PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Computational protein design as a tool for fold recognition

Marcel Schmidt am Busch, David Mignon, and Thomas Simonson*

Laboratoire de Biochimie (CNRS UMR7654), Department of Biology, Ecole Polytechnique, 91128 Palaiseau, France

ABSTRACT

Computationally designed protein sequences have been proposed as a basis to perform fold recognition and homology searching. To investigate this possibility, an automated procedure is used to completely redesign 24 SH3 proteins and 22 SH2 proteins. We use the experimental backbone coordinates as fixed templates in the folded state and a molecular mechanics model to compute the pairwise interaction energies between all sidechain types and conformations. Energy calculations are done with the Proteins@Home volunteer computing platform. A heuristic algorithm is then used to scan the sequence and conformational space for optimal solutions. We produced 200,000—450,000 sequences for each backbone template. The designed sequences ressemble moderately-distant, natural homologues of the initial templates, according to their identity scores and their similarity with respect to the Pfam sets of SH2 and SH3 domains. Standard homology detection tools document their native-like character: the Conserved Domain Database recognizes 61% (52%) of our low-energy sequences as SH3 (SH2) domains; the SUPERFAMILY, Hidden-Markov Model library recognizes 81% (84%). Conversely, position specific scoring matrices (PSSMs) derived from our designed sequences can be used to detect natural homologues in sequence databases. Within SwissProt, a set of natural SH3 PSSMs detects 772 SH3 domains, for example; our designed PSSMs detect 67% of these, plus one additional sequence and two false positives. If six amino acids involved in substrate binding (a selective pressure not accounted for in our design) are reset to their experimental types, then 77% of the experimental SH3 domains are detected. Results for the SH2 domains are similar. Several directions to improve the method further are discussed.

Proteins 2009; 77:139–158.

Key words: directed evolution; molecular mechanics; homology searching; protein folding; distributed computing.

INTRODUCTION

Uncharacterized proteins from complete genome sequencing projects now dominate the relevant public databases and their number continues to grow rapidly.[1–7] At the end of 2004, public databases contained about one million unique protein sequences from ~200 completely sequenced genomes. At the end of 2007, already about five million proteins from ~600 completely sequenced genomes were available.[4–7] The experimental, three-dimensional structure is known for fewer than 1% of the sequences.[5,8] Thus, a major goal of *in silico* approaches in the post-genomic era is the structural and functional characterization of new protein sequences.

To characterize a protein structure, the essential first step is to identify its "fold"; that is, the overall, spatial arrangement of its polypeptide backbone. The fold can be viewed as a medium resolution description of the 3D structure.[9–11] Most protein structures can be subdivided into one or more compact modules called domains, which have their own independent fold. Multidomain proteins can thus be viewed as assemblies of a few (1–5) smaller, well-defined structures.

Current structural data banks contain about 100,000 known structural domains.[8] These can be grouped into a few thousand families, with members of a single family sharing the same fold; that is, having a rather close structural similarity and an evolutionary relationship. The SCOP database, or "Structural Classification of Protein," for example, currently identifies 3464 protein families and 1777 "superfamilies."[9,10] Thanks partly to the progress of structural genomics, the number of known folds appears to be converging towards a limited number of total existing folds, probably on the order of 10,000. Hence, structural biology may become a two-step process, with fold recognition as the first step. Fold recognition assigns an uncharacterized protein to one of the known families, or folds.[12,13] In a second step, the resulting model can be refined using established comparative modeling techniques.[13–15]

Fold recognition is often done using profile-based methods, such as PSI-Blast,[16,17] RPS-Blast,[18] or profile Hidden Markov Models (HMMs).[19–25] These methods rely on virtual consensus sequences,

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Thomas Simonson, Laboratoire de Biochimie (CNRS UMR7654), Department of Biology, Ecole Polytechnique, 91128 Palaiseau, France.
E-mail: thomas.simonson@polytechnique.fr.

Received 28 October 2008; Revised 5 February 2009; Accepted 9 March 2009
Published online 24 March 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22426

which are statistical descriptions of families of homologous protein domains (usually organized into a multiple sequence alignment). To allow fold recognition, one statistical model should be constructed for each fold family. Thus, the SUPERFAMILY resource, used below, provides one multiple sequence alignment (MSA) and one profile-HMM for each structural SCOP domain (92,927 domains and 3464 families).[22,23,26] Similarly, the Protein Family databank (Pfam) groups the known protein domains into 9318 families.[21,27,28] Pfam has two MSAs for each family and as many profile-HMMs.[21,28] The conserved domain database (CDD),[18] also used below, is the protein classification component of NCBI. CDD profiles are constructed using MSAs from a number of resources, including Pfam,[27] SMART,[29] and COG.[30] CDD is applied automatically to many genome annotation projects.

Overall, about two thirds of all protein sequences can be characterized with existing HMM models.[18] Thus, SUPERFAMILY matches 60% of all entries of the Non-Redundant sequence database,[26,31] whereas Pfam matches 75% of all entries when their HMM models are applied to the SwissProt and TrEMBL databases.[27] However, one often needs more than one model for a single homologous family; the constant development in recent years of new HMM libraries for fold recognition and homology searching documents the difficulty of the problem; none of the fold recognition tools has so far converged to completeness.[18,27]

In this work, we examine the usefulness of computationally designed protein sequences as a basis for fold recognition and homology searching, complementing existing tools. This question was examined by Koehl and Levitt[32] and by Larson and Pande.[33,34] We focus on the SH3 and SH2 families of protein domains. SH3 proteins have been extensively studied by computational protein design (CPD), whereas SH2 proteins are larger (∼100 residues) and have rarely been studied. We test the ability of position specific scoring matrices (PSSMs) derived from designed sequences to detect members of each family, and compare their performance to experimental PSSMs, obtained from natural sequences. CPD has been extensively used for protein engineering; it represents a rigorous test of our understanding of the biophysical mechanisms that shape protein sequences and structures.[33,35–59] However, only a few large-scale applications to fold recognition or homologue searching have been reported.[32–34,47,57,58,60] Larson and Pande did the largest study, using designed sequences to do homology searching.[33,34] They selected 253 small proteins from the Protein Data Bank. For each protein, they generated 700–800 low-energy sequences. PSSMs were built and used to find homologues in the PDB and within several fully-sequenced microbial genomes. In general, the designed profiles performed better than a single pairwise BLAST search using a natural query.

Here, CPD is accomplished for 24 SH3 domains and 22 SH2 domains of known 3D structures, using an implementation described recently.[61,62] Each protein is described in atomic detail; the backbone is held fixed, whereas sidechains are allowed to mutate and explore favorable rotameric conformations. The fixed backbone assumption is shown to be an acceptable approximation for this application (see Supporting Information). Protein sequences and structures are selected according to their estimated folding free energy, which relies on a simple unfolded state model. The protein backbone coordinates of all the proteins were determined by X-ray crystallography. In addition, we used one NMR model: the N-terminal SH3 domain of Grb2 (PDB code 1aze). The designed sequences were analyzed and their suitability for homologue detection studied in detail.

The quality of the designed sequences is good (and comparable to other recent work) but not perfect. Similarity to natural sequences is comparable to the similarity between moderately-distant, natural homologues. The SUPERFAMILY fold recognition tool identifies about 80% of the low-energy designed sequences as SH3 or SH2 sequences. Conversely, with PSSMs constructed from the designed sequences, we can retrieve 67% of the SH3 sequences and 56% of the SH2 sequences in SwissProt. If a few functional positions are manually reset to their experimental amino acid types, we retrieve over 70% of the SwissProt sequences. Following the broader but less detailed exploration of Larson and Pande,[33] the analysis reported here for two, well-characterized protein families is a further step towards the large-scale application of CPD in fold recognition.

METHODS

Folded and unfolded states

In our CPD procedure, sequences and structures are selected based on their folding free energies, ΔG_{fold}. In the folded state, the coordinates of the protein backbone are kept fixed, while sidechains occupy rotamers from the backbone-independent Tuffery library.[63] The backbone conformation was obtained by subjecting the crystal structure to 500 steps of conjugate gradient energy minimization. During the minimization, the effect of solvent was represented by a uniform dielectric constant of 20, applied to the Coulomb electrostatic energy term. The minimization typically led to an rms deviation (including backbone and C_β atoms) of 0.7 Å from the experimental structure, and a radius of gyration about 0.1 Å smaller than in the crystal structure.

In the unfolded state, the amino acid sidechains do not interact with each other, but only with nearby backbone and with solvent (through the CASA implicit solvent model; see below). Specifically, for each amino acid type X, we considered a large number of possible tripeptide structures with the sequence Ala-X-Ala, with backbone geometries taken from five proteins. The lowest-energy combination of backbone structure and sidechain rotamer was taken to represent the preferred structure of X in the unfolded state. The corresponding energy, E_X, represents the contribution of X to the unfolded state free energy. An additional (and smaller) contribution, e_X, was determined empirically, so as to obtain reasonable overall amino acid compositions in the final computed sequences; more details are given elsewhere.[61,62] For a given amino acid sequence, the unfolded state free energy is obtained by summing the contributions $E_X + e_X$ of the individual amino acids.

Effective energy function

The effective energy function was described in detail elsewhere.[64] Briefly, we use the Charmm19 molecular mechanics energy function[65] along with the CASA implicit solvent model. With CASA, the solvent contribution is the sum of a screened Coulomb term and a solvent accessible surface term:

$$E_{\mathrm{solv}} = \left(\frac{1}{\varepsilon} - 1\right) E_{\mathrm{coul}} + \alpha \sum_i \sigma_i A_i. \quad (1)$$

Here, E_{Coul} is the usual Coulomb energy, ε is a dielectric constant, equal to 10; the righthand sum is over the protein atoms i, A_i is the solvent accessible surface area of atom i, σ_i is an atomic solvation coefficient that depends on the atom type, and α is an overall scaling factor for the surface term.

The interaction energy between each pair of sidechains, or between a sidechain and the backbone, involved a short energy minimization stage.[42] Each sidechain was first subjected to 15 steps of Powell minimization, with the backbone fixed and inter-sidechain interactions excluded. Then, interactions between the sidechain pair were included and a further 15 steps of minimization performed. The sidechain interaction energy was taken from this last, minimized structure. Interactions between distant groups were omitted through the following cutoff scheme. If the inter-C_β distance was above 15 Å (respectively, below 10 Å), a residue pair was omitted (included). Otherwise (inter-C_β distance between 10 and 15 Å), if the minimum inter-sidechain distance was 9 Å or less, the pair was included.

Surface areas were computed using the Lee and Richards algorithm,[66] using a 1.4 Å probe radius. The atomic solvation coefficients σ_i are the ones used in our previous work: 0.012 kcal/mol/Å^2 for carbons and sulfur; −0.06 kcal/mol/Å^2 for oxygen and nitrogen; zero for hydrogens, and −0.15 kcal/mol/Å^2 for ionized groups.[64] For reasons of efficiency, following Street and Mayo,[67] we assume that A_i can be obtained by summing the contact areas A_{ij} between atom i and its neighbors j, and subtracting the contact, or solvent-inaccessible area $C_i = \Sigma_j A_{ij}$ from the total area of atom i. This approximation has the enormous advantage that the surface energy takes the form of a sum over pairs of amino acids. However, it leads to a systematic error, because the contact areas can overlap: a portion of atom i can be in contact with two atoms j and j' at a time. The systematic error can be largely corrected by applying a scaling factor of 0.5 to contact areas A_{ij} that involve at least one buried atom (i or j).[64,67]

Sequence optimization

We used a heuristic optimization procedure developed by Wernisch *et al.*[42,61] A "heuristic cycle" proceeds as follows: an initial amino acid sequence and set of sidechain rotamers are chosen randomly. These are improved in a stepwise way. At a given amino acid position i, the best amino acid type and rotamer are selected, with the rest of the sequence held fixed. The same is done for the following position $i + 1$, and so on, performing multiple passes over the amino acid sequence until the energy no longer improves (or a set, large number of passes is reached). The final sequence, rotamer set, and energy are output, ending the cycle. The method can be viewed as a steepest descent minimization, starting from a random sequence, and leading to a nearby, local, (folding) energy minimum. For the design calculations below, we perform ∼450.000 (200,000) heuristic cycles for each SH3 (SH2) protein, thus sampling a large number of local minima on the energy surface. Cysteines, glycines, and prolines are expected to have a special effect on the protein's folded and unfolded state structures, which may not be accurately captured by our method. Therefore, if these amino acids are present in the native sequence, they are not mutated; all other amino acids are allowed to mutate freely (but not into Cys, Gly, or Pro).

