *Z*-score higher than 2. In comparison, the native

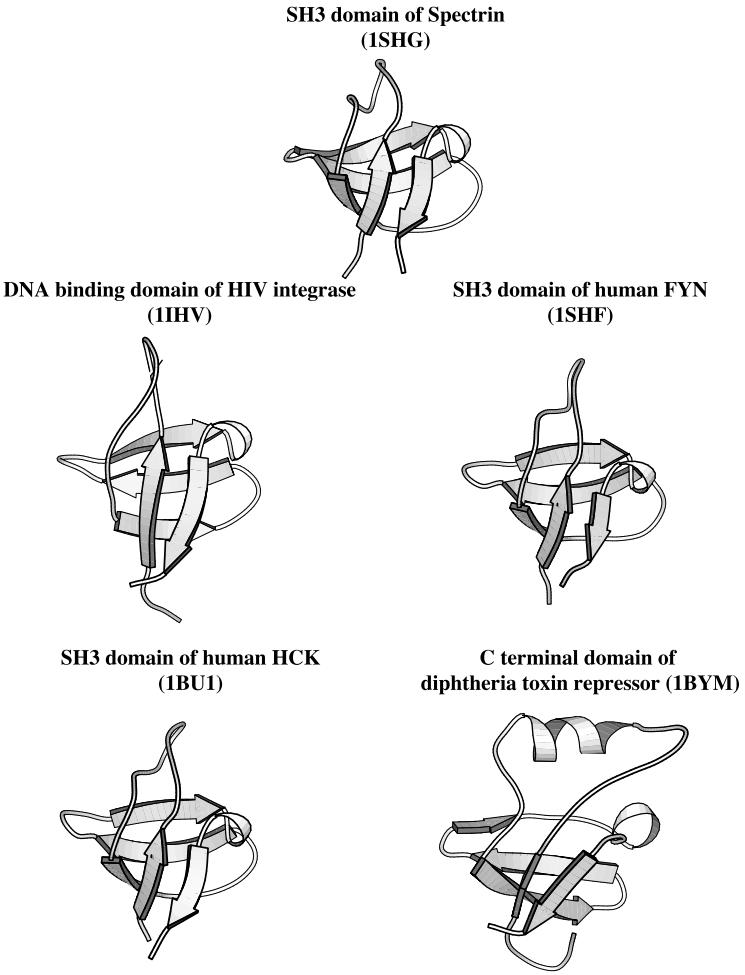

Figure 4. Five SH3 domains. The drawings of the proteins were generated using MOLSCRIPT.[42]

sequence of protein L detects the fold protein G with a low specificity (Z-score = 1.45 < 2).

Protein G and ROP cover different regions of protein sequence space

Protein G and protein L have similar structures, and we have just shown that these structures contain convergent sequence information. A required validation experiment of our procedure is to compare the sequence content of two proteins that are structurally dissimilar. We chose the pair of proteins: (1) protein G and (2) the ROP protein, a transcription regulator of *Escherichia coli*[20] (PDB code 1rop). Protein G is mainly a β protein, while ROP is a fully α protein (Figure 1). Interestingly, protein G and ROP served as templates in the published solution to the Paracelsus challenge: protein G was forced to adopt the ROP fold by changing only 50% of its sequence.[21] For each protein, we design a set of 100 sequences using the SSA procedure described in Methods. The two subsets of 100 optimized sequences represent the regions of sequence space compatible with each protein. The distance in sequence space between these two subsets is plotted against the SSA cycle number in Figure 3. Both distance measures in sequence space (i.e. sequence identity, IS, and sequence similarity, *S*50) show that there is no convergence between the sequence information derived from protein G, and the sequence information derived from ROP.

Structure similarity implies sequence similarity: comparing SH3 domains

SH3 domains are small (55–70 residues) protein modules that mediate protein-protein interactions by binding to Pro-rich peptide sequences. They have no fixed topological position within proteins and are often found in combination with other protein–protein interaction modules such as SH2 domains. As of 1 April 2001, the structures of 75 proteins containing an SH3 domain have been solved using both X-ray crystallography and NMR spectroscopy. SH3 domains share a common fold, a five-stranded β-sandwich. The same fold, referred to as the SH3 fold, is observed in diphtheria toxin repressors, in myosin S1 fragments, in electron transport accessory proteins, in some translation proteins and in the DNA binding domain of retroviral integrase.[22] Examples of the fold are shown in Figure 4. The structure similarities observed among these different protein families do not give rise to detectable sequence similarities. For example, the mean sequence identity between any two SH3 domains in the PDB is

