doi:10.1016/j.jmb.2004.11.062 J. Mol. Biol. (2005) 346, 631–644

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Recapitulation of Protein Family Divergence using Flexible Backbone Protein Design

Christopher T. Saunders[1] and David Baker[2]*

[1]*Department of Genome Sciences, University of Washington, Box 357730 Seattle, WA 98195, USA*

[2]*Department of Biochemistry Howard Hughes Medical Institute, University of Washington, J-567 Health Sciences, Box 357350, Seattle WA 98195-7350, USA*

**Corresponding author*

We use flexible backbone protein design to explore the sequence and structure neighborhoods of naturally occurring proteins. The method samples sequence and structure space in the vicinity of a known sequence and structure by alternately optimizing the sequence for a fixed protein backbone using rotamer based sequence search, and optimizing the backbone for a fixed amino acid sequence using atomic-resolution structure prediction. We find that such a flexible backbone design method better recapitulates protein family sequence variation than sequence optimization on fixed backbones or randomly perturbed backbone ensembles for ten diverse protein structures. For the SH3 domain, the backbone structure variation in the family is also better recapitulated than in randomly perturbed backbones. The potential application of this method as a model of protein family evolution is highlighted by a concerted transition to the amino acid sequence in the structural core of one SH3 domain starting from the backbone coordinates of an homologous structure.

Keywords: flexible backbone protein design; protein evolution; protein structure prediction; sequence space; SH3 domain

Introduction

The evolution of protein families is subject to numerous constraints unified by a simple necessity: the preservation and elaboration of biological function. To represent this evolutionary process a variety of models have been developed which provide insight into relationships among proteins and inference of biological properties for uncharacterized proteins.

For pragmatic reasons, the most popular models of protein evolution use residue sequences without explicitly representing protein structure. Such methods are both efficient and useful for assigning homologous relationships between proteins, and mature instances of these techniques, such as PSI-BLAST[1] and HMMer,[2] form the foundation of modern protein bioinformatics. While these models are effective, they can be improved by explicitly representing protein structure. The inclusion of structure in such models is motivated by the constraint that proteins must fold into a structure which ultimately contributes to its function. For this reason, structure tends to be highly conserved as proteins evolve, which allows homologous relationships to be detected between protein structures even when they cannot be found by sequence alone.[3]

Several methods have been developed to augment sequence-based evolutionary models with structural information; often referred to as fold-recognition or "threading" models, these typically represent protein structure at the residue level.[4–6] A fold-recognition model could, for example, modify the probability of an homologous relationship according to the manner in which a sequence aligns to the pattern of residue burial in the structure of a putative homolog. Despite low resolution and the use of fixed structural representatives, these models can enhance sequence-based search sensitivity.

Given the success of this approach, a subsequent step is to use more detailed, atomic-resolution methods to represent the evolution of protein families. Such techniques do present non-trivial computational problems, however these have been partially addressed by methods developed to

Abbreviation used: RMS, root-mean-square α-carbon deviation.

E-mail address of the corresponding author: dabaker@u.washington.edu

design protein sequences that fold into specific backbone conformations. These protein design methods typically use an atomic-scale structural model to build alternate residue side-chains onto a fixed protein backbone, searching for sequences of residues and corresponding side-chain conformations which form the most stable structures according to a rapidly computable potential function. The efficacy of such methods has been experimentally verified in numerous cases by designing sequences onto the backbones of known structures,[7–9] minor variants of known backbones,[10–12] and backbone conformations not previously observed in nature.[13]

Although these protein design techniques do not explicitly represent an evolutionary process, they model certain constraints on the evolution of natural proteins, such as the thermodynamic constraint that proteins fold to a stable conformation, and the functional constraint that this folded conformation be conserved as the protein evolves. For this reason, protein design can be regarded as a simplified model of protein evolution, and we thus expect that the sequences of the natural protein family represent a subset of the sequences predicted by protein design for the corresponding backbone conformation.

Interest in this similarity between designed and natural sequences has led to several studies evaluating their relationship. These previous studies of protein design as an evolutionary model have either kept the backbone rigid[14–16] or used randomly perturbed backbone ensembles.[17] Due to the small adjustments made in the backbone structure of natural proteins as they evolve, the manner in which design models represent such backbone movements has a potentially large impact on the ability of the model to recapitulate those aspects of protein evolution which are dominated by thermodynamic constraints. We recently described a flexible backbone protein design strategy in which both the sequence and structure evolve by alternately optimizing the amino acid sequence for a fixed backbone structure, and optimizing the structure for a fixed amino acid sequence (the latter is the classical structure prediction problem). This strategy was used to design a protein with a novel fold which was subsequently shown in experimental biophysical and X-ray crystallographic studies to be exceptionally stable and very close in structure, 1.2 Å root mean square (RMS) α-carbon deviation to the design model.[13] Here we investigate the potential of this flexible backbone protein design strategy to model protein family sequence and structure divergence. We show that a design protocol incorporating backbone flexibility by means of an iterative sequence and structure optimization cycle significantly enhances our ability to recreate the sequence diversity of natural families and that structural variations of close natural homologs can be sampled in some cases as well.

Results

The flexible backbone design procedure we use can, in principle, sample all stable protein sequence/structure pairs for a given length starting with a naturally occurring protein backbone, and coupled with sequence insertion and deletion operators could sample all of the naturally occurring homologs for a large protein family. As the sampling method searches for energy minima without any knowledge of functional constraints, the set of sequences generated by such a process would correspond to a superset of the naturally occurring family. Hence, in addition to providing a model of protein evolution under purely thermodynamic constraints, comparison of the natural sequence family to the simulated sequences could highlight functional selection within the protein.

With these ideas in mind, we began by experimenting with the iterative sequence/structure optimization method used to create Top7, a protein designed with a novel backbone topology. In evaluating the ability of this method to recapitulate sequence and structural divergence in naturally occurring protein families, we found that the structures it predicted stayed relatively close to the structure from which the predictions were derived, and that the method had a limited ability to recapitulate the sequence and structure variation of natural protein families (data not shown). We therefore sought to improve approximations in the model expected to influence natural protein family recapitulation, while preserving the iterative sequence/structure optimization approach as our basic flexible backbone design strategy.