Software implementation

As pointed out by Mayo *et al.*[68] the pairwise energy function and discrete conformational space imply that all the relevant energy data can be precomputed and stored in an energy matrix.[42] In effect, we must compute the interactions between all pairs of amino acids in the structure, allowing for all possible pairwise combinations of amino acid types and rotamer values. This calculation is done with the XPLOR program,[69] using a single command script and standard features of the program. Because of its pairwise nature and low communication requirements, this calculation can be done in parallel. We used our Proteins@Home distributed computing platform, which allows us to use the computers of several thousand volunteers in over 100 countries (see the list of participants at biology.polytechnique.fr/proteinsathome). Proteins@Home is based on the Berkeley Open Infrastructure for Network Computing, BOINC.[70] The Proteins@Home platform and project will be described in detail elsewhere.[71]

In a second stage, sequence optimization is done with the heuristic algorithm described above. A C++ program was developed, called Proteus, with core routines taken from the Optimizer program of L. Wernisch.[42] Postprocessing of the computed sequences is done with Proteus and a set of Perl scripts. The XPLOR scripts and Proteus program are available on request (thomas.simonson@polytechnique.fr).

Similarity scores

To measure the quality of the designed sequences, we computed similarity scores between each designed sequence and a multiple sequence alignment (MSA) of experimental sequences. For the SH3 case, the MSA includes the Pfam alignment of 62 SH3 domains, along with the 24 proteins studied here, 13 of which are already part of the Pfam set. This gives an overall, reference MSA of 73 SH3 domains. Below, we refer to this as the Pfam alignment, even though it includes 11 additional proteins. For the SH2 case, the MSA included 133 SH2 sequences. The sequences were collected by doing BLAST searches among the Pfam SH2 sequences, using each of our 22 proteins successively as a query; the homologues obtained this way were pooled and aligned according to the Pfam SH2 alignment, leading to the final MSA. Each of our proteins being aligned with the appropriate MSA, there is a unique correspondence between positions in our designed sequences and the MSA. We then computed the following similarity score:

$$s = \sum_i \sum_a f_{ia} S(x_i, a). \tag{2}$$

Here, i is a position in the designed sequence; x_i is the amino acid type in the designed sequence at that position; a is either one of the 20 amino acid types or a gap symbol; f_{ia} is the frequency (between 0 and 1) of a at the corresponding position in the Pfam MSA; $S(x_i,a)$ is the BLOSUM62 scoring matrix. If a is a gap symbol, $S(x_i,a)$ is set to -5. The first sum is over the designed sequence; the second sum is over the amino acid types (including the gap symbol).

SUPERFAMILY

The SUPERFAMILY database is a library of profile Hidden Markov Models,[72] designed to associate a protein sequence with the most probable 3D structural model. The library is based on the SCOP classification of proteins, with one model for each protein domain in SCOP. We downloaded the set of models (version 1.69) and used them in connection with the Sequence Alignment and Modeling system (SAM, version 3.5), recommended by the creators of the SUPERFAMILY database. For each of our SH3 (SH2) proteins, we used the first 10,000 (8000) designed sequences as queries against the SUPERFAMILY library, and also the 10,000 (8000) lowest-energy designed sequences (out of the full set of 450,000 (200,000) designed sequences). Significant hits were returned with the corresponding E-value and the SH3 (SH2) domain assignment.

CDD: a Conserved domain database for protein classification

The Conserved Domain Database (CDD) is the protein classification component of NCBI's Entrez query and retrieval system. CDD contains protein domain models imported from outside sources, such as Pfam and SMART, and protein domain models curated at NCBI. In total, CDD contains over 12,000 models. Our designed sequences were queried against the CDD database, run locally. For each SH3 (SH2) protein, we analyzed the complete set of designed sequences and the set of 10,000 (8000) lowest-energy sequences.

PSI-BLAST analysis of the designed sequences

To evaluate the native-like character of the designed sequences, for each SH3 domain, we used one of two experimentally-based PSSMs, and a database containing both designed and natural sequences. Specifically, for each of our 24 SH3 domains (respectively, 22 SH2 proteins), we constructed a database containing the 450,000 (200,000) designed sequences along with half of the "Non-Redundant" database, NR01, which is a compilation of coding sequences from GenBank, RefSeq, the PDB, SwissProt, PIR and PRF; redundant sequences are removed, using a threshold of 60% sequence identity to define redundancy. We searched this database using the program BLASTPGP (running locally), with one of two PSSMs. The first PSSM is a "general" PSSM, constructed as follows. We started from an arbitrary SH3 domain and used it to query the NR01 database. Four PSI-BLAST iterations were performed, using a 10^{-3} E-value cutoff to define hits. After the four iterations, we were left with about 1000 homologous sequences and a PSSM. The second PSSM is "backbone-specific". We started from one of our SH3 (SH2) domains and searched SwissProt with a single PSI-BLAST iteration, collecting about 50 sequences that have at least a 45% identity with it, and which define the PSSM. With the PSSM in hand (general or backbone-specific), we searched our database (designed, plus NR01 sequences) using BLASTPGP. Designed sequences returned as hits are deemed to be SH3-like (SH2-like).

Table I
The Test Set of SH3 and SH2 Domains: Identity Scores of the Designed Sequences

SH3 domains				SH2 domains			
PDB code	Chain length	Overall mean[a]	Low energy[b]	PDB code	Chain length	Overall mean[a]	Low energy[b]
1cka	56	34.5	36.7	1cwe	97	27.07	28.18
1fyn	59	25.3	26.6	1k9a	101	27.16	26.87
1abo	58	35.9	36.7	1shd	101	25.81	25.55
1pht	81	36.5	39.4	1g83	104	28.66	28.72
1shg	57	23.6	25.8	1ayd	98	27.87	28.38
1ad5	58	23.5	23.8	1cj1	96	24.34	24.63
1csk	56	37.1	36.4	1ad5	103	28.73	29.86
1fmk	60	26.3	30.1	1ju5	109	30.80	31.47
1gcq(B)	57	37.3	37.9	1mil	101	24.62	22.92
1sem	58	35.1	37.9	2pnb	103	27.11	27.35
1uti	57	28.5	31.6	1nrv	100	26.03	26.09
1lck	59	28.0	29.2	1a81	108	34.65	35.70
1ycs(A)	62	28.9	27.8	1m61	112	40.79	41.26
1bb9	72	22.2	25.2	1lun	107	29.25	28.92
1aoj	114	26.4	26.5	1opk	98	26.65	26.21
1gcq(C)	69	42.5	47.1	1jwo	97	29.54	29.24
1i1i	72	33.2	35.9	1bf5	130	28.67	30.52
1kjw	74	24.9	25.8	1bg1	126	27.60	28.60
1jo8	58	30.0	32.2	1uur	116	25.17	25.11
1ng2	56	31.2	32.4	2cbl	88	31.74	32.65
1jqq	64	28.5	27.9	1m27	104	34.13	34.08
1uj0	55	29.2	34.9	1r1q	93	23.62	23.00
1oot	58	33.7	37.1				
1vyv	68	34.9	35.7				
1aze[d]	56	39.8	42.4				
Average	59	30.7	32.9 (39.2[c])		104	28.6	28.9 (35.0[d])

Mean identities between the designed sequences and the corresponding, native template.
[a] Averages are computed for all designed sequences.
[b] Averages are computed for the 10,000 (8000) lowest-energy SH3 (SH2) sequences.
[c] Mean identity after the peptide binding positions (PBPs) are reset to their experimental types.
[d] NMR structure, with the same sequence as 1sem.

Residual entropy of the natural and designed sequences

To compare the sequence diversity in the designed sequences with the diversity in natural sequences, we used a standard, position-dependent entropy,[72] computed as follows:

$$S_i = -\sum_{i=1}^{6} f_j(i) lnf_j(i) \qquad (3)$$

where $f_j(i)$ is the frequency of residue type j at position i, either in the designed sequences or in the natural sequences (organized into an MSA). Instead of the usual, 20 amino acid types, we use six residue types, corresponding to the following groups: {LVIMC}, {FYW}, {G}, {ASTP}, {EDNQ}, and {KRH}. This classification is obtained by a cluster analysis of the BLOSUM62 matrix,[73] and also by analyzing residue-residue contact energies in proteins.[74] To get a sense of how many amino acid types appear at a specific position i, we report the residue entropy in its exponentiated form, exp(S_i), which ranges from 1 to 6.

Threading analysis of the designed sequences

As a further quality measure, for selected backbone templates (four SH3 proteins and four SH2 proteins), we submitted our low-energy designed sequences to the FROST threading tool,[75] running remotely. Each sequence was threaded onto a large library of backbone models (over 3000 models), including both SH3 and SH2 structures, and chosen by the FROST developers to be representative of a large fraction of all known, non-redundant structures in the PDB (with redundancy defined by a 40% sequence identity). The highest-scoring models and scores were retrieved and analyzed.

RESULTS

Similarity scores of designed sequences

To test the potential of computational protein design (CPD) as a tool for fold recognition, we considered the SH3 and SH2 protein superfamilies in SCOP (Table I). We redesigned 22 SH2 and 24 SH3 proteins for which Xray structures are available; for one, we also did calculations using an NMR structure (the N-terminal domain of human Grb2; PDB code 1aze). For each SH3 (SH2) protein, we performed 450,000 (200,000) heuristic cycles of sequence generation, leading to as many redesigned sequences. All the sidechains were mutated except prolines, cysteines and glycines. Below, we present the SH3 results in detail, and the SH2 results, which are similar, more briefly, with details in Supporting Information.

Identity rates between the designed sequences and the initial, native sequence are commonly used as a first quality check for CPD. Averaging over each set of designed sequences and all SH3 (SH2) proteins, we obtain a mean identity score of 30.7% (28.9%). Taking the lowest-energy sequences for each protein (10,000 for the SH3 and 8,000 for the SH2 proteins), we obtain 32.9% (28.3%) (Table I). For the SH3 domains, these identity rates are comparable to those obtained by other workers, as discussed previously.[61] They are also close to the identity rates between the natural sequences in our test set. Indeed, we compared our 24 SH3 domains to two of the 24: 1cka and 1fyn. The corresponding identity rates range from 18 to 75%, with an average of about 32% (see Supporting Information). We also measured the structural similarity between the native, backbone, SH3 structures, using the structural alignment program MATRAS.[76] The rms deviations, averaged over the aligned C_α atoms (typically, 95% of each sequence), range from 0.8 to 3.2 Å, with an average of 1.6 Å.