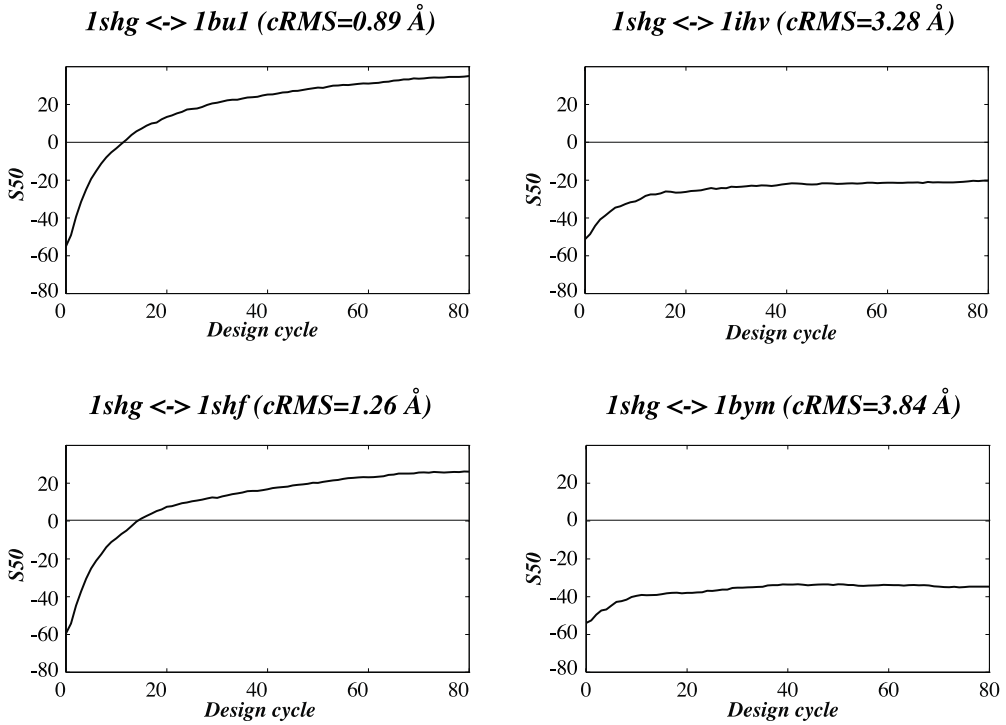

Figure 5. Comparing the subsets of sequence space compatible with the five SH3 domains shown in Figure 4. Five subsets of 100 sequences are designed for the five domains, using the SSA procedure. The mean sequence similarities ⟨*S*50⟩ between the set generated for the SH3 domain of spectrin, 1SHG, and all four other subsets are plotted *versus* the cycle of SSA. The mean similarity score is found to be higher when the cRMS between the two structures is low (i.e. for the two pairs 1SHG-1SHF and 1SHG-1BU1), and conversely lower when the cRMS is high (i.e. for the two pairs 1SHG-1IHV and 1SHG-1BYM).

21% with a minimum of 5% (the sequence identity is computed on the basis of the structural alignment of the two SH3 domains considered).

Our initial test case includes five small proteins covering a large range of structure variation within the SH3 fold: the SH3 domain of spectrin (PDB code 1shg), the SH3 domain of human FYN (PDB code 1shf), the SH3 domain of human HCK (PDB code 1bu1), the DNA binding domain of HIV integrase (PDB code 1ihv), and the C-terminal domain of diphtheria toxin repressor (PDB code 1bym). All five proteins are shown in Figure 4.

We designed a set of 100 sequences for each protein in our test set, using the SSA procedure. Pairs of sequences designed for the SH3 domain of spectrin show 37% sequence identity, on average. This relatively small value suggests that the sequence space compatible with the SH3 fold of 1shg is diverse, and large. The same behavior is observed for all other SH3 domains in our test set. This is in agreement with the fact that SH3 domains represent a very diverse protein sequence family.

The subset of sequences designed for the backbone of the SH3 domain of spectrin is compared to each of the four subsets of sequence space corresponding to the five other proteins in our test set. The evolution of the sequence similarity scores ⟨*S*50⟩ are shown in Figure 5. In all four cases, the sequence similarity between the subsets increases during the sequence optimization. The amplitudes of the increase differ, however, between the different pairs of SH3 domains. We find that the levels of sequence identity and sequence similarity observed between two sequence subsets after convergence of the SSA sequence optimization procedure are correlated with the structural similarity between the corresponding target SH3 domains (Figure 6).

In all test cases presented above, percentage sequence identity measure by ⟨IS⟩ and sequence similarity measured by *S*50 show identical behavior. For simplicity, in the following we only focus to the percentage sequence identity as a measure of sequence similarity.

A linear correlation between protein structure similarity and percentage sequence identity

Is the correlation observed between protein structure similarity and sequence similarity general, or specific to SH3 domains? To attempt to answer this question, we consider a much larger data set of 68 proteins belonging to 14 protein families, corresponding to 179 pairs of structurally similar proteins. These proteins vary in size from 30 residues (1pnh, a scorpion toxin), to 310 residues (1ede, an alkane dehydrogenase from *Xantobacter autotrophicus*), and were chosen to cover all four types of proteins (α, β, α + β and α/β). A high-resolution structure is available in the PDB for all 68 proteins. A full description of the families and proteins is given in Methods.