Sequence design improvement

Using native sequence recovery as a figure of merit, we sought to improve our sequence search for the complete sequence redesign of a test set of 42 small protein domains. The native sequence recapitulation resulting from this test is summarized for several sequence search methods in Table 1. The extent to which native sequences are recovered in protein design calculations on native protein backbones is a useful, albeit approximate measure of design performance, because we expect that a significant fraction of the residues native to a given protein backbone are thermodynamically optimal for that backbone, especially in the core of the structure. This expectation is supported by experimental mutagenesis studies which have shown that mutations of protein core residues are usually destabilizing.[18,19] We also expect that our sequence design methods should more accurately recapitulate the residues in the protein's structural core because the surface residues are largely constrained by solvation, side-chain entropy and biological function, all of which are either poorly approximated or absent in our model. For this reason, we have separately summarized the recapitulation of core residues in Table 1 and

Table 1. Percent identity between native and designed amino acid residues

	% Recapitulation of native amino acid identity	
	All	Structural core
(a) Initial method	32.1	45.6
(b) Method (a) and robust parameterization	33.0	47.7
(c) Method (b) and potential, search and side-chain library modifications	35.1	52.4
(d) Method (c) and extra χ_1 subrotamers	35.9	55.4
(e) Method (c) and extra χ_1 and aromatic subrotamers	37.0	57.1

Recent improvements made to fixed backbone sequence search methods, as approximated by amino acid identity between native and designed sequences calculated over a diverse set of 42 protein backbone structures. Design method (a) is very similar to a previously described protocol.[13] Method (b) incorporates a new parameterization of the energy terms in the potential which optimizes the placement of the amino acid on the protein in both the native and designed environments. Method (c) includes a large number of rotamer, potential and search modifications. Briefly, these include an update of our rotamer database provided by Dunbrack and co-workers,[22] a more accurate procedure for culling very low probability configurations from the library, a statistical approximation of π–π interactions and smoothing all of the energy terms in the potential. Part (d) shows the result of using method (c) with additional rotamers perturbed from each canonical rotamer by ± 1 standard deviation about the χ_1 angle. Part (e) shows the result of using method (c) with our production set of additional rotamers, which includes rotamers perturbed from each canonical rotamer by ± 1 standard deviation about the χ_1 angle, as well as additional perturbations for aromatic residues (see Methods).

we focus on improving the recapitulation of these residues.

As described in Methods, the energy parameterization procedure was modified to partially account for the dependency between the parameterization of the potential and the packing environment used to calculate this potential. This change resulted in a small but robust improvement in both full and core sequence recapitulation, shown in Table 1, part (b). We have also updated the backbone dependent rotamer library, improved our rotamer strain definitions, stabilized the convergence of the sequence search procedure and incorporated a number of refinements to energy function components (see Methods). These changes have resulted in significant additional improvement to the core residue recapitulation of our test set, as is shown in Table 1, part (c).

In practice, expanding the rotamer set used during the sequence search to include subrotamers is an effective means of modeling strained rotamer conformations; we have observed that such subrotamers have a significant effect on sequence prediction. For this reason, a number of improvements were made to the rotamer interaction energy storage method and the low-probability rotamer culling which allowed more subrotamers to be used with the same computational resources. The effect of additional subrotamers is apparent in the results shown in Table 1; using the design protocol introduced in part (c) of this Table, the addition of subrotamers with ± 1 standard deviation about the χ_1 angle of each rotamer improves the native residue recapitulation by 0.8% overall and 3.0% in the core. The design calculations used in this study include an additional expansion of the χ_1 and χ_2 angles for aromatic residue rotamers which results in an additional benefit to the native sequence recapitulation of 1.9% overall and 4.7% in the core.

Backbone search improvement

To model the structural transitions typical of close homologous proteins, we sought simple tests of such transitions for our structure relaxation methods. One such case that has proven useful in refining our methods is an SH3 structure from c-Crk (1cka:A) and a close homologous structure (1awx). We focused on using the backbone of one structure with the residue sequence of the other "threaded" onto it as a test of our ability to shift between homologs. In doing so, we found that it was initially possible to model the transition of the SH3 distal and RT-loops between the two test cases, albeit without energy discrimination; however, the complex transition of the n-src loop between these structures proved more difficult to model. Therefore, we introduced an initial high-temperature Monte Carlo search procedure to sample a larger structure space around the starting structure, as well as allowing the insertion of highly divergent protein fragments followed by compensatory changes in adjacent residues to more aggressively search local conformational space (see Methods). As a result of these search modifications, we sampled a larger conformational space around the starting structure, and thus required improved discrimination of energy minima. For this reason we expanded and improved the fixed geometry energy minimization, as described in Methods, resulting in the detection of lower energy minima by our structure relaxation routines.

The collective effect of these changes was an improvement to our close homolog transition test case, shown in Figure 1. This example demonstrates a transition starting from the backbone of 1awx with the residue sequence of 1cka:A threaded onto it. Our improved structure relaxation procedure not only modeled the transition of the n-src loop, which was not previously possible, but also lowered the RMS to the native structure as a whole. While this case demonstrates an improvement in our sampling of homologous conformational changes, we have not been able to completely discriminate such cases by energy. The decoy shown in Figure 1 was selected for low RMS to the C-terminal region of the native structure, because we were interested in optimizing our structure search techniques such

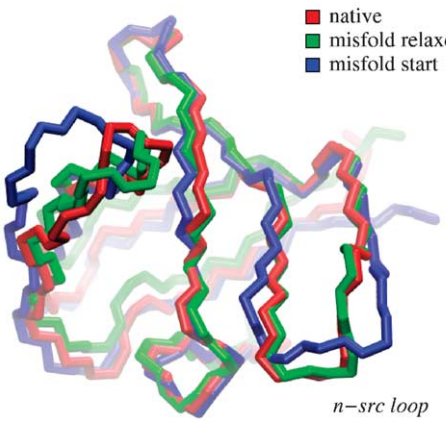

Figure 1. Modeling the structural transition of the n-src loop between SH3 homologs. We constructed a test case using two closely homologous SH3 structures: 1cka:A and 1awk, which share 40% sequence identity and are structurally separated by 1.47 Å RMS. We threaded the sequence of one protein (1cka:A) onto the structure of the second (1awk) and relaxed the threaded structure in an attempt to recover the native conformation, especially for the n-src loop. The backbone of the minimized structure for 1cka:A is shown in red, the backbone of 1awk is in blue, and the green backbone corresponds to a threaded structure following relaxation, selected for low RMS to the C-terminal region of the native structure. The RMS between the native structure and this decoy is 1.13 Å, relaxed from 1.47 Å. For the difficult C-terminal region, including the n-src loop (residues 26–54) the RMS between the native and decoy structures is 0.55 Å, relaxed from 1.19 Å. As discussed in Results, although this decoy was selected for structural similarity to the native structure, it represents an improvement to our structure search methods by modeling the native loop transition even without complete energy discrimination, and it is also found in the lowest 5th percentile by energy of 600 decoys generated, thus it may be possible to discriminate such homologous transitions with further refinements to the design potential.

that these types of structural shifts were possible using our structure search methods. However, this structure is in the most stable 5th percentile of 600 decoys; it is thus our expectation that further refinements to the potential and perhaps successive searches employing an evolutionary algorithm will lead to automated discrimination of such cases by energy.