Similarity scores are a more reliable measure of the native-like character of designed sequences, because they take into account the diversity of the natural sequences and better reflect the experimental, physico–chemical, and geometrical characteristics of each protein. For 14 of the 24 SH3 proteins, we computed similarity scores between each designed sequence and the Pfam SH3 MSA (correctly aligned to the corresponding native sequence; see Methods). The Pfam MSA is shown in Figure 1. Figure 2 shows histograms of the similarity scores, both for the full sets of 450,000 sequences and for low energy sets (the 10,000 lowest-energy sequences). Similarity scores are also shown for the 62 natural sequences in the Pfam alignment and for a larger set of SH3 domains found in Pfam.[77] Notice that only complete domains in the large Pfam SH3 set were used; several incomplete domains were excluded. The large SH3 set, thus curated, contains about 2000 domains.

The histograms for the designed sequences overlap extensively the range of scores obtained with the large Pfam set of SH3 domains (76% overlap). Overlap with the scores in the small Pfam set is much lower (27% overlap on average). For the SH2 case, we considered 20 of the 22 proteins in SCOP. We excluded 1bf5, which has a large insertion and is difficult to align with the Pfam sequences, and 2cbl, which is not classified as an SH2 domain by Pfam. These two SH2 domains are considered "non-classical." Results for the SH2 proteins are somewhat poorer: 40% overlap with the large set of Pfam SH2 domains (Supporting Information). Overall, for both the SH3 and SH2 domains, the designed scores are mostly comparable to moderately-distant natural homologues. Notice that similarity scores were mostly computed using the Blosum62 matrix, for comparison to earlier work.[61,78] If the Blosum45 matrix is used, results are very similar (slightly better, in fact): for 14 SH3 domains, for example, the overlap with the scores in the small Pfam set is 29.1% (27% with Blosum62); the overlap with the scores in the large Pfam set is 78.4% (75.8% with Blosum62).

The influence of core and functional SH3 positions

For six out of the 24 SH3 proteins (1cka, 1fyn 1abo, 1shg, 1ad5, and 1gcq), we considered the effect of two specific groups of positions on the similarity scores. Ten hydrophobic core positions are thought to be required to maintain the fold and stability of SH3 domains (positions 4, 6, 18, 20, 26, 28, 37, 39, 50, and 55, following the numbering of Larson et al.[33]); we refer to these as the "core" positions. Six other positions are closely involved in the function of SH3 domains, because they participate in the binding of a peptide ligand; we refer to these as the "peptide-binding" positions, or PBPs. Figure 2 shows that the designed sequences achieve high similarity scores for the core positions, in the same range as the natural sequences in the small Pfam set. The most common amino acid types found at the core and peptide-binding positions are shown in Supporting Information Material, documenting further the quality of the designed, core positions. Mutations of the original residue to a charged amino acid are never observed, and mutations of a hydrophobic residue to a polar one are very rare.

Similarity scores for the peptide binding positions are considerably worse (Fig. 2). In particular, the similarity scores of the three worst-case proteins (1fyn, 1shg, and 1ad5) show a very weak overlap with the corresponding spectrum of natural similarity scores. This behavior is not surprising, because the peptide binding positions are determined by the proteins' function, whereas our designed sequences are selected for stability. If the PBPs are reset to their experimental types (i.e., taken from the native templates), the overlap with the similarity scores in the large Pfam set is 100%; overlap with the small set is 85% (Table III). For the SH2 case, the PBPs correspond to residues 7, 14, 34, 46, 61, and 88 in c-Src (PDB code 104R).[79] With reset PBPs, overlap with the small (large) Pfam set is 76% (84%) (Supporting Information).

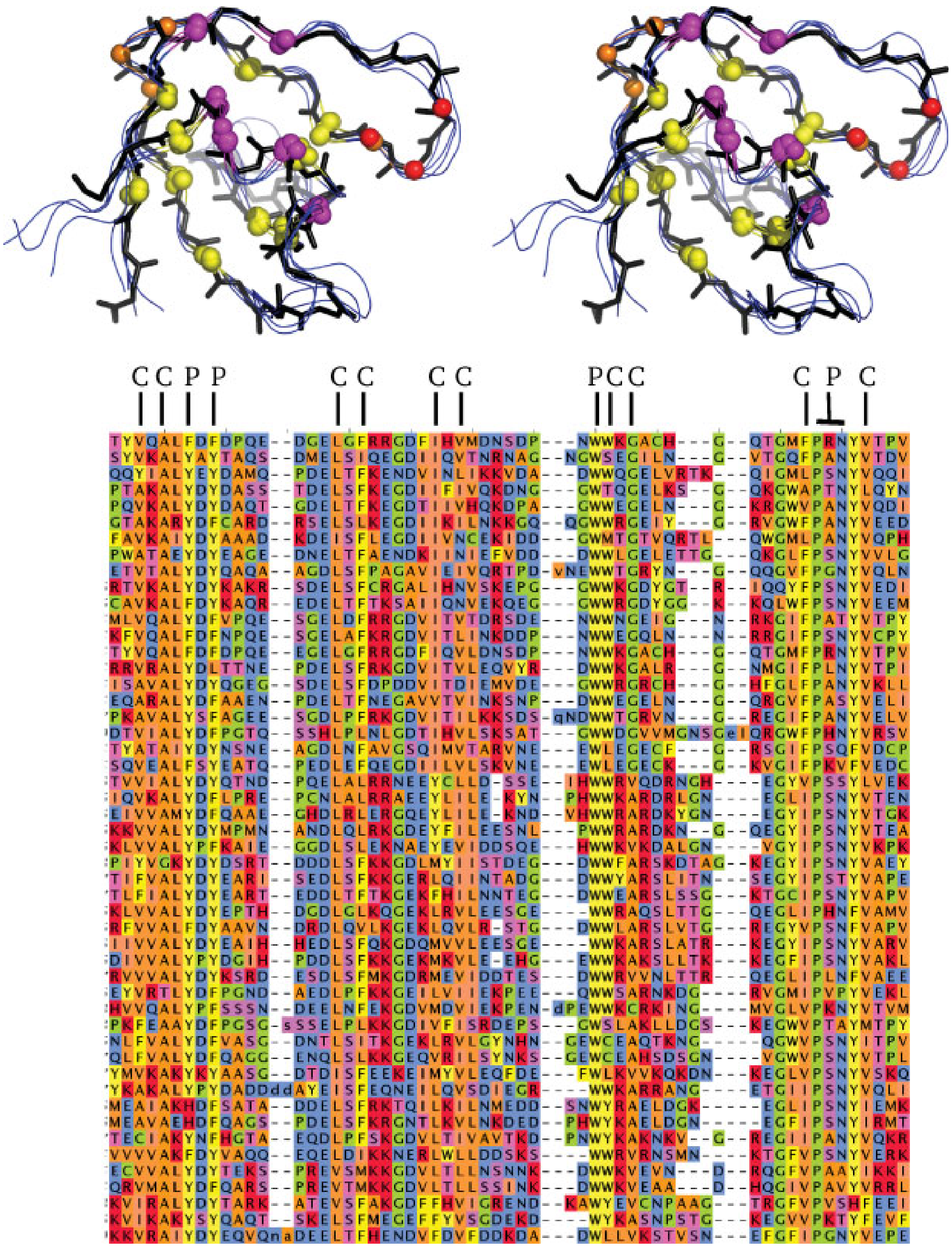

Figure 1

Pfam alignment of SH3 domains (first 50 sequences only, out of 62 in the alignment). Amino acids are colored by physical–chemical groups. Core positions and "peptide-binding positions" (PBPs) are labeled (C and P labels, above). A stereo view of five SH3 domains is shown, above. For one (1cka), the backbone is detailed as black sticks; the others are drawn as blue tubes (1abo, 1shg, 1csk, and 1gcq). Yellow spheres represent core positions; purple spheres represent PBPs. Red spheres are three additional PBPs present in 1cka; orange spheres are three additional PBPs present in 1shg.

Residual entropy

Having characterized the mean similarity between the designed and experimental sequences, we consider now the diversity of the two sets, using a standard entropy measure.[72] In Figure 3, for the SH3 case, we show the (exponentiated) entropy, e^{S_i} as a function of the residue number *i*. The experimental curve corresponds to the Pfam MSA of SH3 domains, supplemented by our own 24 SH3 domains, for a total of 73 unique, experimental sequences. Results for the SH2 case are similar (Supporting Information). The calculation uses a classification of the amino acid types into six groups (see Methods);

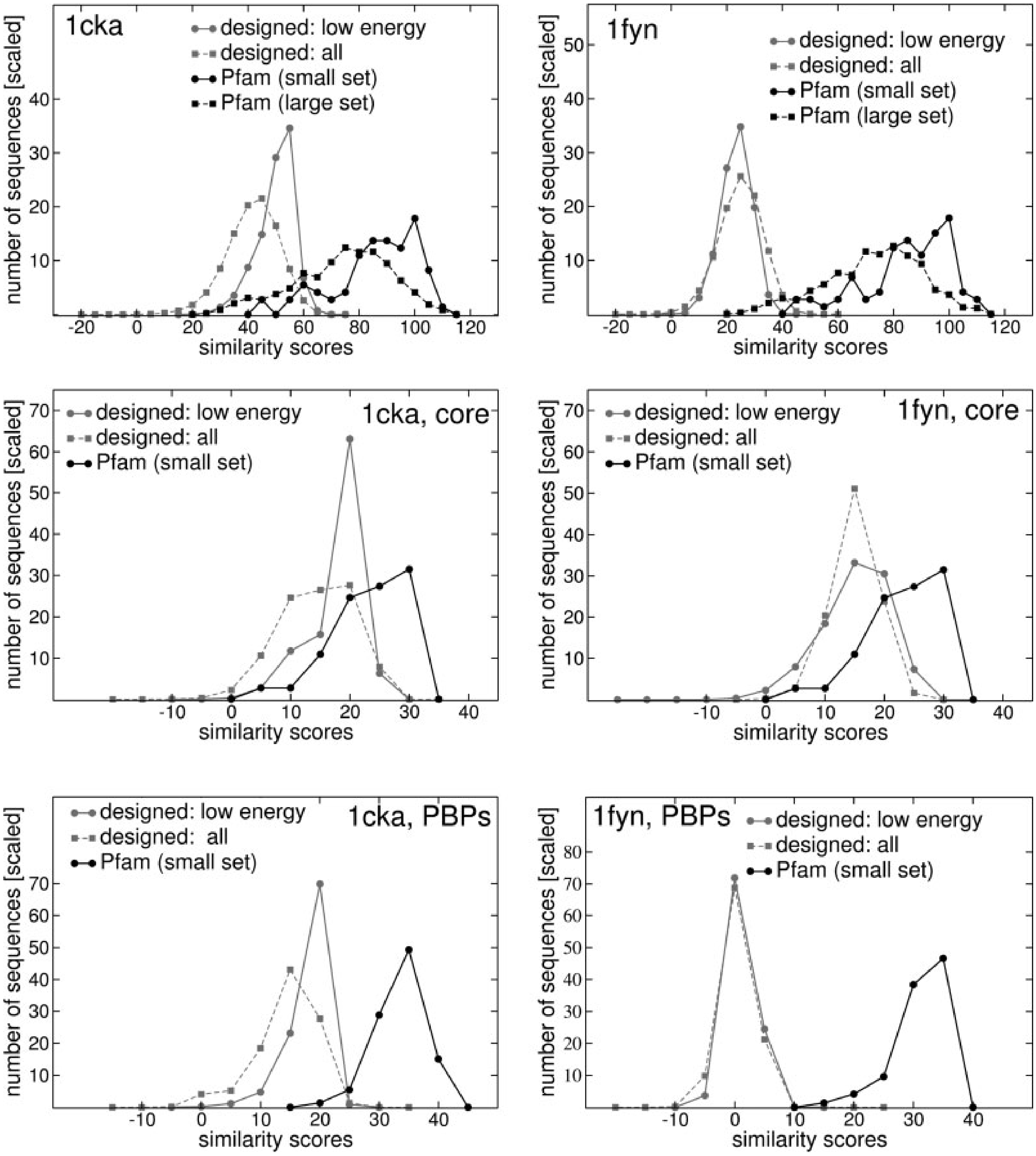

Figure 2

Similarity scores (with respect to the Pfam MSA) for natural and designed SH3 sequences obtained for two representative backbone templates, 1cka and 1fyn. For the natural sequences (either large or small Pfam sets), only positions aligned with the template (1cla or 1fyn) are considered (this explains the slight differences between the natural scores in the left and right upper panels). For the designed sequences, results are shown for all 450,000 sequences (dashed gray) or the 10,000 lowest energy sequences (solid gray). Upper panels: scores for entire proteins; middle panels: 10 core positions only; lower panels: six peptide-binding positions (PBPs). The small “Pfam set” refers in fact to the 62 SH3 sequences in the Pfam MSA, plus our 24 sequences, for a total of 73 unique sequences.