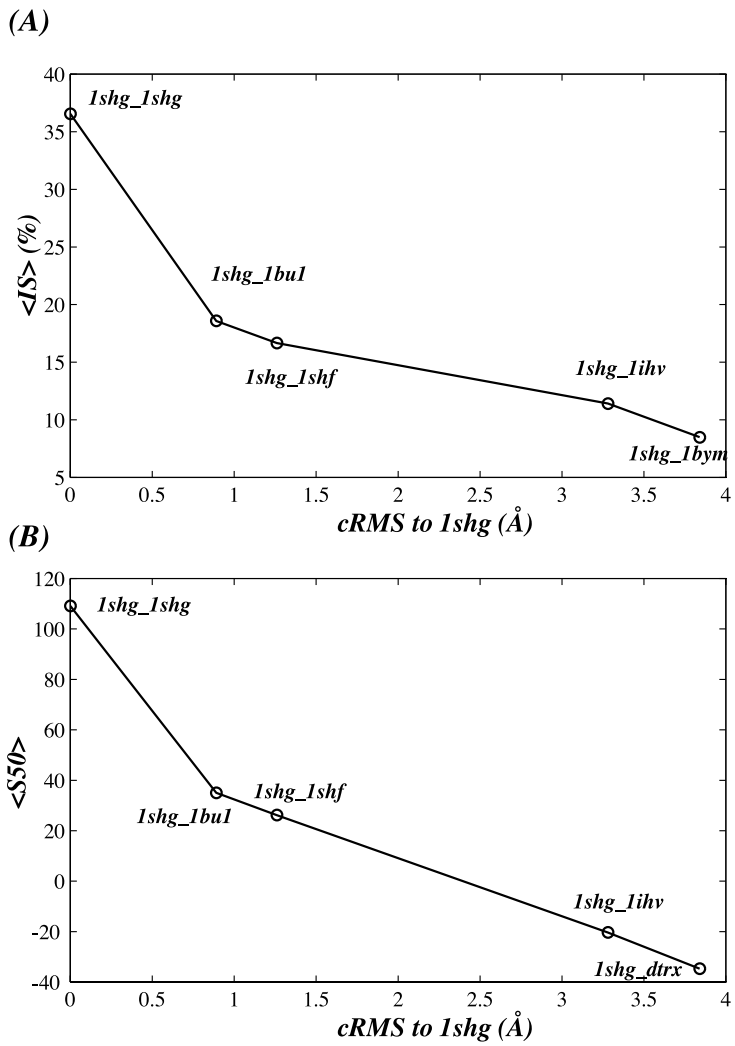

Figure 6. The sequence similarity between two SH3 domains is correlated with their structure similarity. The mean sequence identity ⟨IS⟩ (a) and the mean sequence similarity ⟨*S*50⟩ (b) between the converged sets of sequences designed for two SH3 domains are plotted *versus* the cRMS distance between the structures of these domains. Structural alignments were generated using STRUCTAL.[23]

The overall extent of the structural divergence of each pair of proteins in each family is measured by optimally aligning their two structures and computing the corresponding cRMS distance. The structural alignment was computed using STRUCTAL.[23] The optimal alignment provides a list of equivalent residues, which is used to define the percentage of identical residues, IS_{NAT} between the two proteins. For all 179 pairs of proteins in our data set, IS_{NAT} is found to vary exponentially with cRMS (Figure 7(a)). A least-squares fit to the data gives the relationship:

$$\text{cRMS} = 4.02 \exp(-0.0222 \text{IS}_{\text{NAT}}) \qquad (2)$$

where cRMS is measured in Å and IS_{NAT} is given in percentage. The values of cRMS predicted by equation (2) are, on average, 0.55 Å different of the observed cRMS values. Interestingly, equation (2) can be rewritten as:

$$\text{cRMS} = 0.44 \exp(2.22H) \qquad (3)$$

where H is the fraction of mutated residue (i.e. $H = 1 - \text{IS}_{\text{NAT}}/100$). Equation (3) reveals an asymptotic behavior between structure similarity and native sequence similarity, which was originally observed by Chothia & Lesk on a smaller test set[11] (see equation (1), $\text{cRMS} = 0.4 \exp(1.87H)$). Using equation (1), the predicted cRMS values are, on average, 0.70 Å different from the observed cRMS value. A similar non-linear relationship is observed between cRMS and sequence similarity defined as the raw FASTA score of the alignment of the native sequences (Figure 7(b)).

We designed 68 sets of 100 sequences for the 68 proteins in our data sets, using the SSA procedure described in Methods. All distances between all pairs of sequence sets were computed, and the mean sequence identities ⟨IS⟩ are compared to the cRMS distances between the corresponding proteins. Results for the 179 pairs of proteins in our data sets are shown in Figure 7(c). A linear relationship is observed between sequence similarity and structure similarity for all 14 families of proteins. A least-squares fit to the data gives:

$$\text{cRMS} = -0.154\langle \text{IS} \rangle + 4.14 \qquad (4)$$

with a correlation coefficient $R = 0.80$. The cRMS values computed from equation (4) differs from the observed cRMS values by 0.4 Å, on average.