Natural family sequence recapitulation

Recapitulation of natural protein designability

After incorporating various sequence and structure search improvements into the flexible backbone design procedure, we studied the recapitulation of natural sequence families for ten diverse protein structures and compared this performance to two alternate design methods previously discussed in the literature. The first of these is a fixed backbone design procedure which searches for low energy protein sequences compatible with the exact backbone of the starting structure. In this case the repulsive van der Waals energy and sequence search have been modified such that greater sequence diversity can be produced in spite of the fixed backbone restriction. The second procedure searches for the optimal protein sequence for each of an ensemble of backbones randomly perturbed from the starting structure (see Methods). This strategy of incorporating backbone flexibility through randomized structure ensembles is similar to that developed by Desjarlais and used by Larson *et al.* to model natural backbone flexibility in a large-scale design study.[17]

To better understand how each of these methods was able to recapitulate the qualities of natural sequence families, we first examined whether the design methods could recreate the characteristic sequence diversity of each family. Such characteristic sequence diversity, or designability, reflects the size of sequence space compatible with a protein's backbone architecture. We express this diversity by calculating a "seqeunce diversity score" for each domain, which is the exponential of the average residue entropy for each site in a sequence alignment; a value which approximately expresses the average number of residues allowed at each position in the protein (see Methods). This diversity score is used to characterize the natural sequence variability for the ten protein families in our test set (Table 2). The family members were identified using PSI-BLAST, with near duplicate and significantly gapped sequences removed, and the sequence diversity score obtained after weighting to deemphasize large groups of similar sequences (see Methods).

Using each of the three design methods discussed above, we produce alignments of designed sequences for each domain by taking the 100 lowest energy sequences from a total of 300 designs, weighting to deemphasize highly similar groups of sequences, and calculating the sequence diversity score as with the natural sequence alignments. In Figure 3, we plot the natural *versus* designed sequence diversity score for each design method as well as the best fit by linear regression analysis. Of the three methods, flexible backbone design was the only one to produce sequence diversity scores which could account for a portion of the variability in the natural sequence diversity. The linear regression of the sequence diversity score from flexible backbone design to that of natural family members was found to be significant, and has an R^2 value of 0.475. A similar analysis of the sequence diversity score produced by fixed backbone and randomized backbone design did not yield any significant explanation of the natural sequence diversity. Hence, it appears that some component of the natural protein designability can be accounted for by the iterative flexible backbone design method, and that this ability depends on the evolution of the protein backbone under a physical potential during the design process.

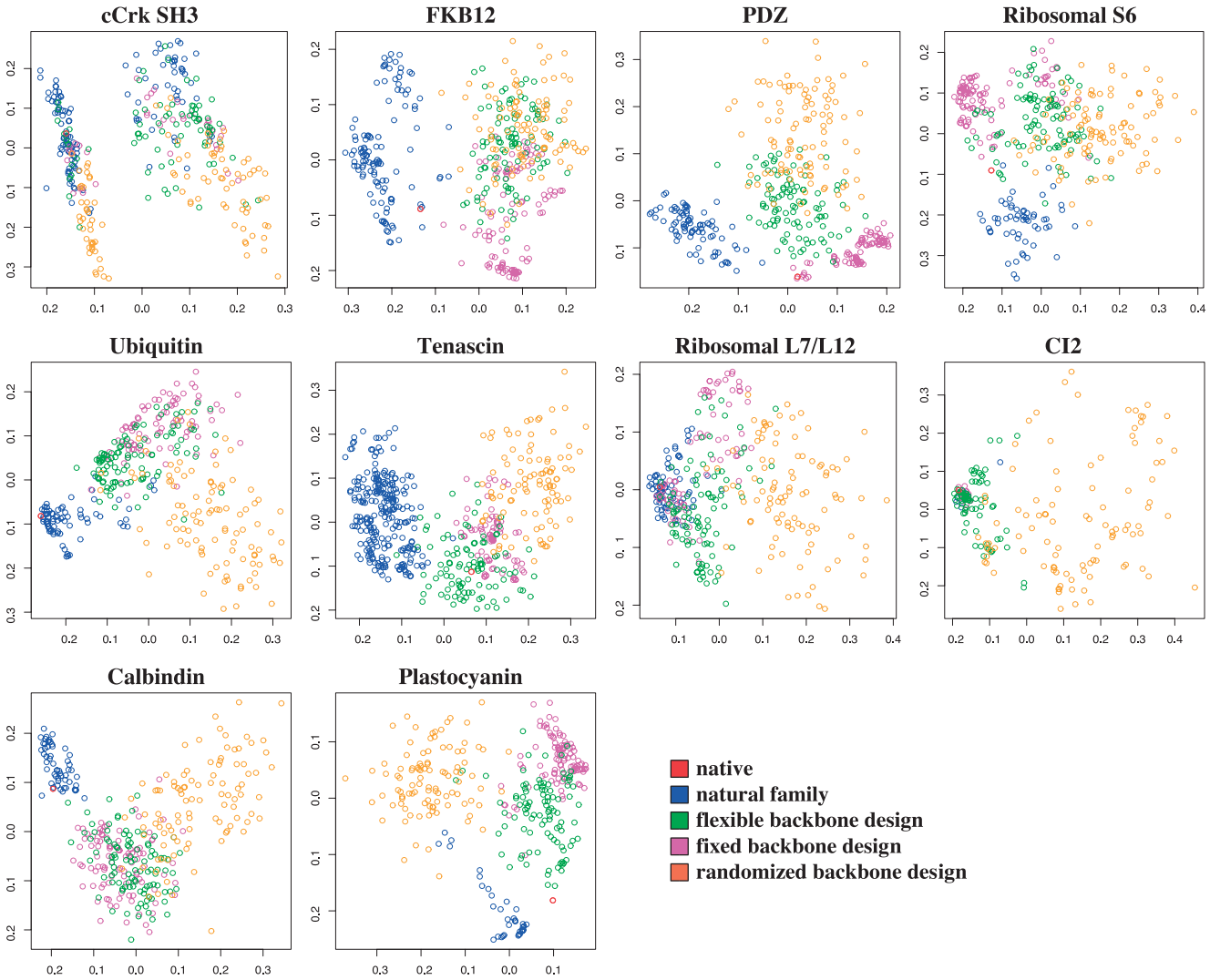

Figure 2. Comparison of designed and natural homologous sequences for each member of the test domain set. An evolutionary sequence distance is calculated between all designed and natural sequences for each family as described in Methods. For all sequences, the distance between the core residues of each domain are represented in a two-dimensional projection using metric multidimensional scaling, which selects the projection axes to maximize the sequence distance variance. In this plot, sequences are represented as points and the separation between points approximates sequence distance; the native sequence is shown in red, natural homolog sequences are blue, flexible backbone design decoys are shown in green and those for fixed backbone and randomized backbone design are shown in magenta and orange, respectively.

Table 2. Protein domain test set

Protein	pdb-id	Residues	Sequence count	Sequence diversity score
c-Crk SH3	1cka:A	134–189	512	5.697
FKB12	1fkb	1–107	356	4.775
PDZ	1qau:A	14–102	244	6.023
Ribosomal S6	1ris	1–94	138	5.255
Ubiquitin	1ubq:A	1–76	342	4.974
Tenascin	1ten	803–890	642	7.723
Ribosomal L7/L12	1ctf	7–74	196	3.051
CI-2	2ci2:I	19–83	110	3.226
Calbindin	4icb	1–76	180	4.651
Plastocyanin	7pcy	1–98	132	4.239

For each domain, the family alignment was found by PSI-BLAST and filltered for deletions and high similarity to the sequence of the test domain to produce the final aligned sequence set, the size of which is shown. To express the diversity of each sequence family, we calculate the exponential of the average site entropy of the sequence alignment after weighting to deemphasize highly similar sequence groups (see Methods).