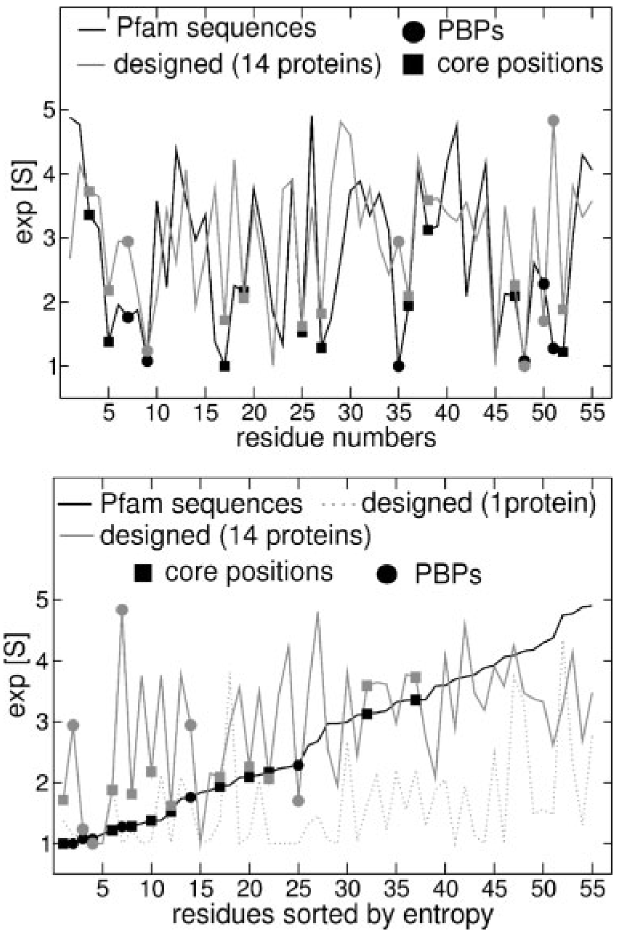

Figure 3

Sequence entropy S of natural and designed SH3 sequences, exponentiated, for clarity, and computed with a six-class amino acid alphabet; see text. Residues are numbered according to the Pfam MSA (upper panel) or by increasing entropy (in the natural sequences; lower panel). Core positions and peptide-binding positions (PBPs) are highlighted. For the designed sequences, we considered the 10,000 lowest-energy sequences from either 14 backbone templates (the "classic" SH3 domains; see text), or from a single backbone template (1cka; dotted line, lower panel).

e^{S_i} can be interpreted as the number of groups that appear at position i in the MSA. The theoretical entropy curve corresponds to the low energy, designed sequences from the 14 classical SH3 domains, aligned according to the Pfam MSA (see Methods). The mean experimental and theoretical entropies are very similar: 2.8 and 3.0, respectively (Table II). Thus, in both sets of sequences, there are about three different amino acid groups sampled at each position. The slightly greater diversity in the designed sequences arises mainly from the six peptide binding positions (PBPs), whose mean entropy is 2.4, compared with 1.4 in the Pfam sequences. This is not surprising, because the PBPs are constrained by their functional role in the experimental sequences, but not in the designed sequences, where they are selected for protein stability only. For the SH2 sequences, the mean entropy of the designed sequences is 2.8, in agreement with the experimental value, 2.8. Notice that the designed sequences for a single protein (Grb2, in Fig 3) underestimate notably the experimental diversity. Using an ensemble of protein backbones, obtained either from a normal mode calculation or a set of different Xray structures of the same protein, yields results that are better,[47] but still inferior to the ensemble of 14 SH3 domains or 20 SH2 domains (see Supporting Information).

The variation of the entropy along the polypeptide chain (excluding the PBPs) is qualitatively similar in Pfam and the design (considering again the 14 low energy SH3 sets), though the detailed behavior is different. Similar agreement was found in other recent studies.[60] The level of agreement is seen most clearly in Figure 3(B), where the amino acids are sorted by increasing (experimental) entropy. The agreement for the core positions is very good. In contrast, three of the six PBPs have very large computed entropies; experimentally, five of the twelve lowest entropies are at PBPs [Fig. 3(B)]. The most experimentally-diverse positions (right of Fig. 3B) are also diverse in the design, but the computed diversity is somewhat underestimated. Overall, the agreement with experiment, PBPs excluded, is quite good.

Fold recognition tools confirm the natural character of designed sequences

CDD and SUPERFAMILY detection

Before using the designed sequences for homology detection, we tested their quality further, subjecting them to four standard fold recognition and homology detection tools: PSI-BLAST, SUPERFAMILY, CDD, and FROST.[75] With the CDD library, 49% of the designed SH3 sequences and 61% of the low energy sequences are

Table II

Entropy of Natural and Designed SH3 and SH2 Sequences

	Amino acids	Pfam sequences[a]	Designed[b]	Designed[b] (1 protein)
SH3 domains	Core	1.9	2.4	1.4
	PBPs	1.4	2.4	1.2
	Remaining	3.2	3.2	1.8
	All	2.8	3.0	1.7
SH2 domains	Core	2.0	2.6	2.1
	PBPs	1.7	2.4	1.5
	Remaining	3.0	2.9	1.9
	All	2.8	2.8	1.9

We report exponentiated entropies, exp(S), computed using a simplified amino acid alphabet with six amino acid classes: {LVIMC}, {FYW}, {G}, {ASTP}, {EDNQ}, and {KRH},[74] instead of a detailed, 20-class alphabet.

[a] Entropy calculated from 73 natural SH3 sequences or 133 natural SH2 sequences.

[b] Entropy of the designed sequences, using the 10,000 lowest-energy sequences from either the 14 classic SH3 domains, a single SH3 domain (1cka), the 22 SH2 domains, or a single SH2 domain (1cwe).

Table III
Similarity Scores and Detection Rates for the Designed SH3 Sequences

					PSI-BLAST[c]		
PDB code	Identity score[a]	Similarity overlap[b]	SUPERFAMILY[c]	CDD[c]	Close homologues	General	[d]Reset PBPs
1cka	36.7	98.8	100.0	100	99.2	87.2	95.3 (93.2)
1fyn	26.6	96.5	25.5	44.2	12.6	0.0	76.1 (92.4)
1abo	36.7	96.9	99.4	72.0	94.6	12.8	86.6 (96.6)
1pht	39.4		97.0	8.1		90.3	
1shg	25.8	31.0	75.9	50.2	0.0	0.1	0.2 (99.9)
1ad5	23.8	82.6	39.4	41.1	1.3	0.7	79.3 (94.9)
1csk	36.4	88.8	97.5	82.0	59.2	8.4	82.7 (100)
1fmk	30.1	99.5	98.4	77.8	36.4	36.4	95.5
1gcq(B)	37.9	100.0	100.0	100.0	99.9	96.3	100.0
1sem	37.9	97.4	99.7	98.9	92.9	48.8	98.2
1uti	31.6	99.9	99.6	99.6	22.0	47.6	96.3
1lck	29.2	35.1	78.4	43.5	11.1	3.0	53.9
1ycs(A)	27.8		44.8	9.3		0.4	
1bb9	25.2		43.5	7.0		0.5	
1aoj	26.5		52.3	60.2		6.2	
1gcq(C)	47.1		98.5	72.0		83.9	
1i1i	35.9		52.3	33.5		43.1	
1kjw	25.8		75.0	12.8		1.6	
1jo8	32.2		99.9	99.7		73.2	
1ng2	32.4	68.6	94.0	64.3	70.5	38.1	91.0
1jqq	27.9		67.4	56.9		36	
1uj0	34.9	99.9	99.8	100.0	62.8	69.8	95.8
1oot	37.1		98.4	79.8		47.5	
1vyv	35.7		93.1	1.3		48.6	
1aze	42.4	99.9	100.0	99.5	96.1	84.1	99.2
[e]mean (14)	33.5	85.4 (99.6[f])	86.3	76.6	54.2	38.1	82.3
mean (25)	32.9		81.2	61.0		39.5	

[a] Identity scores of the low energy designed sequences.

[b] Overlap (%) between the similarity scores and the scores of natural sequences in the small Pfam set of SH3 domains. In the designed sequences, the six conserved PBPs have been reset to their experimental values. A 100% overlap means that the designed sequences all have scores within the range spanned by the natural sequences' scores.

[c] Detection rates (%) for the 10,000 lowest-energy designed sequences, using SUPERFAMILY, CDD, or PSI-BLAST. With PSI-BLAST, two PSSMs are compared for each backbone template: a PSSM constructed from close homologues, and a more general PSSM for SH3 domains (see text).

[d] The more general PSSM is also applied to designed sequences with the six conserved PBPs reset to their experimental values (in parentheses, results with another four, less-conserved PBPs also reset).

[e] Mean values are computed over all 25 templates or over the 14 "classic" SH3 templates (see text).

[f] The mean overlap with the large Pfam set.

identified as SH3 sequences; 52% of the low energy SH2 sequences are correctly identified. The SUPERFAMILY library of Hidden Markov Models (HMMs) correctly identified 81% of the low energy SH3 sequences and 83% of the low energy SH2 sequences (Supporting Information). Thus, the mean identification rate is quite high when an SH3- or SH2-specific HMM is used. For the individual backbone SH3 templates, the SUPERFAMILY detection rates vary between 100% and 25.5% (Table III). When the mean identify score is above 30%, the SUPERFAMILY detection rate is above 90%, except for 1i1i. Three SH3 families are detailed in Table IV. For the

Table IV
SUPERFAMILY Assignment of Designed SH3 Sequences

PDB code	Correct family[a]	Correct domain (name)[b]	Alternate domain (name)[c]	Sequences tested[d]
1cka	98.7	96.6 (1cka)	0.25 (1shf)	All sequences
1cka	100	99.8 (1cka)	0.01 (1gbr)	Low-energy sequences
1gcq	99.0	98.0 (Grb2, Human)	1.0 (Grb2, Caeno rh.)	All sequences
1gcq	100	98.9 (Grb2, Human)	1.1 (Grb2, Caeno rh.)	Low-energy sequences
1fyn	48.8	6.1 (1fyn)	13.5 (1fmk)	All sequences
1fyn	25.5	3.0 (1fyn)	2.5 (1fmk)	Low-energy sequences

[a] Percentage of designed sequences correctly assigned to the corresponding backbone template's SCOP family.

[b] Percentage of designed sequences assigned to the correct domain.

[c] Percentage assigned to the next most abundant, alternate domain.