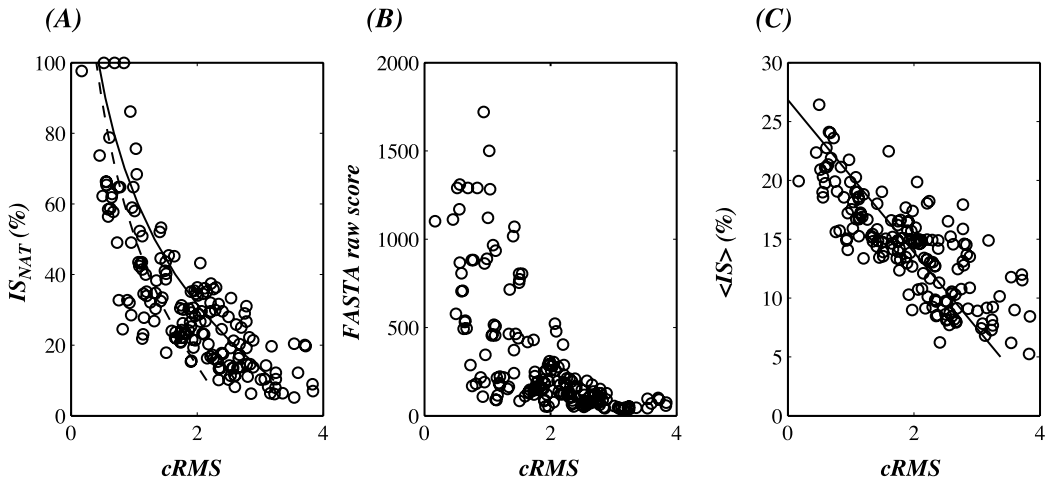

Figure 7. The relationship between sequence similarity and structure similarity is studied over 14 protein structure families (see Table 1). (a) For each pair of proteins in each family, the percentage sequence identity between their native sequences IS_{NAT} is plotted *versus* the cRMS distance between their structures. A least-squares fit to the data gives the relationship $cRMS = 4.02 \exp(-0.0222 IS_{NAT})$. The continuous line shows the fit to the data, while the discontinuous line shows the relationship between IS_{NAT} and cRMS predicted from equation (1) derived from Chothia & Lesk.[11] (b) The sequence of each protein is given as input to a FASTA search for sequence similarity over the database of protein sequences derived from the PDB. For each pair of proteins considered in (a), the raw FASTA score is plotted *versus* the cRMS distance. A non-linear between the FASTA score and cRMS is observed. (c) Sets of 100 sequences are designed for all proteins in the three structure families. For each pair of proteins, the mean percentage sequence identity ⟨IS⟩ between the converged sets of sequences designed based on their structures is plotted *versus* the cRMS distance. A linear relationship is observed, with a correlation coefficient $R = 0.80$. The continuous line shows the fit to the data.

Discussion

Proteins with similar sequences adopt similar structures.[6,11,24,25] This observation has given hope that it is possible to build model structures for all protein sequences, without going into the expensive procedures of systematic experimental structure determination: this is the justification of structural genomics. Interestingly, the reverse correlation is not always true: systematic comparison of protein structures included in the PDB database[26] have uncovered homologous protein structure pairs with less than 10% pairwise sequence identity.[5] These counter examples emphasize the need for a better understanding of the relationship between protein sequence similarity and structure similarity. Here, we find that structural changes in a protein family, as measured by cRMS, is linearly related to changes in sequence similarity, measured as the mean distance between the subsets of sequence space compatible with the pair of proteins considered. This relationship extends over a much larger range of sequence similarities than the relationship obtained when using the native protein sequence alone.

Sequence and structure similarity

Three levels of protein sequence similarities are usually considered, i.e. above 30% sequence identity, the region 20–30% (the so-called twilight zone[6]), and the region below 20% (the midnight zone[10]). More than 90% of all protein pairs with a sequence identity larger than 30% are found to be structurally similar, while in the midnight zone this number falls below 10%.[8] These observations favor a strongly non-linear relationship between sequence and structure similarity. Early studies by Chothia & Lesk have shown the existence of this non-linear behavior on a small data set of 36 proteins.[11] Their findings were confirmed by more recent studies on larger data sets.[12–14] We have found the same non-linear behavior in this study, both qualitatively and quantitatively, when plotting the percentage sequence identity between the native sequences of a pair of proteins *versus* the cRMS distance between their two structures (Figure 7(a) and equation (3)). All these studies, including ours, have used the percentage sequence identity and the cRMS difference of superimposed C^{α} atoms to measure sequence and structure similarity, respectively. After it was shown, however, that these two measures do not provide a reliable assessment of similarities[5,9] Wood & Pearson[15] revisited the relationship between sequence and structure similarity, using statistical Z-scores rather than raw scores to assess similarity. On the basis of 36 protein families, they found that most of the structural changes in a protein family are linearly related to changes in sequence similarity, when plotted in terms of statistical significance of the cRMS and percentage sequence identity.