Recapitulation of natural family amino acid distributions

If the flexible backbone design procedure can partially recapitulate the characteristic level of sequence diversity associated with each protein family, how well can this method recreate the natural family sequence space? To address this question, we first examined the degree of similarity between the amino acid frequency distributions at each site in the designed and natural protein alignments. To do so, we calculated the relative entropy of the designed amino acid distribution compared to the natural family amino acid distribution at every site in each of our test set proteins and averaged this value over all sites in each protein. This average relative entropy is shown for each of the ten test structures in Table 3, together with the average relative entropy of residue positions in the structural core of each protein. The relative entropy approaches zero as the designed and natural amino acid distributions come closer to matching, thus a lower value for the relative entropy indicates a closer fit of the designed to the natural amino acid distribution. In all cases it appears that the flexible backbone design method tends to produce site amino acid distributions closer to those of the natural sequence family than the fixed backbone method, and with only one exception the flexible backbone method is superior to randomized backbone ensemble design as well. These results suggest that the iterative sequence/structure optimization protocol is capable of more accurately recapitulating the naturally observed sequence family than fixed or randomized backbone design methods.

Given that flexible backbone design can replicate natural amino acid site distributions more closely than our alternate design methods, does the same relationship hold for the replication of natural family sequences? Our design methods appear to more accurately replicate the distribution of amino acids at sites in the core of protein structures, where evolutionary pressures are more closely related to the design model. Therefore, we expect that our methods may be capable of replicating correlated residue changes which occur in the protein's structural core as well. To gain a clearer understanding of how the core positions of designed and natural sequences are related, we created two-dimensional projections of the sequence similarities between all designed and natural structural core

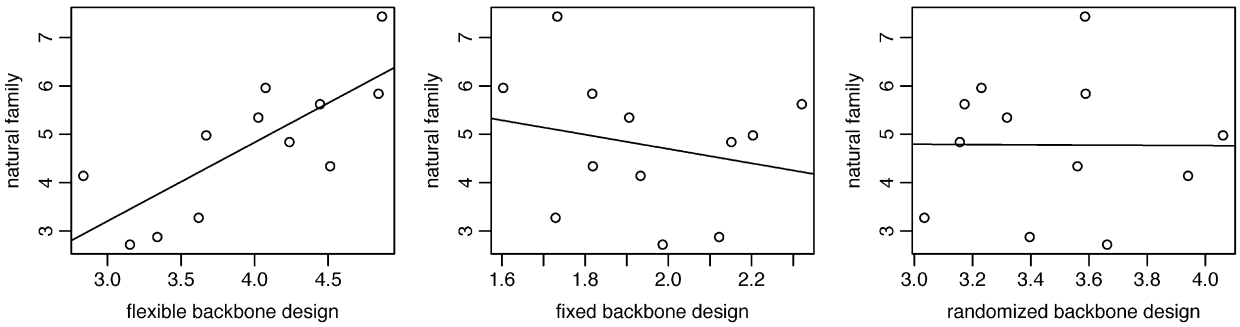

Figure 3. Comparison of natural family and designed sequence diversity. Shown is the best linear fit of all three design methods to the natural sequence diversity, expressed here as the "sequence diversity score", or the exponential of the average site entropy (see Methods). A linear fit of the flexible backbone design sequence diversity score explains the natural family variation with an R^2 value of 0.475; this regression is significant at the 0.05 level ($P = 0.0275$). No significant linear relationship is found for the equivalent regression using sequence diversity scores generated from fixed backbone design ($P = 0.5235$) and randomized backbone design ($P = 0.8586$).

Table 3. Mean amino acid relative entropy between designed and natural sequence alignments

	Mean amino acid relative entropy					
	All			Structural core		
Protein	Flexible	Flxed	Random	Flexible	Fixed	Random
c-Crk SH3	0.98	1.57	1.44	0.07	0.10	0.15
FKB12	0.89	1.64	1.21	0.17	0.24	0.20
PDZ	0.98	1.75	1.32	0.15	0.30	0.21
Ribosomal S6	1.27	1.68	1.31	0.15	0.21	0.16
Ubiquitin	1.19	1.53	1.33	0.12	0.19	0.18
Tenascin	0.91	1.65	1.27	0.14	0.22	0.21
Ribosomal L7/L12	1.18	1.38	1.27	0.16	0.19	0.20
CI-2	1.14	1.75	1.50	0.07	0.09	0.12
Calbindin	1.19	1.55	1.49	0.22	0.25	0.26
Plastocyanin	1.25	1.70	1.22	0.24	0.32	0.24

Amino acid distributions for designed sequences are taken directly from designed sequence alignments for each variant design method, in each case using the most stable 100 sequences of 300 computed, and weighting the designed sequences using the Henikoff algorithm to deemphasize similar sequence groups. Natural amino acid distributions are extracted from the position specific scoring matrix produced by a PSI-BLAST search of the native protein sequence, including pseudo-counts. The relative entropy for each site in the protein family alignment is $-\sum_{aa} p_{aa} \log(p_{aa}/q_{aa})$, where p_{aa} is the frequency of each amino acid in designed sequences and q_{aa} is the frequency in homologous sequences. For each domain we show the average of the relative entropy over all sites and over sites in the structural core of each domain.

sequences for each of the proteins in our test set (Figure 2). From these projections, it appears that both the flexible and fixed backbone design methods model the natural core residue sequences of each domain more closely than randomized backbone design. It is also apparent that, although the flexible and fixed backbone design methods often cover similar areas of sequence space, there are a number of cases, such as for SH3, ribosomal S6, ubiquitin, tenascin and plastocyanin, where a subset of the sequences produced using the flexible backbone protocol more closely resemble the homologous core sequences than the sequences produced using the fixed backbone method.

Although the performance of flexible backbone design in the replication of natural core sequences is encouraging, it is apparent from these sequence projections that this is a more difficult problem than the replication of residue distributions at individual sites. This may be explained by the design method's prediction of sequences that represent analogs or novel representatives of the protein domain, which are not subject to any constraints of function or evolutionary pathway. For this reason, we do not necessarily expect that designs which are dissimilar to the natural core sequences of each domain are invalid. Despite this, not all of these distant sequences encode folded proteins, as found in recent large-scale experimental design efforts, due to inaccuracies and approximations in the design model.[9] These projections also reveal that our flexible backbone design method is not simply increasing the sequence diversity by allowing sequence search to be run on a greater variety of backbones. This does, however, appear to be the case for the randomized backbone method, as sequences produced with this method often occupy a unique area of sequence space without significant overlap to known homologs or flexible backbone design predictions.

Among all the test structures, the flexible backbone design procedure appears to have an exceptional ability to represent the core of the SH3 domain, and we have thus chosen this domain for further study of both the types of structures and core sequence combinations our design procedure is able to predict.

