Computational protein design: Software implementation, parameter optimization, and performance of a simple model

ARTICLE *in* JOURNAL OF COMPUTATIONAL CHEMISTRY · MAY 2008

Impact Factor: 3.59 · DOI: 10.1002/jcc.20870 · Source: PubMed

4 AUTHORS, INCLUDING:

Anne Lopes
Université Paris-Sud 11
14 PUBLICATIONS **142** CITATIONS
SEE PROFILE

Thomas Simonson
École Polytechnique
110 PUBLICATIONS **20,143** CITATIONS
SEE PROFILE

Computational Protein Design: Software Implementation, Parameter Optimization, and Performance of a Simple Model

MARCEL SCHMIDT AM BUSCH, ANNE LOPES, DAVID MIGNON, THOMAS SIMONSON
Laboratoire de Biochimie (CNRS UMR7654), Department of Biology, Ecole Polytechnique, 91128 Palaiseau, France

Received 13 July 2007; Revised 28 September 2007; Accepted 7 October 2007
DOI 10.1002/jcc.20870
Published online 10 December 2007 in Wiley InterScience (www.interscience.wiley.com).

Abstract: Computational protein design will continue to improve as new implementations and parameterizations are explored. An automated protein design procedure is implemented and applied to the full redesign of 16 globular proteins. We combine established but simple ingredients: a molecular mechanics description of the protein where nonpolar hydrogens are implicit, a simple solvent model, a folded state where the backbone is fixed, and a tripeptide model of the unfolded state. Sequences are selected to optimize the folding free energy, using a simple heuristic algorithm to explore sequence and conformational space. We show that a balanced parametrization, obtained here and in our previous work, makes this procedure effective, despite the simplicity of the ingredients. Calculations were done using our Proteins @ Home distributed computing platform, with the help of several thousand volunteers. We describe the software implementation, the optimization of selected terms in the energy function, and the performance of the method. We allowed all amino acids to mutate except glycines, prolines, and cysteines. For 15 of the 16 test proteins, the scores of the computed sequences were comparable to those of natural homologues. Using the low energy computed sequences in a BLAST search of the SWISSPROT database, we could retrieve natural sequences for all protein families considered, with no high-ranking false-positives. The good stability of the designed sequences was supported by molecular dynamics simulations of selected sequences, which gave structures close to the experimental native structure.

 J Comput Chem 29: 1092–1102, 2008

Key words: protein engineering; molecular mechanics; inverse protein folding; distributed computing; combinatorial optimization

Introduction

With the development of genomics and rapidly growing databases of protein sequences, protein structure prediction is increasingly important. The protein folding problem, or prediction of structure from sequence alone, remains a major challenge. In the 80's, the inverse folding problem was formulated: instead of predicting the 3D structure from the sequence, one considers a given backbone structure and predicts the amino acid sequences that fold into it.[1–4] Computational protein design addresses this inverse problem. The methods developed over the last two decades and their applications[4–26] represent rigorous tests of our understanding of the mechanisms that shape protein sequences and structures. They provide not only tools for engineering new proteins,[27–32] but also methods for structure prediction. Indeed, protein design can be seen as an evolutionary model that evaluates the sequence evolution within a given protein family when a structural or functional constraint is applied.[15,33–35] A 25–30% sequence homology between proteins is generally sufficient for them to share a common fold, and computational design methods have the potential to become a fold-recognition method that can categorize unknown proteins into previously determined fold families.[15,22,34,35] As new implementations and parameterizations are explored, computational protein design will continue to improve.