Wood & Pearson[15] consider the apparent contradiction between the linear correlation observed for measures of similarities on the basis of statistical significance, and the non-linear relationship observed between cRMS and percentage sequence identity to be due to three technical problems.

Firstly, protein structure data determined under different conditions can cause the same protein sequence to have structural variants. This problem occurs in our test case on the basis of the globin family, which contains three structural isoforms of a myoglobin, namely 5mbn (sperm whale myoglobin), 1bvd (complex of sperm whale myoglobin with biliverdin), 1vxb (sperm whale myoglobin at pH 4). 1bvd and 1vxb differ from 5mbn by 0.69 Å and 0.52 Å, respectively, even though they have exactly the same sequence. Secondly, techniques for protein structure comparison do not always generate the best alignment. In order to reduce the effect of this serious problem, we decided to include in our study only pairs of structurally similar proteins with a cRMS value lower than 4.0 Å, whose structural alignment covers at least 50% of the smallest of the two proteins considered. Thirdly, sequence alignment techniques often fail to correctly align two distantly related proteins. This was not a direct problem in our study, since we relied on the structure alignment to derive the percentage sequence identity.

Though these are real technical problems that prevent accurate measurements of cRMS and sequence identity, we believe that there is a more fundamental reason for the non-linear relationship between the measures of structure similarity and sequence similarity. The percentage sequence identity is computed on the basis of the native sequences of the proteins considered. We believe that the relationship between the structure similarity and the sequence similarity of two proteins is meaningful when the sequence similarity is defined as the overlap between the subsets of sequence space compatible with these two proteins. On the basis of the results shown in Figure 7(c), we conclude that there is a linear relationship between protein structure similarity measured by cRMS and protein sequence similarity measured by percentage sequence identity, if the latter reflects the mean distance between the two subsets of sequences compatible with the two structures.

The linearity of the structure/sequence relationship is observed for protein pairs with cRMS values between 0 Å and 4 Å, whose native sequences are 10–100% identical, i.e. well within the twilight zone. We do not show any results for weaker structural similarities, i.e. for cRMS values larger than 4 Å: it is more than likely that for these values, the linearity breaks. As mentioned above, however, it is difficult to provide an accurate quantitative measure of weak structural similarity. Consequently, the relationship between structure and sequence similarity would be difficult to assess. This limit in our analysis emphasizes the need to develop better protein structure alignment programs.[27]

A local or global model for the protein folding code?

There are two prevailing models that explain how the tertiary structure of a protein is encoded in its linear sequence: the local model, in which fold specificity is coded in just a few critical residues, and the global model in which the fold is formed by interactions involving the entire sequence.[28] Wood & Pearson stated that the linear relationship observed by between sequence and structure similarity measured by statistical significance is more consistent with the global model than with the local model.[15] On the basis of our results, we cannot assert that the linearity of the structure/sequence relationship is inconsistent with the local model. Our protein design procedure generates sequences for a protein on the basis of all residue interactions in that protein. We have found that among all sequences designed for a protein, some positions are more conserved than others. This is the case, for example for residues like Gly and Pro that are structurally important.[29,30] This heterogeneity in the amino acid specificity of some residues in a protein is implicitly encoded in our measure of sequence similarity, and we do observe a linear relationship between sequence and structure similarity.

The fold recognition problem

Understanding the relationship between structural similarity and sequence similarity in protein families is a fundamental problem that has practical applications to the fold recognition problem, namely find a structural fold that best matches a given sequence. Sequence alignment techniques used for fold recognition unambiguously distinguish between protein pairs of similar and non-similar structures, as the similarity between the native sequences of the proteins considered is high. The use of sequence similarity to assess structure similarity becomes statistically less reliable when the percentage sequence identity decreases. Methods often fail to correctly align protein pairs in the twilight zone (i.e. with 20–30% pairwise sequence identity), resulting in detection of structural homology with low statistical significance.

We have shown that when sequence similarity is defined on the basis of the subset of sequences compatible with a protein structure rather than the native sequence alone, we observe a linear sequence/structure relationship. We also observe that this relationship is better defined, with less dispersion. The percentage sequence identity, IS_{NAT}, between the native sequences of two proteins whose structures are 1 Å cRMS apart can be anywhere between 20% and 100% (Figure 7(a)), while the mean sequence identity, $\langle IS \rangle$, between the sequence spaces compatible with these two proteins is between 15% and 25% (Figure 7(c)). This result suggests that more reliable detection of structure similarity can be achieved if sequence similarity is defined on the basis of families of sequences, rather than on the basis of the native sequence alone. This result is not new: it is in fact at the root of all profile methods used in modern database searching programs such as PSIBLAST[31]

and HMMER.[32] This paper provides, however, the structural validation for these methods, in that a single protein structure is better defined by the set of sequences compatible with its fold than by its native sequence alone.