Natural family structure recapitulation

If flexible backbone design can be used to more accurately recapitulate the natural sequence diversity of protein families, we would like to know if it could be used to recreate the natural structure variation of these families as well. This question is somewhat more problematic than its sequence counterpart, largely because less structural data are available and it is more difficult to objectively define similarity between structures.

Due to the encouraging core sequence design results and large number of structures available for SH3, we examined the extent of structure recapitulation by our flexible backbone design process for this family. We identified natural analogs of the starting SH3 structure (1cka:A) and aligned them using CE; this alignment is plotted for visual comparison with alignments of design decoys in Figure 4(a). This shows that the iterative sequence/structure optimization method recreates some basic features of natural analogs: specifically, accumulation of most structural variation in the surface loops of the protein. To gain a more quantitative view of structural similarity we evaluated the closest RMS decoy to each natural analog and show the distribution of this value over all analogs for each design method in Figure 4(b). For the fixed backbone design case, we examined the distribution of RMS between each natural analog and the starting backbone conformation. All alignments between analogous structures and designed decoys

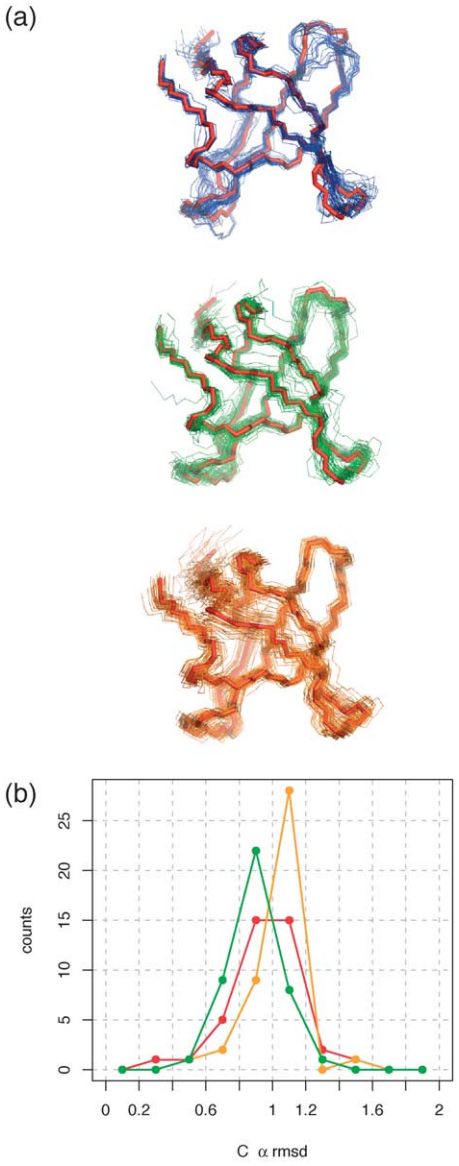

Figure 4. Comparison of designed and natural structures for the SH3 domain. (a) A least-squares alignment of designed and analogous backbones to the starting structure (1cka:A) is shown. The backbone of the starting structure is shown in red and the closest naturally occurring structures are shown in blue, these are the top 40 structures by CE alignment *Z*-score[37] from all proteins in the pdb. In green and orange are the 40 lowest energy decoys generated, respectively, by flexible backbone and randomized backbone ensemble design. (b) For each naturally occurring SH3 structural analog, the RMS of the closest design structure is found, and the distribution of this closest RMS value over all analogs is plotted for each design method, with flexible backbone design shown in green, randomized backbone ensemble design shown in orange and fixed backbone design (or simply the distance to the native structure) shown in red.

were made with CE, and the RMS values calculated from this alignment were used to generate the distributions shown in this Figure. It is apparent that the decoys generated by iterative sequence/structure optimization are distributed closer to the analogs of the starting backbone than the structure from which they started, and that this increased similarity to analogous structures cannot be recreated by small random backbone perturbations. This result is consistent with our observations of designed SH3 core sequences, which show that designs made to randomized backbone ensembles have less similarity to natural family sequences than those made with iterative sequence/structure optimization.

Modeling an exact homologous core transition

Due to the relatively strong sequence and structure recapitulation observed for the SH3 domain using flexible backbone design, we investigated individual decoys to better understand the source of this similarity. This revealed an interesting case wherein an exact recapitulation of the core residue sequence was made for one structure (1sem) starting from the backbone structure of its homolog (1cka:A). By our definition, there are eight residues in the core of the SH3 domain, three of which are mutated between the sequences of these two proteins. We show a structural alignment of the three mutated residue positions in Figure 5, comparing the residue positions of the starting structure (1cka:A), its natural homolog (1sem), and the designed structure which correctly replicated the core sequence of this homolog. It is evident from this Figure that the backbone shifts among these core residues are relatively small, yet it is probable that these play an important role in the replication of the homologous residue pattern as the sequence design method is unable to replicate this transition when building side-chains onto the fixed starting structure. An examination of the decoy backbone shows that a shift of the backbone away from the core of the domain around position 49 made the I49F substitution sterically feasible. Notably, this core arrangement was found without using the sequence of the starting structure or any of its homologs, and without an explicitly evolutionary model. Thus it is significant that the model was able to replicate all eight core residues of this domain, which included sampling backbone modifications that allowed the model to replicate the three residue substitutions between the starting structure and its homolog.

Discussion

The results presented here show that our design methods recreate certain aspects of protein evolution. We find that a flexible backbone design protocol using iterative sequence and structure optimization appears to sample protein family sequence diversity more accurately than fixed backbone design and better recapitulates the site residue distributions from natural family sequence alignments. Moreover, this improvement does not

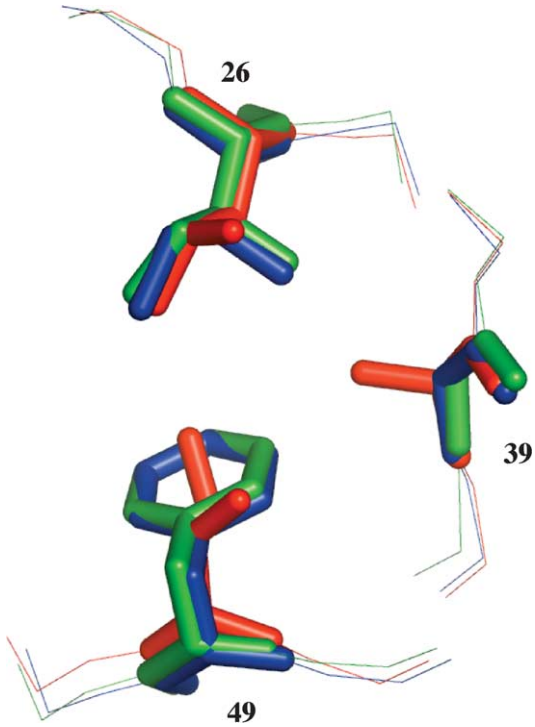

Figure 5. Prediction of the amino acids in the core of one SH3 structure (1sem) by flexible backbone design starting from the backbone structure a second, homologous SH3 structure (1cka:A). The starting structure (1cka:A) is shown in red, the replicated structure (1sem) is shown in blue and the design decoy is shown in green. The residues shown here are the three core residues which have mutated between the two homologous SH3 structures and were correctly predicted in the design decoy shown; a total of eight core residues exist in this structure by our definition and the other five were also replicated by the decoy, thus all core residues of 1sem were predicted starting from the backbone of 1cka:A. On inspection, it seems plausible that the residue substitution at 1cka:A position 49 (I49F) was replicated by the design model because of steric accommodation from backbone movement.