We use simple, existing ingredients to implement a method whose performance appears to be competitive with several existing, more sophisticated methods.[10,36] We benefit from our previous testing and optimization of a simple implicit solvent model, and we describe the further optimization of other parameters in the

This article contains supplementary material available via the Internet at http://www.interscience.wiley.com/jpages0192-8651/suppmat

Correspondence to: T. Simonson; e-mail: thomas.simonson@polytechnique.fr

Contract/grant sponsor: Agence Nationale pour la Recherche

energy function. Folded conformations have the experimental, native backbone conformation, in contrast to some recent methods that use multiple backbone conformations.[8,11,15,22–24] Sidechains occupy rotamers from a simple, backbone-independent rotamer library.[37] Sequences are selected to optimize the folding free energy, using a simple, heuristic algorithm to explore sequence and conformational space.[10] The method employs a molecular mechanics description of the protein, with a force field that treats nonpolar hydrogens implicitly.[38] Solvent is described implicitly, using a very simple model that includes a screened Coulomb energy and a solvent-accessible surface energy. This Coulomb/Accessible Surface Area, or CASA model was recently parameterized and tested for protein stability and for sidechain reconstruction, which are key steps in protein design.[39] The unfolded state model is also very simple, with each amino acid interacting only with nearby backbone groups (tripeptide model) and with solvent. An additional, empirical correction is derived here (similar to earlier work[9,11,40]), which depends on the amino acid type, and is chosen to provide a realistic overall amino acid composition. Energy calculations are performed using our Proteins @ Home distributed computing platform (biology.polytechnique.fr/proteinsathome), which will be described in detail elsewhere.[41]

We test the capabilities of the model for the nearly-complete redesign of 16 proteins (only glycines, prolines, and cysteines are not allowed to mutate). The test set includes eight SH3 domains and a diverse set of eight other proteins, from eight different families in the SCOP classification.[42] A longer-term goal is to study the inverse folding problem and the use of protein design for protein fold recognition.[22,41] For this, we want to produce sets of sequences that are realistic but also large and diverse. This contrasts with some applications where only a few, highly-optimized sequences are sought. For 15 of the test proteins, the computed sequences have similarity scores comparable to natural homologues. They also display a good structural stability when subjected to molecular dynamics simulations using a high quality, generalized Born, implicit solvent model.[39,43]

Overall, the method implemented here is similar to several existing methods but uses somewhat simpler ingredients. For example, our method ressembles that proposed by Wodak and coworkers,[10,36] but uses a simpler force field and a simpler rotamer library. Nevertheless, our implementation achieves a performance that is distinctly improved compared to Wodak et al. and appears to be competitive with several other, still more complex implementations, including methods that use a flexible backbone and/or a generalized Born solvent. The improvements can be partly attributed to our reparametrization of the CASA solvent model[39,44] and a simple but carefully optimized unfolded state model.

Methods

Folded and Unfolded States

Sequences and structures are selected based on their folding free energies, ΔG_{fold}, the difference between the free energy of their folded and unfolded states. In the folded state, the coordinates of the protein backbone are kept fixed, while sidechains occupy rotamers from the backbone-independent Tuffery library.[37] The backbone conformation was obtained by subjecting the protein 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 led to an rms deviation from the experimental structure of 0.56—0.90 Å (depending on the protein) and a protein radius of gyration about 0.1 Å smaller than 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). 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 (lysozyme, ribonuclease A, bovine pancreatic trypsin inhibitor, Staphylococcal nuclease, and the α-toxin). 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; the optimization of e_{X} is described later (Empirical correction to the unfolded state energy section). For a given amino acid sequence, the unfolded state free energy is obtained by summing the contributions $E_{\text{X}} + e_{\text{X}}$ of the individual amino acids.

In protein design, we perform rounds of random mutagenesis, transforming a given sequence A into a new sequence B. By comparing the folding energies for A and B, we can determine which sequence is most favorable. Because our energy function is pairwise additive (see later), and because the backbone structure is fixed in the folded state, we account correctly for ΔG_{fold} if we include all pairwise interactions between sidechains and between each sidechain and the backbone. In particular, interactions between different portions of the protein backbone cancel when two sequences are compared, both in the folded or the unfolded state, so that no important interactions are missed through the tripeptide unfolded model.