The protein sequence profiles used in fold recognition techniques such as PSIBLAST and HMMs are generated from multiple sequence alignments of proteins expected to share a similar fold. Most of these methods do not take into account structural information, and therefore rely on stringent tests of sequence similarity when selecting sequences that belong to the same family (usually E-value cutoff of 10^{-4}). These methods might consequently generate too restrictive profiles for structural folds that are shared by very diverse sequences, with small sequence similarities (in the twilight zone for example). The SSA technique described here offers an alternative approach for generating multiple sequence alignments. Starting from an exhaustive library of protein structure representatives, we could run SSA to generate for each representative an alignment of 100 sequences compatible with its structure. These alignments would then be used to generate a library of profiles, which in turn would serve as a database for profile-based fold recognition techniques. The current version of SSA is too slow for such a large-scale experiment, but we are currently working on improving its performance in terms of CPU usage, such that this experiment will become computationally tractable.

Methods

The sequence information contained in a protein structure X is defined by the set $S(X)$ of protein sequences that adopt this fold. We propose to explore $S(X)$ by designing a large number, N, of sequences compatible and specific to X. For two protein structures X and Y, two subsets of sequences are thereby generated, $S(X)$ and $S(Y)$. The similarity between $S(X)$ and $S(Y)$ provides an estimate of the sequence similarity between the two proteins; it is defined as the mean distance between the two subsets in sequence space. Each step of the procedure is detailed below.

Defining the sequence space compatible with a protein structure

Subsets of sequences compatible with a given protein fold are derived using a genetic algorithm in sequence space. This method has been fully described elsewhere.[33] For a target protein structure, C, SSA starts with a set of N sequences, all with the same given amino acid composition (this is a consequence of using the random energy model (REM); see below). For each sequence, A_i, a full atom model structure is built, and its energy is evaluated and stored as $E(A_i, C)$. We use a semi-empirical energy function, which includes a Lennard Jones potential for van der Waals (vdW), a Coulomb term for electrostatics, and a free energy term for the solvent. Optimization is performed as follows. Starting with sequence A_1, M new sequences B_m are generated, each derived by random exchange of the amino acid types of K randomly chosen positions in A_1. Models structures are built for each B_m, and the corresponding energies are stored in $E(B_m, C)$. Each new sequence B_m is characterized by the sequence A_c in the initial set that is closest to B_m. If the distance between A_c and B_m is smaller than a given cutoff value and $E(B_m, C)$ is smaller than $E(A_c, C)$, A_c is replaced by B_m in the sequence set. If, on the other hand, B_m and A_c are further apart than the cutoff, B_m is added to the set. The sequences in the set are then sorted on the basis of energy, and only the N best are kept. The procedure is repeated for all B_m and all A. The updated set serves then as input to the following cycle, and the full procedure is repeated until the system has equilibrated and the variance of the sequence space described by the set remains steady. We refer to this procedure as SSA, in analogy to the "conformational space annealing" or CSA technique introduced by Scheraga and co-workers[34,35] for exploring protein structure space. SSA maintains a full atom representation of the model proteins, and has a built-in procedure to ensure the specificity of the designed sequences for their target structures.

Building full-atom model proteins

The compatibility of a sequence A with a protein X is tested by first building a model of A onto X. Side-chains are positioned using our self-consistent mean field approach.[36] The procedure iteratively refines a conformational matrix of the side-chains of the protein, CM, such that its current element at each cycle, CM(i,j), is the probability that side-chain i of the protein adopts the conformation of its possible rotamer j. Interactions and hence probabilities depend solely on a Lennard Jones function for van der Waals interactions (electrostatics and solvent interactions are ignored). The rotamer with the highest probability in the optimized conformational matrix defines the conformation of the side-chain in the final model. This model is characterized by an energy, $E(A, X)$, corresponding to the energy of the chimeric protein obtained by threading the side-chains corresponding to sequence A onto the backbone of fold X. This energy is derived from estimates of the physical forces that stabilize native protein structures: it includes van der Waals interactions, electrostatics and an environment free energy.[37]

Reaching specificity

A sequence, A, designed for a target fold, X, must be stable for that conformation. This is reached by minimizing $E(A, X)$. Sequence A must also be specific to X, i.e. incompatible with competing folds. A rigorous solution to this problem requires simultaneous and complete explorations of sequence space and conformation space. We have shown, however, that under the approximation of the REM,[38] specificity can be achieved by optimization in sequence space alone, provided that the amino acid composition of the sequence is held constant.[39,40]

Comparing two subsets of protein sequence space

Let $S(X)$ and $S(Y)$ be the sets of N sequences compatible with two protein structures X and Y, respectively, and let A_i and B_j be a pair of sequences in $S(X) \times S(Y)$. The similarity between A_i and B_j is defined from the

structural alignment of X and Y, and a measure of amino acid similarity.