appear to be explained by chance variation in the protein backbone, as shown by a test of sequence designs made to randomized backbone ensembles. Furthermore, in examining projections of sequence space for the structural core of our test proteins, we observe that in many cases iterative flexible backbone design produces sequences which replicate the natural sequence variation more closely than the alternate design methods we have discussed. In one case, our flexible backbone design method predicted the exact residue identities in the core of one SH3 sequence utilizing only the backbone coordinates of a second, homologous SH3 structure, a promising indication of this method's potential as an atomic scale model of protein evolution. It also appears that our methods can recreate the patterns of structural variation in protein families over short evolutionary distances with greater accuracy than simple backbone randomization.

While our efforts have focused on comparing the performance of flexible and fixed backbone design in the recapitulation of protein families, a number of studies have previously examined fixed backbone protein design as an evolutionary model,[14–16] and found that this method produces sequences that are significantly more like the native sequence than random, especially in the core of the structure. Although the relationship of designed sequences to the native sequence is clear in these studies, the relationship to natural family sequences is less certain. For instance, Raha and co-workers demonstrated a significant profile score of designed sequences aligned to protein family profiles; however, it is unclear whether this was due to recapitulation of the native sequence or reflected sequence patterns in the family. Koehl *et al.* used profiles of sequences created with a fixed backbone design method for the TIM fold to search for natural homologs and found that the natural family members could be recovered with a significant Z-score when conducting a profile search using designed sequences. Although this is an indirect test, Koehl's results show that the structural information used to generate the designed sequences could be recapitulating the natural family patterns. Larson *et al.* used a method which emulates the small backbone variations observed among natural family structures by generating randomized backbone ensembles from a known structure and designing sequences onto these structures, a procedure similar to the randomized backbone ensemble control we use in this study. This design procedure was evaluated using a homology search test similar to that discussed by Koehl and co-workers. In this case, a large set of 264 structures was used, and it was found that the natural homologs could be recovered using profiles of designed sequences for roughly half of these proteins. In addition, the search coverage using randomized backbone ensembles was higher than for fixed backbone design.[17] These results suggest that some form of flexibility is necessary in the design process to advance from recapitulation of the native sequence to that of the native sequence family, a finding which is echoed here by the improved recapitulation of protein family diversity using our flexible backbone design protocol compared to fixed backbone design method.

While our design methods implicitly reflect certain aspects of protein evolution, they could be adapted in a number of ways to more explicitly model natural evolutionary phenomena. The design process could be restrained to produce sequences within a certain evolutionary distance of a starting sequence, and evolutionary pathways could be modeled using this technique by allowing the design process to iteratively explore sequence space in progressive steps restrained to this distance. The constraints of known biological function could be represented in the design process as well, most easily for known protein and ligand binding, in which case the structure of the binding

partner could be represented explicitly in the model. Functional constraints could also be represented by taking a hybrid approach: sampling the residue distribution of natural family alignments in critical surface regions while designing the rest of the sequence with a thermodynamic model. It is also possible to make the structural evolution more realistic by including residue insertion and deletion operators to represent greater structural diversification than is possible with the fixed-length design process.

The applications for such evolutionary design methods are quite diverse. We have discussed previous work, which uses protein design to search for homologous relationships between proteins, to infer biological properties of the protein domain as a whole. It is also possible for these same design techniques to be used for the inference of functional regions within the protein, by predicting the patterns of residue variation expected under thermodynamic constraints and comparing these to the sequence variation of the natural protein family. Interestingly, useful progress in both of these applications has recently been reported using a residue-scale design model,[20] indicating the potential of more precise models in these areas. As these evolutionary phenomena are addressed, flexible backbone methods may further enhance our understanding of protein evolution and the role that thermodynamic stability played in selecting the modern natural protein universe, extending the considerable progress already made in this area.[21] Perhaps the most useful future application of flexible backbone design will be engineering novel structures: the recapitulation of natural protein family sequence and structure distributions will undoubtedly complement experimental results in this endeavor by providing rapid methodological feedback, helping to free a multitude of medical and bioengineering applications from the constraints of known protein structures in the course of pursuing customized function.

Methods

Design algorithm

The basis of our flexible backbone design methods have been recently described by Kuhlman *et al.*,[13] to which a number of modifications were made to improve the recapitulation of natural protein families. Both the sequence optimization and structure relaxation procedure have been improved, as described below.

Sequence optimization

For sequence design we employ a quenched Monte Carlo search of rotamer space seeking a low energy rotamer configuration according to a pairwise decomposable potential function. We have made a number of improvements to the previously reported implementation of this approach.

Rotamer library. A rotamer library calculated from an improved version of the method described by Dunbrack & Cohen[22] was incorporated into our sequence search algorithm. Among the improvements made to this library by Dumbrack and co-workers are the introduction of a prior distribution for the variance so that reasonable estimates can be made for undersampled cases, filtering of high *B*-factors and clashes in the input data, and optimization of flip states for asparagine and histidine χ_2 angles, as well as glutamine χ_3 angles, with a consequent treatment of these angles in 360° (no longer assuming symmetry).[23,24] It should also be noted that more structural data was available to calculate this new rotamer database, which should yield more accurate rotamer statistics. We use the backbone dependent rotamer library calculated in May 2002, from the Dunbrack lab rotamer website†.

We implemented a number of changes in our sequence search methods to take advantage of new library features. Among these are the inclusion of χ_3 and χ_4 angles in the calculation of rotamer strain, due to the availability of standard deviations for these angles, as well as using backbone dependent standard deviations for all χ angles in this calculation. We also filter for rare rotamers with greater accuracy than our previous method, by filtering based on the joint probability of all χ angle states conditioned on ϕ and ψ, instead of using only χ_1 and χ_2 to make this decision.

Pair potential. The statistical pair potential used to approximate electrostatics was calculated from an updated set of high resolution crystal structures using the distances between the center of mass of the polar groups for each side-chain, with appropriate pseudo-counts for rare cases. The role of this pair potential was expanded to crudely represent π–π interactions by including all interactions between aromatic side-chains in addition to polar side-chain interactions.

Rotamer interaction energy storage. A method to more efficiently store rotamer interaction energies was developed which eliminates any computational storage of interaction energy between non-interacting amino acid residues, while retaining a constant look-up time for rotamer interaction energies, which is critical for efficient execution of the Monte Carlo search procedure. The increased storage efficiency allowed for greater subrotameric detail to be represented on the protein domains we have studied using relatively low-cost computational hardware.