[d] Results for all sequences or low-energy designed sequences.

templates 1cka and 1gcq, 100% of the low energy sequences were assigned to the SH3 family in SCOP, and 99% were correctly assigned to the correct domain within this family. The remaining low energy sequences were assigned to another, closely-related family member. The lowest rate overall (25.5%) is for the 1fyn backbone, which also yields one of the lowest identity scores for its low energy sequences (Table I). For 1fyn, only 25.5% of the low energy sequences are assigned by SUPERFAMILY to the correct, SH3 family. Surprisingly, the detection rate for all 1fyn sequences was much higher: 48.8% (Table IV). This is the only backbone template for which the low energy detection rate is lower than the overall rate.

The designed SH3 sequences that are not detected by SUPERFAMILY are characterized in Supporting Information by their similarity scores relative to the Pfam SH3 domains. Results are shown for four backbone templates. In all four cases, the histogram of scores is down-shifted with respect to the remaining, detected sequences. However, the shift is not very large (especially for the 1fyn case) and the histograms of detected and non-detected sequences overlap extensively. We also calculated the detection rates for random sequences, randomized so as to have a chosen, overall, mean identity to one of the 25 backbone templates. Strikingly, for random sequences with a 43% mean identity, the SUPERFAMILY detection rate is only 11%. 43% is close to the mean identity of our designed SH3 sequences (39% when the PBPs are reset to their experimental values), which have a 99% detection rate. At a 33% identity level, random sequences have a 1.5% detection rate; at the 53% identity level, the rate is still only 41%. The Pfam similarity scores of the random sequences are also quite low (Supporting Information).

Some backbone templates lead to poor detection by CDD of the designed sequences. This may reflect the uncertainty of domain classification for some proteins and/or an uneven quality of the CDD model set for SH3 or SH2 domains. The lowest CDD detection rates of the designed SH3 sequences are for 1pht, 1ycs, 1bb9, 1kjw, and 1vyv (Fig. 4), which give an average detection rate of just 7.7%. Notice that the mean identity score for the low energy sequences of these proteins is a respectable 33%. The rate of SUPERFAMILY detection for these five proteins is 78%, somewhat below the overall average. Evidently, the location, and not just the number of identical positions is important for CDD and SUPERFAMILY detection. All five proteins depart from the "classical" SH3 sequence length, with a mean length of 71.5 amino acids compared with about 57 for a "classical" SH3 domain. 1pht and 1kjw are the second- and third-longest proteins in our set. The protein 1vyv is not classified as an SH3 domain by the Pfam database. Thus, CDD (unlike SUPERFAMILY) does not include a model that directly represents it. Averaging over the other 19 SH3 proteins yields a CDD detection rate of 74%, instead of 61%.

PSI-BLAST detection

PSI-BLAST was used with several different position specific scoring matrices (PSSMs). The first was constructed from SH3 domains in the NR01 database of natural sequences (see Methods). With this "general" matrix, 30% of the designed SH3 sequences and 38% of the low-energy sequences were identified as SH3 sequences (using an *E*-value threshold of 0.1; see Table III and Fig. 4). A second set of 24 "backbone-specific" PSSMs was constructed: one for each SH3 domain. Each PSSM was constructed using a database of close homologues of the corresponding SH3 protein (with at least 45% identity; see Methods). With these PSSMs, the detection rate is much higher: 54% instead of 38% for the 14 "classical" SH3 proteins (Table III). Taken together, the hits recorded with the two types of PSSMs correspond to a detection rate for low energy sequences of 57.5%. The detection rate for 1fyn, 1shg, and 1ad5 remains poor, ~4.5%, consistent with their low calculated similarity scores. If we eliminate these three cases, the mean PSI-BLAST detection rate is a respectable 72% (Table III). For the SH2 domains, the overall detection rate is 77% (Supporting Information).

Finally, we studied the effect of the peptide binding positions (PBPs) in the SH3 proteins on the detection rates. These positions are responsible for the function of SH3 proteins, driving the binding of their polyproline ligands. For the 14 "classical" SH3 proteins, for these six positions, we systematically replaced the designed amino acids by the native ones. The resulting sequences were analyzed with CDD and PSI-Blast. PSI-Blast was run using the general PSSMs. For the low energy sequences with reset PBPs, the average detection rate jumps to 82%, compared with 38% for the unaltered sequences (see Table III and Fig. 4). Interestingly, 76% and 79% of the low-energy sequences from 1fyn and 1ad5 are now detected as SH3 sequences, compared with 0.1% previously. The detection rate for the 1shg sequences remains low, just 0.2%, but resetting four additional positions increases the detection rate to nearly 100% (Table III). Similar tendencies are seen for CDD and SUPERFAMILY detection. The overall SUPERFAMILY detection rates of the modified sequences is close to 99% (or 97% for the SH2 case; Supporting Information).

FROST threading analysis

A sample of designed sequences were evaluated by the FROST library of threading models.[75] For 14 SH3 and 14 SH2 templates, we evaluated 100 low energy sequences each (chosen randomly). Of the 2800 SH3 sequences, 433 (15.5%) were considered by FROST to be unlike natural

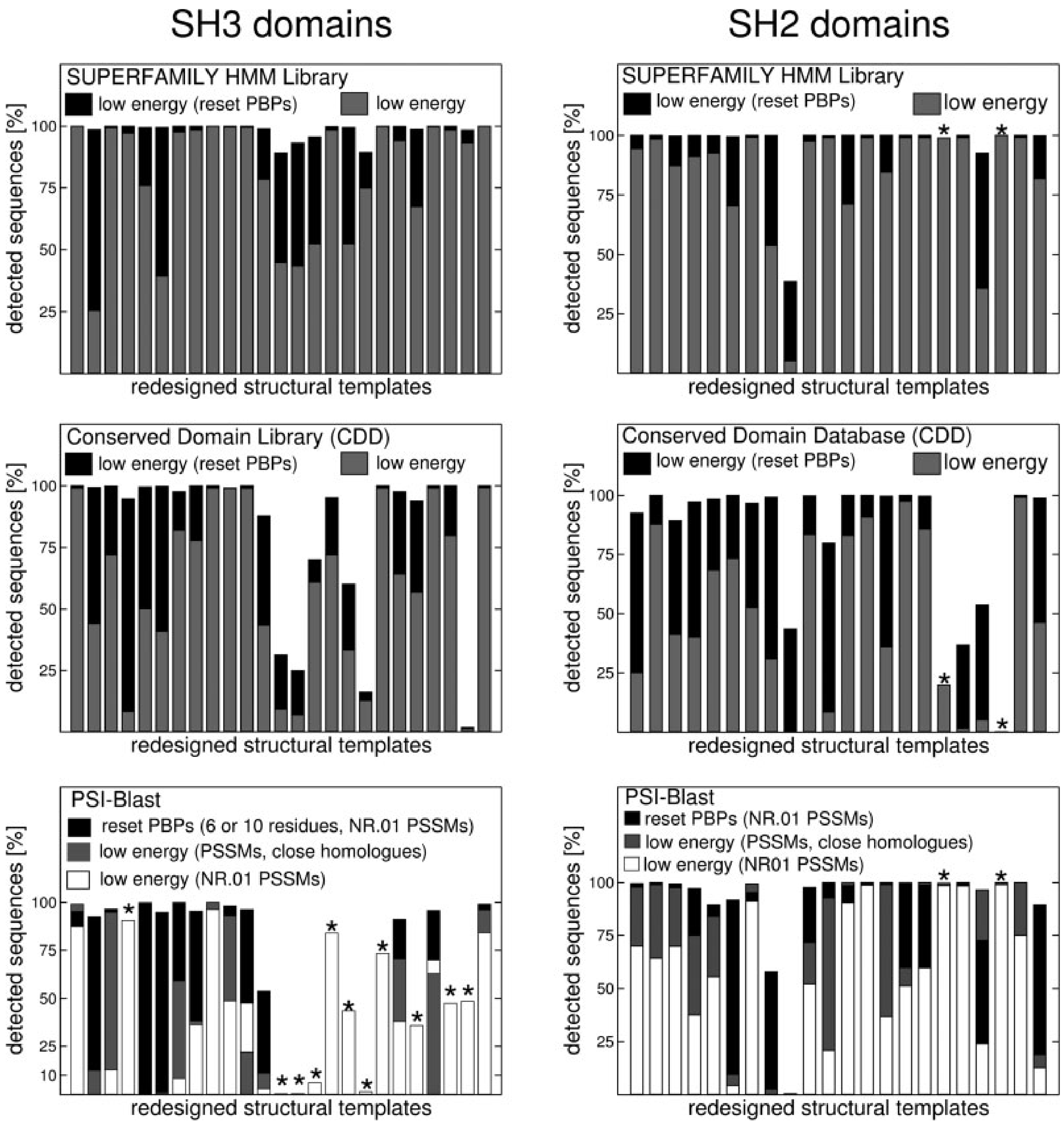

Figure 4

Designed sequences detected as SH3/SH2 domains by SUPERFAMILY (above), CDD (middle), or PSI-BLAST (below). Each column corresponds to one of the 24 SH3 templates (left) or the 22 SH2 templates (right), ordered as in Table I. Results are shown for 10,000 low energy sequences and for 10,000 "average" designed sequences (8000 for the SH2 domains). In some cases, the PBPs are reset to their experimental values. With PSI-BLAST (below), results are also shown for the SH3 domaines with four additional (less-conserved) PBPs reset, and two different PSSMs are used, constructed from the NR.01 database or a database of close homologues. Ten SH3 domains and two SH2 domains (1bf5 and 2cbl)) that are not "classical" are marked by stars (above each bar).

protein sequences. Another 8 (0.3%) were assigned to an erroneous 3D model (not an SH3 domain). For the other 2359 (84%), the program returned an average of 5.2 high-scoring hits, all belonging to the SH3 family of structures, along with 0.1 incorrect hits (not an SH3 structure). For the 2800 SH2 sequences, 5.6% were found to be unlike natural proteins; 2.5% were assigned to an erroneous 3D model; for the remainder (92%), the program returned 4.2 high-scoring, SH2 hits, and 1.2 hits from another structural family. Thus, for this sample of designed SH2 and SH3 sequences, the consensus prediction of FROST is overwhelmingly correct for 88% of the sequences. Notice that we did not reset PBPs for this test.

Homologue searching using designed sequences and PSSMs

We now consider the usefulness of the designed sequences for detecting SH3 and SH2 domains in sequence databanks. Indeed, our longer-term goal is to use computationally-designed sequences as a tool for fold recognition. Here, as a first step, we follow Larson et al.,[33,57] constructing "theoretical" PSSMs from the designed sequences and using them for homologue searching. We use both the unmodified, low energy, designed sequences, and modified sequences, where the peptide binding positions (PBPs) are reset to their experimental amino acid types. To construct the PSSMs, low-energy designed sequences from each SH3 or SH2 backbone are first pooled. The resulting database (6250 sequences for the SH3 case) is searched with PSI-BLAST, using one of the 24 native sequences as an initial query. Alternatively, a low-energy designed sequence, corresponding to a particular backbone template, can be used as the initial query. The resulting profile is refined through four iterations of PSI-BLAST. We thus obtain 25 theoretical SH3 PSSMs (one per SH3 backbone template, including the 1aze NMR structure). The whole procedure is typically repeated several times, using different sets of low energy sequences, chosen randomly; we refer to each repetition as a "cycle."