We have used STRUCTAL[23] for protein structure superposition. The protein structure alignment between X and Y provides a list of P equivalent positions in the sequences A_i and B_j. For each equivalent position p, the corresponding positions in A_i and B_j are referred to as $EA(p)$ and $EB(p)$, respectively.

We use two different measures of similarity: sequence identity, I, and sequence similarity, SIM50. The sequence identity (in percent) between A_i and B_j is defined by:

$$I(A_i, B_j) = \frac{100}{P} \sum_{p=1}^{P} \delta[A_i(EA(p)), B_j(EB(p))] \quad (5)$$

The sum extends over all P equivalent positions in the structural alignment of X and Y, $EA(p)$ and $EB(p)$ are the positions is A_i and B_j of the pth equivalent position in the alignment, and δ is the Dirac function ($\delta(x,y) = 0$ if $x \neq y$ and $\delta(x,x) = 1$).

The sequence similarity between A_i and B_j is given by:

$$\mathrm{SIM}(A_i, B_j) = \frac{1}{P} \sum_{p=1}^{P} \mathrm{BL50}(A_i(EA(p)), B_j(EB(p))) \quad (6)$$

where BL50 stands for the Blosum 50 substitution matrix.[41]

Both I and SIM50 are distance measures in protein sequence space.

The similarity between the complete subsets $S(X)$ and $S(Y)$ is characterized by the means and standard deviations of the distribution of distances over all pairs of sequences (A_i, B_j) of $S(X) \times S(Y)$:

$$\langle \mathrm{IS}(X, Y) \rangle = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} I(A_i, B_j) \quad (7)$$

$$\sigma\mathrm{IS}(X, Y)^2 = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} I(A_i, B_j)^2 - \mathrm{IS}(X, Y)^2 \quad (8)$$

for the distance based on sequence identity, and

$$\langle S50(X, Y) \rangle = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{SIM50}(A_i, B_j) \quad (9)$$

$$\sigma 50(X, Y)^2 = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{SIM50}(A_i, B_j)^2 - S50(X, Y)^2 \quad (10)$$

for the distance based on sequence similarity.

Our protein data set

The sequence/structure relationship was studied over a set of 14 structural families: alkane dehalogenases (ADH), cyanins/azurins (AZN), cystein proteases (CPR), fatty acid binding proteins (FAP), ferredoxins (FXN), globins (GLB), glutathione transferases (GLT), proline isomerases (ISM), protein-G like domains (PTG), SH3 domains (SH3), serine proteases (SPR), scorpion toxins (STX), thioredoxins (THX) and triose phosphate isomerase (TPI). These families were chosen to cover all four types of protein fold: α (GLT, GLB), β (AZN, FAP, FXN, ISM, SPR, SH3), α/β (THX, TPI, EDE) and $\alpha + \beta$ (CPR, FXN, PTG), as well as one example of a family of so-called small proteins in the SCOP database[33] (STX).

The PDB structure identifiers as well as the lengths of the 68 proteins in these 14 families are listed in Table 1. With the exception of four families, ADH, GLB, PTG and SH3, all the other proteins in the remaining ten families were chosen from the Appendix by Wood & Pearson.[15] Our globin (GLB) data set contains five myoglobins (1bvd, 1lht, 1myt, 1vxb and 5mbn), two hemoglobins (2hbg, 3sdhA) and one leghemoglobin (2gdm). The SH3 data set contains ten proteins, chosen from the SH3 family defined in the SCOP[33] database. The two alkane dehydrogenases are the largest proteins in our dataset. We also included the pair of proteins 1pgb and 2ptl for the PTG family.

Because protein structure alignments are often inaccurate in the absence of significant structure similarity, the 14 families considered here were defined such that all proteins within a family are structurally similar. Our family classification follows the classification by Wood & Pearson,[15] except for their thioredoxin/glutathione transferases, which we divided into two sub-families, thioredoxins (THX) and glutathione transferases (GLT). The need for this division was triggered by the fact that mixed pairs of proteins from THX and GLT show only marginal structural similarity.

Acknowledgements

This work was supported by grants to M.L. from the Department of Energy (DE-FG03-95ER62135) and the National Institutes of Health (GM63817).

References

1. Chothia, C. (1992). One thousand fold families for the molecular biologist? *Nature (London)*, **357**, 543.
2. Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature (London)*, **372**, 631–634.
3. Govindarajan, S., Recabarren, R. & Goldstein, R. (1999). Estimating the total number of protein folds. *J. Mol. Biol.* **35**, 408–414.
4. Wang, Z. X. (1998). A re-estimation of the total numbers of protein folds and superfamilies. *Protein Eng.* **11**, 621–626.
5. Gerstein, M. & Levitt, M. (1998). Comprehensive assessment of automatic structural alignment against a manual standard; the scop classification of proteins. *Protein Sci.* **7**, 445–456.
6. Doolittle, R. F. (1986). *Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences*, Series University Science Books, Mill Valley, CA.
7. Chung, S. & Subbiah, S. (1996). A structural explanation for the twilight zone of protein sequence homology. *Structure*, **4**, 1123–1127.
8. Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94.
9. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.
10. Rost, B. (1997). Protein structures sustain evolutionary drift. *Fold. Des.* **2**, 519–524.