Forcefield parameterization. The forcefield parameterization used here is an extension of the technique originally described by Kuhlman,[14] wherein the energy terms are set to maximize the probability of the native amino acid in its native

† http://dunbrack.fccc.edu/bbdep

packing environment for all residue positions in a large set of protein structures. The extension to this method attempts to compensate for the approximation of using the native packing environment at each residue position. For consistency, one would like the potential to maximize the probability of the native amino acid at each position in the protein when the surrounding environment was designed using the same potential; however, this is not computationally feasible. We therefore take an iterative approach, such that the terms of the potential are first set to maximize the probability of the native amino acid in a fixed packing environment, then the protein sequence is redesigned using this new potential; the side-chain packing from this sequence redesign is then used as the fixed packing environment in the next iteration to reweight the potential.

A second minor departure from the original parameterization technique described by Kuhlman is to optimize not only the recovery of native sequences but also the overall native amino acid composition. We defined an amino acid distribution error term that reflected how far the designed residue distribution was from the distribution of residues in the proteins used to train the potential. The error term is the sum over all residues of the squared difference between the designed and test set frequencies for that residue. The inverse of this term was empirically scaled and added to the primary term used to train the potential: the log probability of the native amino acid in a fixed packing environment, and this combined term was maximized to find the potential.

Using such an iterative technique, we found that the potential converged after the third iteration. However, no significant improvement in the recapitulation of the native protein sequence was observed after the second iteration, therefore a single round of redesign was used to generate the potential for this study. We implement this parameterization procedure using a set of 46 diverse small protein domains. This procedure is expected to yield a potential that is more robust to certain approximations of the model, such as the discretization of side-chain conformational space and physical constraints that are poorly represented in the potential.

Subrotameric states. A large number of subrotameric states were included in this model to represent strained side-chain torsion angles. For the design calculations in this study the subrotamers included ± 1 standard deviation for each χ_1 angle and an additional ± 0.5 standard deviation about each χ_1 angle of each aromatic residue in combination with ± 1 standard deviation for each χ_2 angle of each aromatic residue. This expanded rotamer set represents the highest detail with which we can model any typical 100 residue protein using less than 500 megabytes of storage for the sequence design procedure.

Sequence search stringency. With the goal of consistently reaching lower energy minima during sequence search, the sequence search annealing schedule was slowed and allowed to reach a lower final temperature than that used previously. These parameters were empirically adjusted such that for a typical small (60-residue) protein, the same energy minima would be reached from a random starting point greater than 50% of the time.

Structure optimization

All decoys were initially subjected to a high-temperature Monte Carlo melting procedure in which the structure relaxation protocol (random torsion moves and insertions of three residue segments of known protein backbones) was implemented on a lower resolution model of the structure wherein side-chains are represented by a single center of mass pseudoatom. This high-temperature simulation continued until the backbone RMS to the starting structure was equal to an amount uniformly selected from 0 Å to 4 Å at random, followed by iterations of sequence design and structure relaxation as described.

We use Monte Carlo minimization[25] as the basis of our structure relaxation method. Here we have made several modifications to this method with the goal of improving the representation of transitions between close homologs.

ω Angle search. An ω angle strain energy was incorporated into the relaxation potential, assuming a Gaussian distribution for this torsion angle with a mean of 179° and standard deviation of 5.6°, as derived from high-resolution crystal structures.[26] This allowed small random ω angle variations to be searched during the relaxation procedure.

Complete torsional minimization. An enhanced rigid geometry minimization of the entire protein heavy atom torsion space was implemented for the relaxation procedure applied during flexible backbone design. This required the addition of all heavy atom χ angles and backbone ω angles to the backbone ϕ and ψ angles used in previous implementations of the relaxation procedure. The energy gradients for the entire torsion space were calculated, as in previous versions, using the efficient recursive calculation methods of Gō and co-workers.[27] To improve the efficiency of energy minimization in this expanded torsion space, the torsional derivatives were modified to include all terms used in the potential, including a novel technique to find the analytic derivatives of the orientation dependent hydrogen-bond term.[28] In a complementary modification, all terms in the potential were smoothed to have a finite first derivative with respect to all free torsion angles, which prevented the minimization from becoming trapped on artifactual roughness in the energy surface. This torsional minimization is applied with progressively increasing stringency over the course of the relaxation procedure, leading to the detection of significantly lower energy minima.

Transitional search moves. Several fragment insertion search moves were modified such that they

would more aggressively search for alternate local structures by attempting to insert protein fragments into a structure that would cause the greatest disruption of the global structure, followed by compensatory changes in adjacent residues as in the "wobble" method previously described by Rohl *et al.*[29]

Variant design methods

We study and compare three design methods which differ primarily in their treatment of the protein backbone. For each of these variants, we calculate 300 designed proteins for each test domain and select from among this set the 100 sequences which are the most stable according to the design potential. In all design methods, the sequence optimization step is started from a random residue sequence, without using the native sequence information from the starting structure.

Flexible backbone design. By flexible backbone design, we refer to the procedure of iterative sequence and structure optimization derived from the method of Kuhlman *et al.*,[13] with the addition of the sequence and structure search modifications enumerated above. Each design run starts from the input structure without use of the native residue sequence and by iterative sequence and structure optimization searches for a low energy backbone structure and residue sequence combination. This is a stochastic process that is expected to predict a different low-energy sequence/structure combination in each run.

Randomized backbone ensemble design. Randomized backbone ensemble design was adapted from the methods of Desjarlais and co-workers;[30] the backbone torsion angles of the starting protein are randomly perturbed by $\pm 5°$ followed by a Monte Carlo procedure which executes a series of random backbone torsion angle moves with the goal of reducing the RMS to within 1 Å of the starting structure. If this goal is reached, then the standard sequence search procedure is executed on the randomly perturbed backbone.

Fixed backbone design. The fixed backbone design procedure is identical with the sequence search step used during flexible backbone design, except that the repulsive van der Waals energy and search stringency have been modified such that reasonable sequence variation can be produced from multiple runs of the algorithm. The van der Waals repulsive energy is represented by a standard 12-6 Lennard Jones potential except there is a cutoff distance below which the repulsive energy is extrapolated linearly to lower the energy of close-contact repulsion. For flexible and randomized backbone design this cutoff is set to 0.6 times the sum of the van der Waals radii for two atoms, for the fixed backbone protocol, a cutoff of 0.86 times the sum of the radii is used instead.