We systematically compare the performance of the theoretical PSSMs to "experimental" PSSMs, constructed from natural sequences. The same procedure is used, except that the NR01 database of natural sequences is searched, instead of the collection of designed sequences; we thus obtain 25 experimental SH3 PSSMs. We also compare to "random" PSSMs, constructed from sets of random sequences (see above). The sequences are randomized in such a way that the mean sequence identity to a particular native template is equal to a chosen value; say, 33%.

Figure 5 and Table V compare the performance of the experimental and theoretical SH3 and SH2 PSSMs within the PDB and the SwissProt databases. We report domains detected with an *E*-value of 0.1 or less. Within the PDB (removing redundant sequences, using a 95% identity threshold), the experimental PSSMs (applied successively) detected 67 SH3 domains, out of 69 in the PDB. With each theoretical PSSM, to increase the number of hits, we actually perform five "cycles", or rounds of searching; see above. The exact number of hits with the theoretical profiles depends on their detailed construction. If each PSSM is seeded by a designed sequence and the PBPs are unmodified (i.e., they are taken from the design), we retrieve 59 SH3 domains (including the 24 used here as backbone templates), and two false positives. If each theoretical PSSM is seeded with the native sequence instead of a designed sequence, we retrieve 61 SH3 domains, with the same two false positives. If the PBPs are reset to their experimental values, we retrieve 61 SH3 domains, with the same two false positives. For the SH2 case (Table V), the experimental PSSMs retrieve 131 SH2 domains; the theoretical PSSMs with reset PBPs retrieve 120, plus four false positives. We retrieve no "new" SH2 or SH3 sequences, not found by the experimental PSSMs.

Within SwissProt, the experimental PSSMs detected 772 SH3 sequences and 413 SH2 sequences, respectively. A PSSM constructed from only the 24 native SH3 templates behaved almost as well, detecting 763 sequences. With the theoretical PSSMs, we systematically performed 90 cycles. With PSSMs using designed queries and unmodified PBPs (and an *E*-value threshold of 0.1), we detected 462 SH3 sequences plus nine false positives, or 190 SH2 sequences plus seven false positives. If the PBPs are reset to their experimental values, 597 SH3 domains are detected, with three false positives, or 299 SH2 domains with nine false positives. Thus, our best PSSMs identify 77% of the SH3 domains and 72% of the SH2 domains. If we accept a confidence level of 95% by allowing up to 5% of false positives, we can raise the *E*-value threshold to 1. This leads to 675 SH3 hits, along with 29 false positives. Finally, the PSSMs can be improved by using designed sequences that have not only low energies but also high SUPERFAMILY scores. By selecting designed sequences with a SUPERFAMILY criterion, we combine the theoretical sequence model with experimental information. This improves the detection rate significantly, with 683 out of 772 SH3 domains now retrieved from SwissProt (88% detection rate) and 353 out of 413 SH2 domains (85% detection rate; Fig. 5).

In Figure 5, bottom panel, we report the number of retrieved sequences on a template-by-template basis. For 15 of the SH3 templates, the individual, designed PSSMs retrieve about half of the experimental homologues, if reset PBPs are used. For three SH3 templates (1fyn, 1shg, and 1ad5), the results are very poor, with almost no homologues retrieved. In SwissProt, once the first five designed PSSMs have been applied (the five leftmost SH3 templates in Fig. 5), the cumulated number of retrieved sequences reaches 514, compared with 614 when all 25 PSSMs are applied. Note that the order of the templates in Figure 5 is arbitrary. This suggests that in future applications, a particular SCOP family could be described using a subset of its member templates, perhaps as few as 20%. This subset could then lead to a reasonable detection rate in homologue searching. Results for the SH2 case are similar.

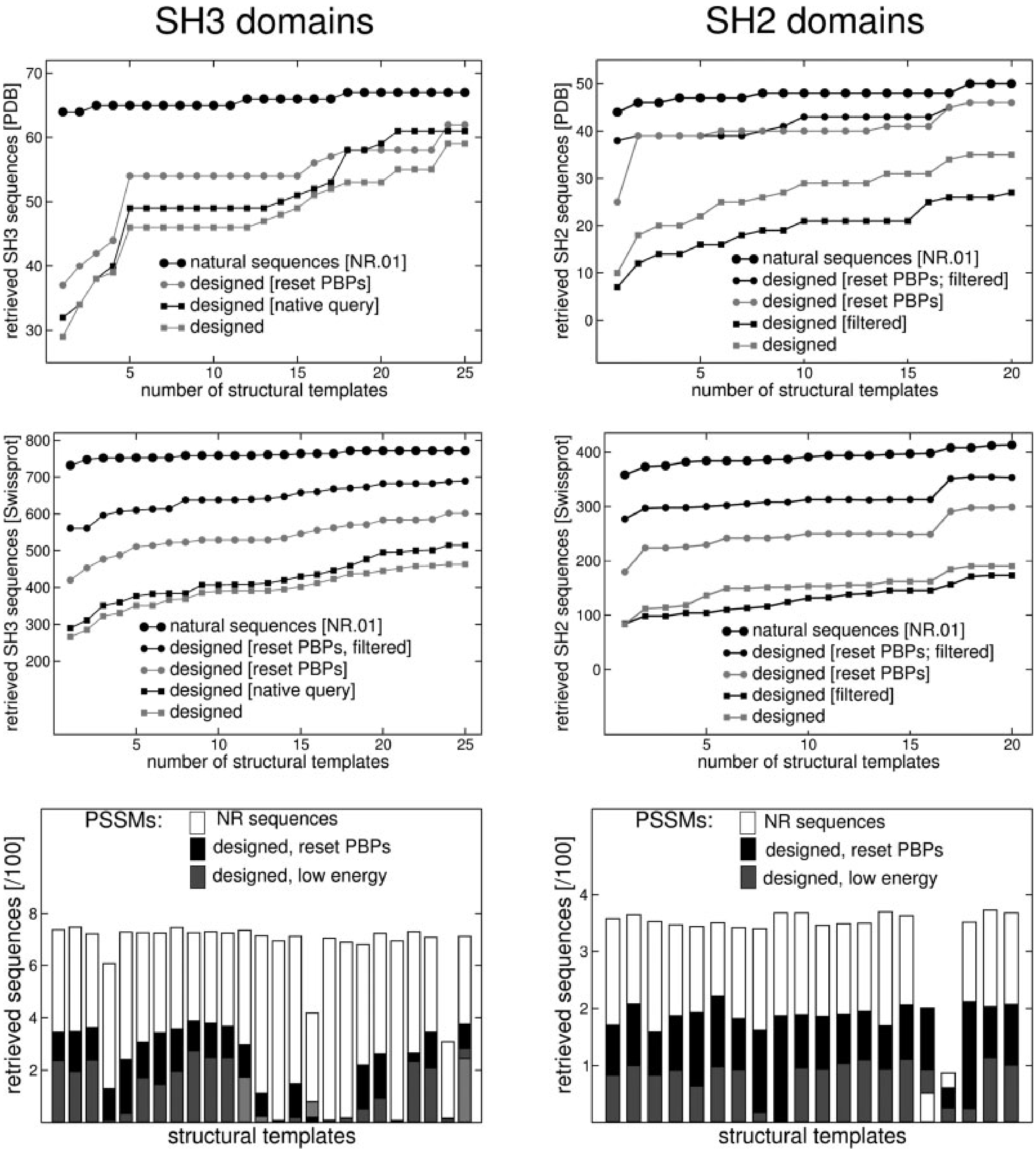

Figure 5

Detection of natural homologues using designed PSSMs. Detection in the PDB (top) and SwissProt (middle). Results are cumulated over the 24 SH3 templates (left) or the 22 SH2 templates (right) (ordered as in Table I) and their associated PSSMs. Bottom: The SwissProt results on a template-by template basis (non-cumulated). Homologues detected with experimental PSSMs are shown for comparison (large black dots in upper panesl). With the designed sequences, the PBPs can be reset to their experimental values or not, as indicated.

Table V
Sequences Retrieved from Experimental Databases Using Designed PSSMs

		SH3 domains			SH2 domains		
Experimental database[a]	[b]Sequences used	NR.01	Designed, Reset PBPs	Designed	NR.01	Designed, Reset PBPs	Designed
PDB	Homologues detected	67	62	59	131	120	102
	False positives	0	2	2	4	4	2
SwissProt	Homologues detected	772	597	462	413	299	190
	False positives	0	3	9	0	9	7

[a]Experimental database that is searched for SH3 or SH2 homologues.
[b]The sequences used to construct the PSSM are either experimental sequences from the NR.01 database, or designed sequences (with or without reset PBPs).

Comparing to random sequences

Figure 6 shows the number of sequences retrieved by several random PSSMs (see also Supporting Information). Results are compared with the designed PSSMs, constructed with a designed query and designed or reset PBPs. Either five or 90 cycles of searching were performed with each PSSM. Several comments can be made. First, the level of performance of the designed sequences is much better than that of random sequences having the same identitity level (when compared with their SH3 or SH2 templates). This is consistent with the very low level of detection of the random sequences by SUPERFAMILY (about 1% in the SH3 case; see above).

Second, the performance of the designed SH3 sequences is comparable to random sequences that have a 47% sequence identity to their templates. In the SH2 case, the designed sequences have a mean identity of 29% with their templates; random sequences with a 33% identity give a similar level of SwissProt detection (but many more false positives; see Supporting Information). The detailed variations of the curves corresponding to the random and designed sequences (Fig. 6) are roughly similar; in particular, the templates that lead to very poor homologue detection are mostly the same in the two cases. Third, if the PBPs are reset to their experimental types, the performance of the designed SH3 (SH2) sequences is comparable to random sequences that have a 47% (43%) sequence identity to their templates. With reset PBPs, the designed SH3 sequences have a mean identity of about 39% to their templates (Table I). Fourth, the relative performance of the random sequences is better with 90 cycles than with five. The diversity of the random sequences is much higher than that of the designed sequences, with mean entropies of 4.5–5, compared with ~2.9 for the designed and experimental SH3 sequences. Thus, additional cycles are more likely to yield significantly different PSSMs that give somewhat different hits.

Finally, we should bear in mind that we compare here the performance of the random and designed PSSMs for PSI-Blast searching. It is likely that PSSMs do not fully exploit all the information contained in the designed sequences (see Discussion); in this sense, the quality of the designed SH3 sequences may actually be superior to the random, 53%-identity ones (as suggested by their much higher SUPERFAMILY detection rate; see above).

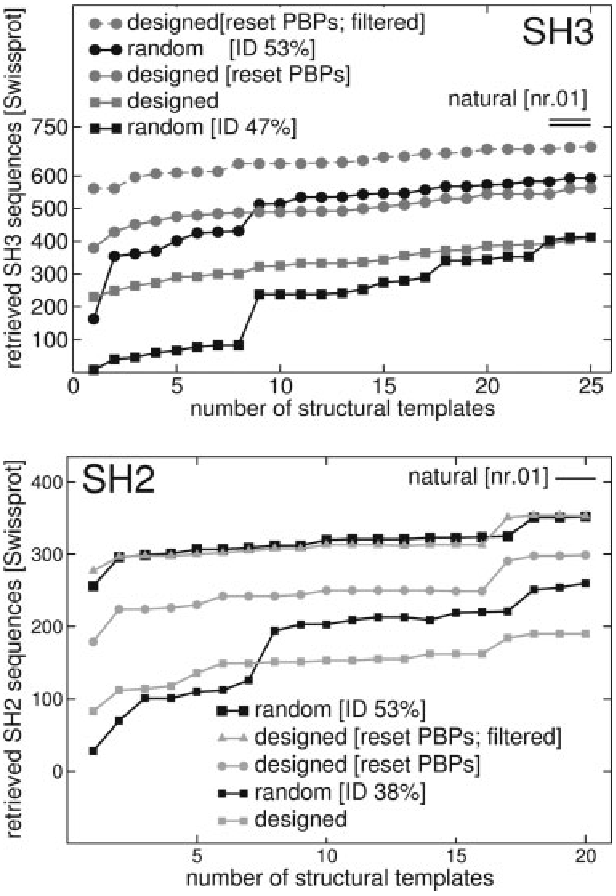

Figure 6
Comparing random and designed PSSMs for the detection of natural homologues in SwissProt; same representation as in Fig. 5. The sequences are randomized so as to have a set identity with their templates, as indicated. Horizontal lines (upper right) indicate the number of sequences retrieved by two natural PSSMs, constructed respectively from NR01 (as in Fig. 5) and from the 25 backbone templates (lower of the two bars, SH3 panel). In some cases, the designed sequences have been "filtered": the corresponding PSSMs are constructed from designed sequences that have high SUPERFAMILY scores, as well as low energies.