11. Chothia, C. & Lesk, A. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.
12. Flores, T. P., Orengo, C. A., Moss, D. S. & Thornton, J. M. (1993). Comparison of conformation characteristics in structurally similar protein pairs. *Protein Sci.* **2**, 1811–1826.
13. Russell, R. B., Saqi, M. A., Sayle, R. A., Bates, P. A. & Sternberg, M. J. (1997). Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *J. Mol. Biol.* **269**, 423–439.
14. Sauder, J. M., Arthur, J. W. & Dunbrack, R. L. (2000). Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins: Struct. Funct. Genet.* **40**, 6–22.
15. Wood, T. C. & Pearson, W. R. (1999). Evolution of protein sequence and structures. *J. Mol. Biol.* **291**, 977–995.
16. Gronenborn, A. M., Filpula, D. R., Essig, N. Z., Achari, A., Whitlow, M., Wingfield, P. T. *et al.* (1991). A Novel: highly stable fold of the immunoglobulin binding domain of streptococcal protein-G. *Science*, **253**, 657–661.
17. 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.
18. Wikstrom, M., Drakenberg, T., Forsen, S. & Bjork, L. (1994). 3-Dimensional solution structure of an immunoglobulin light chain-binding domain of protein L: comparison with the IGG binding domains of protein G. *Biochemistry*, **33**, 14011–14017.
19. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature (London)*, **358**, 86–89.
20. Banner, D., Kokkinidis, M. & Tsernoglou, D. (1987). Structure of the COL-E1 ROP protein at 1.7 Å resolution. *J. Mol. Biol.* **196**, 657–675.
21. Dalal, S., Balasubramanian, S. & Regan, L. (1997). Protein alchemy: changing beta-sheet into alpha-helix. *Nature Struct. Biol.* **4**, 548–552.
22. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). Scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
23. 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.
24. Zuckerlandl, E. & Pauling, L. (1965). *Evolutionary Divergence and Convergence in Proteins*, Series Academic Press, New York.
25. Doolittle, R. (1981). Similar amino acid sequences: chances or common ancestry. *Science*, **214**, 149–159.
26. Bernstein, F. C., Koetzle, T. F., Williams, G., Meyer, D. J., Brice, M. D., Rodgers, J. R. *et al.* (1977). The protein databank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.
27. Koehl, P. (2002). Protein structure similarities. *Curr. Opin. Struct. Biol.* **11**, 348–353.
28. Lattman, E. E. & Rose, G. D. (1993). Protein folding—what's the question? *Proc. Natl Acad. Sci. USA*, **90**, 439–441.
29. Koehl, P. & Levitt, M. (1999). *De novo* protein design. I. In search of stability and specificity. *J. Mol. Biol.* **293**, 1161–1181.
30. Koehl, P. & Levitt, M. (1999). *De novo* protein design. II. Plasticity in sequence space. *J. Mol. Biol.* **293**, 1183–1193.
31. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
32. Eddy, S. R. (1996). Hidden Markov models. *Curr. Opin. Struct. Biol.* **6**, 361–365.
33. Koehl, P. & Levitt, M. (2002). Protein topology and stability define the space of allowed sequences. *Proc. Natl Acad. Sci. USA*, **95**, 1280–1286.
34. Lee, J., Scheraga, H. & Rackovsky, S. (1997). New optimization method for conformational energy calculations on polypeptides: conformational space annealing. *J. Comp. Chem.* **18**, 1222–1232.
35. Lee, J., Liwo, A. & Scheraga, H. (1999). Energy-based *de novo* protein folding by conformational space annealing and an off-lattice united-residue force field: application to the 10–55 fragment of staphylococcal protein A and to apo calbindin D9K. *Proc. Natl Acad. Sci. USA*, **96**, 2025–2030.
36. Koehl, P. & Delarue, M. (1994). 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.
37. Koehl, P. & Delarue, M. (1994). Polar and non-polar atomic environment in the protein core: implications for folding and binding. *Proteins: Struct. Funct. Genet.* **20**, 264–278.
38. Pande, V. S., Grosberg, A. Y. & Tanaka, T. (1997). Statistical mechanics of simple models of protein folding and design. *Biophys. J.* **73**, 3192–3210.
39. Shakhnovich, E. I. & Gutin, A. M. (1993). A new approach to the design of stable proteins. *Protein Eng.* **6**, 793–800.
40. Shakhnovich, E. I. & Gutin, A. M. (1993). Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl Acad. Sci. USA*, **90**, 7195–7199.
41. Henikoff, S. & Henikoff, J. G. (1992). Amino-acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
42. Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946–950.

Edited by J. Thornton

(Received 1 February 2002; received in revised form 1 July 2002; accepted 5 September 2002)