Natural family alignments

For each test set protein, a natural family alignment was made using PSI-BLAST; aligned sequences with greater than 5% deletions relative to the test domain were removed, as well as sequences with greater than 99% sequence identity to the test domain sequence. The sequence diversity score was derived from the resulting natural family sequence alignment after weighting all sequences by the Henikoff position-based weighting algorithm.[31]

Sequence diversity score

Given an alignment of natural homologs or designed sequences for any protein, we express the characteristic sequence diversity of the alignment by calculating the site entropy s_i of the residue distribution at each site i in the sequence alignment as: $s_i = -\sum_{aa} p_{i,aa} \log(p_{i,aa})$, where $p_{i,aa}$ refers to the frequency of each amino acid at site i, after accounting for sequence weights. The sequence diversity score is the average of the site entropy over the entire protein alignment, expressed in exponential form so that it conveniently approximates the average number of amino acid residues tolerated at each site in the alignment. This diversity score is quite similar to the "sequence entropy" used in previous work by Larson *et al.*[32]

Core residue definition

For each test domain, we defined the set of core residues as those residue positions which had a 10 Å C^β density greater than 20. The C^β density is a residue contact score which measures the number of C^β atoms within a 10 Å radius of the C^β atom for the residue in question; for glycine, the C^α atom is counted instead.

Sequence distance

To express the evolutionary divergence of sequences, we compute a sequence distance by negating and scaling all BLOSUM62[33] values to the range 0–1, such that they correspond to a substitution cost; and a fixed gap cost of 0.9 is additionally defined. The sequence distance is the average residue substitution cost for any pair of aligned sequences. For all sequence comparisons in this study, natural sequences are aligned to the native sequence of each domain studied using the Smith–Waterman algorithm[34] implemented as a post-processing step by PSI-BLAST. Distances between homologs are found using the transitive alignment of each homolog to the native sequence. Designed sequences are not realigned to the native sequence, rather every residue is considered aligned to its position in the starting structure used for design.

Acknowledgements

The authors thank Brian Kuhlman and Ora Schueler-Furman for helpful comments on the manuscript, and Roland Dunbrack for providing early access to rotamer databases and detailed method descriptions. Figure 1 was prepared using VMD;[35] all other molecular graphics were created with PyMOL.[36] This work was supported by NIH training grant T32 HG00035 and the HHMI.

References

1. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., 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.
2. Eddy, S. R. (1998). Profile hidden markov models. *Bioinformatics*, **14**, 755–763.
3. Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO*, **5**, 823–826.
4. Bowie, J. U., Luthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.
5. Jones, D. T. (1999). Genthreader: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**, 797–815.
6. Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299**, 499–520.
7. Dahiyat, B. & Mayo, S. (1997). *De novo* protein design: fully automated sequence selection. *Science*, **278**, 82–87.
8. Marshall, S. A. & Mayo, S. L. (2001). Achieving stability and conformational specificity in designed proteins *via* binary patterning. *J. Mol. Biol.* **305**, 619–631.
9. Dantas, G., Kuhlman, B., Callender, D., Wong, M., Baker, D. & large, A. (2003). Scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol.* **332**, 449–460.
10. Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T. & Kim, P. S. (1998). High-resolution protein design with backbone freedom. *Science*, **282**, 1462–1467.
11. Ross, S. A., Sarisky, C. A., Su, A. & Mayo, S. L. (2001). Designed protein G core variants fold to native-like structures: sequence selection by ORBIT tolerates variation in backbone specification. *Protein Sci.* **10**, 450–454.
12. Offredi, F., Dubail, F., Kischel, P., Sarinski, K., Stern, A. S., Van de, C. *et al*. (2003). *De novo* backbone and sequence design of an idealized alpha/beta-barrel protein: evidence of stable tertiary structure. *J. Mol. Biol.* **325**, 163–174.
13. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science*, **302**, 1364–1368.
14. Kuhlman, B. & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. USA*, **97**, 10383–10388.
15. Koehl, P. & Levitt, M. (1999). *De novo* protein design. II. Plasticity in sequence space. *J. Mol. Biol.* **293**, 1183–1193.
16. Raha, K., Wollacott, A. M., Italia, M. J. & Desjarlais, J. R. (2000). Prediction of amino acid sequence from structure. *Protein Sci.* **9**, 1106–1119.
17. Larson, S. M., Garg, A., Dejarlais, J. R. & Pande, V. S. (2003). Increased detection of structural templates using alignments of designed sequences. *Proteins: Struct. Funct. Genet.* **51**, 390–396.
18. Eriksson, A. E., Baase, W. A., Zhang, X. J., Heinz, D. W., Blaber, M., Baldwin, E. P. & Matthews, B. W. (1992). Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*, **255**, 178–183.
19. Baldwin, E., Xu, J., Hajiseyedjavadi, O., Baase, W. A. & Matthews, B. W. (1996). Thermodynamic and structural compensation in "size-switch" core repacking variants of bacteriophage T4 lysozyme. *J. Mol. Biol.* **259**, 542–559.
20. Pei, J., Dokholyan, N. V., Shakhnovich, E. I. & Grishin, N. V. (2003). Using protein design for homology detection and active site searches. *Proc. Natl Acad. Sci. USA*, **100**, 11361–11366.
21. Xia, Y. & Levitt, M. (2004). Simulating protein evolution in sequence and structure space. *Curr. Opin. Struct. Biol.* **14**, 202–207.
22. Dunbrack, R. L. & Cohen, F. E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **6**, 1661–1681.
23. Lovell, S. C., Word, J. M., Richardson, J. S. & Richardson, D. C. (2000). The penultimate rotamer library. *Proteins: Struct. Funct. Genet.* **40**, 389–408.
24. Dunbrack, R. L. J. (2002). Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.* **12**, 431–440.
25. Li, Z. & Scheraga, H. A. (1987). Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl Acad. Sci. USA*, **84**, 6611–6615.
26. Network, E.-D.V. (1998). Who checks the checkers? Four validation tools applied to eight atomic resolution structures. *J. Mol. Biol.* **20**, 417–436.
27. Abe, H., Braun, W., Noguti, T. & Gō, N. (1984). Rapid calculation of first and second derivatives of conformational energy with respect to dihedral angles for proteins. general recurrent equations. *Comput. Chem.* **8**, 239–247.
28. Wedemeyer, W. J. & Baker, D. (2003). Efficient minimization of angle-dependent potentials for polypeptides in internal coordinates. *Proteins: Struct. Funct. Genet.* **53**, 262–272.
29. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. (2004). Protein structure prediction using Rosetta. *Methods Enzymol.* **383**, 66–93.
30. Desjarlais, J. R. & Handel, T. M. (1999). Side-chain and backbone flexibility in protein core design. *J. Mol. Biol.* **290**, 305–318.
31. Henikoff, S. & Henikoff, J. G. (1994). Position-based sequence weights. *J. Mol. Biol.* **243**, 574–578.
32. Larson, S. M., England, J. L., Desjarlais, J. R. & Pande, V. S. (2002). Thoroughly sampling sequence space: large-scale protein design of structural ensembles. *Protein Sci.* **11**, 2804–2813.
33. Henikoff, S. & Henikoff, J. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
34. Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.

35. Humphrey, W., Dalke, A. & Schulten, K. (1996). VMD—visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38.
36. Delano, W. L. (2002). *The Pymol User's Manual*, DeLano Scientific, San Carlos, CA, USA.
37. Shindyalov, I. N. & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11**, 739–747.

Edited by M. Levitt

(Received 17 June 2004; received in revised form 18 November 2004; accepted 22 November 2004)