In particular, we can establish that the designed sequences contain information on the covariances between different amino acid positions. Indeed, we have generated random sequences using the experimental, SUPERFAMILY, SH3 sequence profile (i.e., the probabilities of each amino acid type at each position are given by SUPERFAMILY). The resulting sequences were then evaluated using our energy function. Specifically, we did extensive rotamer exploration for each sequence, as for the designed sequences, above (see Methods). The resulting energies were typically several hundred kcal/mol higher than the energies of the designed sequences, because of steric conflicts in the protein interior (data not shown). Thus, random sequences that present not only a high average sequence identity (compared to the experimental template sequence) but actually obey the detailed SUPERFAMILY profile, usually do not pack into the correct 3D structure, unlike the designed sequences. Details of this analysis will be published elsewhere. This is a strong indication that there is covariance information in the designed sequences, which is not exploited in PSI-Blast searching.

Comparing retrieved and nonretrieved sequences

To improve the potential of CPD for homologue detection, it is important to understand why certain natural sequences are retrieved by the designed PSSMs but others are not. Thus, 155 SH3 sequences were retrieved from SwissProt by the experimental PSSMs but not the theoretical PSSMs (with reset PBPs). For these 155 "non-retrieved" sequences, we examined the level of confidence with which the experimental PSSMs retrieve them. Similarly, we computed their *E*-values for detection by CDD and SUPERFAMILY. These *E*-values are reported in Supporting Information, along with the PSI-BLAST *E*-values of the full set of 772 retrieved sequences. Of the 155 "non-retrieved" sequences, about half have PSI-BLAST *E*-values greater than 0.0001 (with the experimental PSSMs); one third (50 sequences) have *E*-values of about 0.001 or more. The CDD *E*-values are in the same range. The SUPERFAMILY *E*-values are lower, reflecting the better detection achieved with a large library of HMM models (one model for each SCOP domain). The bulk of the sequences retrieved by the natural and designed PSSMs have much lower *E*-values (with both PSI-BLAST and CDD). Thus, many of the sequences that are missed by the designed PSSMs do not have a very strong SH3 character, as measured by the PSI-BLAST and CDD *E*-values.

In a similar way, we can characterize many of the nonretrieved sequences by the *E*-values attributed by the designed PSSMs. Indeed, as mentioned above, many of these sequences are detected by our designed PSSMs, but with *E*-values greater than 0.1 (our nominal detection threshold). Thus, 56 of the 155 nonretrieved sequences have *E*-values between 0.1 and 1. Another 26 have *E*-values between 1 and 4. More generally, although the designed PSSMs perform rather well, they systematically return higher *E*-values than the natural PSSMs (Supporting Information).

CONCLUDING DISCUSSION

The design method used here has significant limitations, similar to several other implementations. They include the simple molecular mechanics force field and the sidechain rotamer approximation. The fixed backbone approximation is also potentially serious,[47,80–82] but is compensated for here by considering over 20 representatives of each protein family, each having a somewhat different backbone structure. Additional tests show, for example, that using multiple backbones from normal mode calculations does not improve the similarity scores and entropy estimates, when compared with the family approach used here (Supporting Information). More serious limitations are the implicit solvent model, whose nature is partly dictated by the need for pairwise additivity, and the crude, tripeptide model of the unfolded state. Finally, the selection criterion, maximizing protein stability, is plausible but not ideal. Indeed, the timescales of cellular dynamics, and adaptation of the cell to a fluctuating environment require that proteins should not be too stable, and functional residues, such as those in enzyme active sites, are often thermodynamically destabilizing.[83,84] In theory, we should also include negative design in the selection process, to exclude sequences that might prefer to adopt another, non-SH3 or non-SH2 fold. In practice, by testing our sequences against the full library of SUPERFAMILY models, we ensure that over 80% are not suited to other folds (over 97% when the PBPs are reset).

Despite these limitations, the quality of the designed sequences is good. Their identity scores with respect to the corresponding native templates are comparable to other recent work (see discussion in Ref. 61), and their Pfam similarity scores overlap extensively those of the natural SH3 and SH2 sequences found in the large Pfam sets. The SUPERFAMILY library of HMMs recognizes 81% of our low energy SH3 sequences as SH3-like; if the six PBPs are reset to their experimental types, the detection rate is 99%. In contrast, for random sequences with a 43% mean identity level, the SUPERFAMILY detection rate is 11%. The designed SH3 and SH2 sequences also capture the diversity of the natural sets, as measured by the sequence entropy. This overall sequence quality partly reflects the careful parameterization of the energy function and solvent model.[62,64] It also reflects the importance of protein stability in the evolution of natural sequences. Indeed, the similarity scores are especially high for the core positions, where the most probable amino acids are very similar in the designed and experimental sets. Clearly, these positions play an important role in stabilizing the core 3D structure of the protein. The concept of core positions has been illustrated by recent experimental and computational studies of protein stability and design.[46,85,86]

The performance of the designed sequences for homologue detection was investigated by PSI-Blast searching in the PDB and SwissProt. The designed SH3 PSSMs typically detect about 1/2 of the SH3 domains detected by the native backbone template sequences. When the results are cumulated over several cycles and all 25 sets of sequences, the SH3 detection rate is about 60% (462 out of 772 sequences), with nine false positives. With reset PBPs and a native initial query, the detection rate rises to 77%. The SH2 results are slightly poorer (72% detection). The designed SH3 (SH2) sequences have an average identity of 33% (29%) to their templates, or 39% (35%) when the PBPs are reset. To obtain the same SH3 detection level with random sequences, the identity rate should be about 45–50%. The number of false positives with the random sequences is notably higher than with the designed ones: 20 at the 47% identity level, compared with 3–9 with the designed SH3 sequences. Although the detection level is much lower than 100% for the SH3 and SH2 families, it may be that the method can already help detect new homologues in some other families that are not as well-studied. In particular, new protein folds are still being discovered.[87,88]

An important limitation of the designed sequences is that they do not include explicit selection for biological function. To alleviate this, we manually reset a few peptide binding positions to their experimental types after the design was performed. An alternative would have been to constrain them from the outset, during the design calculations.[60] This might improve the quality of the designed sequences, because the PBPs could indirectly affect other, nearby positions. In particular, we pointed out above that some of the overestimated residue entropies might be due to the lack of constraints on the PBPs during the design. Because the PBPs are not strictly conserved, it might also be preferable to bias them but not completely fix their identities. In this work, we identified the functional positions from the literature. To apply our methodology to other protein families on a large scale, it would be preferable to use an automatic criterion. One possibility would be to perform an initial round of design and identify a subset of residues with the largest discrepancies between experimental and computed entropies, then fix a subset in their experimental identities while a second round of design is performed.

More generally, a direction to be explored in the future is the combined use of experimental and designed sequences. Here, we infuse experimental sequence information either by constraining functional positions, or by extracting from the design sequences those that not only have a low energy but provide a high SUPERFAMILY score (Fig. 5). There are many other possibilities to explore. One idea would be to bias the stochastic exploration of sequence space, in order to search more thoroughly the vicinity of the experimental sequences. This could be done through some form of umbrella sampling;[89–91] another possibility would be to initiate the heuristic cycles of our method not with completely random sequences (see Methods), but with sequences generated by one of the SUPERFAMILY HMMs.

The potential of the method for general application is supported further by preliminary results for two additional protein families: PDZ domains and Kunitz domains. We have done calculations for 18 and 9 members of these families, respectively, which will be described elsewhere. The proteins are distinctly larger than the SH3 domains. The quality of the designed sequences appears to be comparable to the SH2 and SH3 cases.

It is striking that the random sequences perform moderately well for PSI-Blast searching, even though very few of them are recognized as SH3 or SH2 domains by SUPERFAMILY. Clearly, SUPERFAMILY relies more on the core positions for its detection. It may be that the designed sequences would outperform the random sequences more strongly in homologue searching if a more sophisticated method were used. For example, the designed sequences could be used to parameterize a profile HMM, similar to those used by SUPERFAMILY. Modeling the designed sequences through a PSSM has at least one severe limitation, because it effectively replaces the sequence collection by a single, average sequence. This averaging destroys the information on covariances between individual positions in the protein, such as those induced by the 3D protein structure. Indeed, we generated random sequences that obeyed the SUPERFAMILY SH3 and SH2 profiles and studied their 3D packing and folding free energies. The energies were typically several hundred kcal/mol worse than for the designed sequences, precisely because the 3D interactions and correlations were not taken into account (data not shown). In contrast, the designed sequences are based on a sophisticated and rather accurate model of the 3D structure and interactions, so that it is plausible that their performance is degraded, relative to random sequences, when part of this information is averaged out. HMM models probably will not adequately describe such long-range correlations either, so that sequence models of a different type should be explored; stochastic grammars are one possibility.[72] In the future, we hope to combine the considerable computer power provided by Proteins@Home with improved statistical models of the designed sequences, and to realize the potential of CPD as a powerful new tool for protein fold recognition.

ACKNOWLEDGMENTS

We thank the many volunteers who have participated in the Proteins@Home project and contributed computer cycles to this work. See biology.polytechnique.fr/proteinsathome for a complete list of participants. We thank the BOINC development community for testing the alpha version of Proteins@Home. We thank Jean-François Gibrat, Joël Pothier, Yves-Henri Sanejouand, Anne Lopes, and Yann Ponty for discussions, Yves-Henri Sanejouand for performing normal mode calculations on some of our proteins, and the ANR High Performance Computing program for support.

REFERENCES

1. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho C, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons J, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rotherberg J. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005;437:376–380.
2. Todd AE, Marsden RL, Thornton JM, Orengo CA. Progress of structural genomics initiatives: an analysis of solved target structures. J Mol Biol 2005;348:1235–1260.
3. George RA, Spriggs RV, Bartlett GJ, Gutteridge A, MacArthur MW, Porter CT, Al-Lazikani B, Thornton JM, Swindells MB. Effective function annotation through catalytic residue conservation. Proc Natl Acad Sci USA 2005;102:12229–12304.
4. Sillitoe I, Dibley M, Bray J, Addou S, Orengo C. Assessing strategies for improved superfamily recognition. Prot Sci 2005;14:1800–1810.
5. Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure protein families. Nat Rev Mol Cell Biol 2007;8:995–1005.
6. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler L. GenBank. Nucleic Acids Res 2006;34:D16–D20.
7. Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC. The genomes on line database (GOLD) v.2: a monitor of genome projects worldwide. Nucleic Acids Res 2006;34:D332–D334.
8. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
9. Andreeva A, Howorth D, Brenner SE, Hubbard JJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 2004;32:D226–229.
10. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 2008;36: 419–425.
11. Pearl F, Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, Bennett C, Marsden R, Grant A, Lee D, Akpor A, Maibaum M, Harrison A, Dallman T, Reeves G, Diboun I, Addou S, Lise S, Johnston C, Sillero A, Thornton J, Orengo C. The CATH domain structure database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res 2005;33:D247–251.
12. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known 3-dimensional structure. Science 1991;253:164–170.
13. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 2000;29:291–325.
14. Venclovas C. Comparative modeling of CASP4 target proteins: Combining results of sequence search with three-dimensional structure assessment. Proteins 2001;45 (Suppl 5):47–54.
15. Schwede T, Kopp J, Guex N, Peitsch MC. Swiss-Model: an automated protein homology-modeling server. Nucleic Acids Res 2003;31:3381–3385.
16. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zang Z, Miller W, Lippman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
17. Schaffer AA, Aravind L, Madden TL, Shavirin JL, Spouge S, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001;29:2994–3005.
18. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWweese-Scott C, Geer LY, Gwadz M, He SQ, Hurwitz DI, Jackson JD, Ke ZX, Lanczycki CJ, Liebert CA, Liu CL, Lu F, Marchler GH, Mullokandov M, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Yamashita RA, Yin JJ, Zhang DC, Bryant SH. CDD: a conserved domain database for protein classification. Nucleic Acids Res 2005;33:D192–D196.
19. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. Hidden Markov models in computational biology: applications to protein modelling. J Mol Biol 1994;235:1501–1531.
20. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res 1998;26:320–322.
21. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res 1998;26:260–262.
22. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 2001;313:903–919.
23. Madera M, Gough JA comparison of profile hidden Markov model procedures for remote homology detection. Nucleic Acids Res 2002;32:4321–4328.
24. Brown DP, Krishnamurthy N, Sjoelander K. Automated protein subfamily identification and classification. Plos Comp Biol 2007;3:1526–1538.
25. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, Sillitoe I, Yeats C, Thornton JM, Orengo CA. The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res 2007;35:D291–297.
26. Madera M, Vogel C, Kummerfeld SK, Chothia C, Gough J. The SUPERFAMILY database in 2004: additions and improvements. Nucleic Acids Res 2004;32:D235–D239.
27. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. Nucleic Acids Res 2004;10:D138–D141.
28. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A. The Pfam protein families database. Nucleic Acids Res 2008;36:D281–D288.
29. Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P. SMART 4.0: towards genomic data integration. Nucleic Acids Res 2004;10:D142–D144.
30. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 2003;4:41.
31. Wilson D, Madera M, Vogel C, Chothia C, Gough J. The SUPERFAMILY database in 2004: additions and improvements. Nucleic Acids Res 2007;35:D308–D313.

32. Koehl P, Levitt M. De novo protein design. II. Plasticity in sequence space. J Mol Biol 1999;293:1183–1193.
33. Larson S, Garg A, Desjarlais J, Pande V. Increased detection of structural templates using alignments of designed sequences. Proteins 2003;51:390–396.
34. Larson S, England JE, Desjarlais J, Pande V. Thoroughly sampling sequence space: Large-scale protein design of structural ensembles. Protein Sci 2002;11:2804–2813.
35. Ponder J, Richards FM. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. J Mol Biol 1988;193:775–791.
36. Hellinga H, Richards F. Optimal sequence selection in proteins of known structure by simulated evolution. Proc Natl Acad Sci USA 1994;91:5803–5807.
37. Dahiyat B, Mayo S. Protein design automation. Protein Sci 1996;5:895–903.
38. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. High-resolution protein design with backbone freedom. Science 1998;1998:1462–1467.
39. Dokholyan NV, Shakhnovich EI. Understanding hierachical protein evolution from first principles. J Mol Biol 2001;312:289–307.
40. Desjarlais J, Handel T. Sidechain and backbone flexibility in protein core design. J Mol Biol 1999;289:305–318.
41. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 2000;97:10383–10388.
42. Wernisch L, Héry S, Wodak S. Automatic protein design with all atom force fields by exact and heuristic optimization. J Mol Biol 2000;301:713–736.
43. Kuhlman B, Dantas G, Ireton G, Varani G, Stoddard B, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science 2003;302:1364–1368.
44. Dwyer M, Looger L, Hellinga H. Computational design of a biologically active enzyme. Science 2004;304:1967–1971.
45. Havranek J, Harbury P. Automated design of specifity in molecular recognition. Nat Struct Biol 2003;10:45–52.
46. Ventura S, Serrano L. Designing proteins inside out. Proteins 2004;56:1–10.
47. Saunders C, Baker D. Recapitulation of protein family divergence using flexible backbone protein design. J Mol Biol 2005;346:631–644.
48. Wollacott AM, Zanghellini A, Murphy P, Baker D. Prediction of structures of multidomain proteins from structures of the individual domains. Protein Sci 2007;16:165–175.
49. Swift J, Wehbi WA, Kelly BD, Stowell XF, Saven JG, Dmochowski IJ. Design of functional ferritin-like proteins with hydrophobic cavities. J Am Chem Soc 2006;128:6611–6619.
50. Kang SG, Saven JG. Computational protein design: structure function and combinatorial diversity. Curr Opin Chem Biol 2007;11:329–334.
51. Koehl P, Levitt M. De novo protein design. I. In search of stability and specificity. J Mol Biol 1999;293:1161–1181.
52. Koehl P, Levitt M. Structure-based conformational preferences of amino acids. Proc Natl Acad Sci USA 1999;96:12524–12529.
53. Hubner IA, Deeds EJ, Shakhnovich EI. Understanding ensemble protein folding at atomic detail. Proc Natl Acad Sci USA 2006;103:17747–17752.
54. Pokala N, Handel T. Energy functions for protein design I: Efficient and accurate continuum electrostatics and solvation. Protein Sci 2004;13:925–936.
55. Pokala N, Handel TM. Energy functions for protein design: Adjustement with protein-protein complex affinities models for the unfolded state and negative design of solubility and specificity. J Mol Biol 2005;347:203–227.
56. Chowdry AB, Reynolds KA, Hanes MS, Voorhies M, Pokala N, Handel T. An object-oriented library for computational protein design. J Comp Chem 2007;28:2378–2388.
57. Larson S, Pande V. Sequence optimization for native stability determines the evolution and folding kinetics of a small protein. J Mol Biol 2003;332:275–286.
58. Dantas G, Kuhlman B, Callender D, Wong M, Baker D. A large test of computational protein design: Folding and stability of nine completely redesigned globular proteins. J Mol Biol 2003;332:449–460.
59. Raha K, Wollacott AM, Italia MJ, Desjarlais JR. Prediction of amino acid sequence from structure. Protein Sci 2000;9:1106–1119.
60. Ding F, Dokholyan NV. Emergence of protein fold families through rational design. PLoS Comp Biol 2006;2:e85.
61. Schmidt Am Busch M, Lopes A, Mignon D, Simonson T. Computational protein design: software implementation parameter optimization and performance of a simple model. J Comp Chem 2008;29: 1092–1102.
62. Schmidt am Busch M, Lopes A, Amara N, Bathelt C, Simonson T. Testing the coulomb/accessible surface area solvent model for protein stability ligand binding and protein design. BMC Bioinformatics 2008;9:148–163.
63. Tuffery P, Etchebest C, Hazout S, Lavery R. A new approach to the rapid determination of protein side chain conformations. J Biomol Struct Dyn 1991;8:1267.
64. Lopes A, Aleksandrov A, Bathelt C, Archontis G, Simonson T. Computational sidechain placement and protein mutagenesis with implicit solvent models. Proteins 2007;67:853–867.
65. Brooks B, Bruccoleri R, Olafson B, States D, Swaminathan S, Karplus M. Charmm: a program for macromolecular energy minimization and molecular dynamics calculations. J Comp Chem 1983;4: 187–217.
66. Lee B, Richards F. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.
67. Street A, Mayo S. Pairwise calculation of protein solvent-accessible surface areas. Folding Des 1998;3:253–258.
68. Dahiyat B, Mayo S. De novo protein design: fully automated sequence selection. Science 1997;278:82–87.
69. Brünger AT. X-PLOR version 3.1, A system for X-ray crystallography and NMR. Yale New Haven: University Press; 1992.
70. Anderson DP. BOINC: A system for public-resource computing and storage. In 5th IEEE/ACM International Workshop on Grid Computing, USA: IEEE Computer Society Press; 2004.
71. Simonson T, Mignon D, Schmidt Am Busch M, Lopes A, Bathelt C. The inverse protein folding problem: structure prediction in the genomic era. In Distributed & grid computing – science made transparent for everyone principles applications and supporting communities. Berlin; Tektum Publishers; 2008.
72. Durbin R, Eddy SR, Krogh A, Mitchison G. Biological sequence analysis. Cambridge University Press, Cambridge, 2002.
73. Murphy LR, Wallqvist A, Levy RM. Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng 2000;13:149–152.
74. Launay G, Mendez R, Wodak SJ, Simonson T. Recognizing protein-protein interfaces with empirical potentials and reduced amino acid alphabets. BMC Bioinformatics 2007;8:270–291.
75. Marin A, Pothier J, Zimmermann K, Gibrat JF. FROST: a filter-based fold recognition method. Proteins 2002;49:493–509.
76. Kawabata T. MATRAS: a program for protein 3D structure comparison. Nucleic Acids Res 2003;13:3367–3369.
77. Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer ELL, Bateman A. Pfam: clans web tools and services. Nucleic Acids Res 2006;34:D247–251.
78. Jaramillo A, Wernisch L, Héry S, Wodak S. Folding free energy function selects native-like protein sequences in the core but not on the surface. Proc Natl Acad Sci USA 2002;99:13554–13559.
79. Waksman G, Shoelson S, Pant N, Cowburn D, Kuriyan J. Binding of a high affinity phosphotyrosyl peptide to the src SH2 domain: crystal structures of the complexed and peptide-free forms. Cell 1993;72:779–790.
80. Smith CA, Kortemme T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant sidechain prediction. J Mol Biol 2008;380:742–756.

81. Friedland GD, Linares AJ, Smith CA, Kortemme T. A simple model of backbone flexibility improves modeling of sidechain conformational variability. J Mol Biol 2008;380:757–774.
82. Fung HK, Floudas CA, Taylor MS, Zhang L, Morikis D. Towards full-*sequence de novo* protein design with flexible templates for human beta-defensin-2. Biophys J 2008;94:584–599.
83. Shoichet BK, Baase WA, Kuroki R, Matthews BW. A relationship between protein stability and protein function. Proc Natl Acad Sci USA 1995;92:452–456.
84. Elcock AH. Prediction of functionally important residues based solely on the computed energetics of protein structure. J Mol Biol 2001;312:885–896.
85. Koehl P, Levitt M. Protein topology and stability define the space of allowed sequences. Proc Natl Acad Sci USA 2002;99:1280–1285.
86. Morrra G, Colombo G. Relationship between energy distribution and fold stability: Insights from molecular dynamics simulations of native and mutant proteins. Proteins 2008;72:660–672.
87. Liu X, Fang K, Wang W. The number of protein folds and the distribution over families in nature. Proteins 2004;54:491–499.
88. Guerler A, Knapp EW. Novel protein folds and their nonsequential structural analogs. Protein Sci 2008;17:1374–1382.
89. Frenkel D, Smit B. Understanding molecular simulation. New York: Academic Press; 1996.
90. Lartillot N, Philippe H. Computing Bayes factors using thermodynamic integration. Syst Biol 2006;55:195–207.
91. Kleinman CL, Rodrigue N, Bonnard C, Philippe H, Lartillot N. A maximum likelihood framework for protein design. BMC Bioinformatics 2006;7:Art. 326.