Effective Energy Function

The effective energy function was described in detail elsewhere.[39] Briefly, we use the Charmm19 molecular mechanics energy function[38] 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:

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

Here, E_{coul} is the usual Coulomb energy, ϵ is a dielectric constant, 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 (measured in kcal/mol/Å^2), which depends on the atom type, and α is an overall scaling factor for the surface term.

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,[45] implemented in the XPLOR program,[46] using a 1.4 Å probe radius. For reasons of efficiency, following Street and Mayo,[47] 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 $C_i = \sum_j A_{ij}$ 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, since the contact areas can overlap: a portion of atom i can be in contact with two atoms j and j' at a time. Street and Mayo showed, and we confirmed[39] that 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); for details, see.[39]

The interaction energy between each pair of sidechains, or between a sidechain and the backbone, involved a short energy minimization stage.[10] 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.

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.[39] In the previous work, we did extensive testing and comparison of several different sets of surface parameters, based on sidechain reconstruction, protein solvation energies, and mutations of over 1000 sidechains (including buried sidechains).[39] Thus, the atomic solvation coefficients and the surface calculations used here can be viewed as extensively optimized and tested.

Sequence Optimization

We used a heuristic optimization procedure developed by Wernisch et al.[10] One of the goals of this work is to continue to test the performance of this method. 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 typically perform ∼450,000 heuristic cycles for each 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).

Empirical Correction to the Unfolded State Energy

Given the simplicity of the effective energy and the unfolded state model, it was necessary to add an empirical correction to the unfolded state energies. For each amino acid I, a correction e_X is defined, where X represents the current amino acid type at position I. There are 17 values to optimize, corresponding to the 17 amino acid types that are free to mutate. We optimized the e_X so as to obtain reasonable overall amino acid compositions for the designed sequences of four test proteins: Grb2 (1gcqB), Vav (1gcqC), c-Src (1cka), and SHG (1shg), which are all SH3 domains (see Table 1). Homologous proteins were found by a BLAST search of the SWISSPROT database, using an identity threshold of at

Table 1. Test Set of Proteins.

PDB	Short name	Full name	SCOP family
1gcqB	Grb2	Growth factor receptor-Bound protein 2	SH3 domain
1gcqC	Vav	Vav proto-oncogene	SH3 domain
1cka	c-Crk		SH3 domain
1shg	Alpha-spectrin		SH3 domain
1abo	ABL kinase	ABL tyrosine kinase	SH3 domain
1ad5	HCK kinase	Haematopoietic cell kinase	SH3 domain
1csk	Csk	c-Src specific tyrosine kinase	SH3 domain
1fmk	c-Src	c-Src tyrosine kinase	SH3 domain
1lz1	Lysozyme		C-type lysozyme
4pti	BPTI	Pancreatic trypsin inhibitor	Kunitz-type inhibitors and BPTI-like toxins
1ctf	L7/L12	Ribosomal protein L7/L12	ribosomal protein L7/L12, C-terminal domain
1enh	Homeodomain	Engrailed homeodomain	homeodomain
1pgb	Protein G	Protein G, Ig binding domain	immunoglobulin binding domains
1zaa	Zif268		classic zinc finger, C2H2
1c9o	CSP	Cold shock protein	cold shock DNA-binding domain-like
1bdd	Protein A	Protein A, B-domain	Ig binding protein A modules

least 35% to the native, query protein, for a chain length no less than 90% of the native length. The mean amino acid frequencies $f_X^{\exp}$ were computed by averaging over this data set. We then proceeded iteratively, with the e_X initially set to zero. At each iteration, 30,000 sequences were computed for each protein (through 30,000 heuristic cycles). The corresponding amino acid frequencies, f_X^{calc}, averaged over all sequences, proteins, and amino acid positions, were compared to the experimental frequencies $f_X^{\exp}$. The energy correction e_X was then modified according to the Boltzmann-like relation:

$$e_X^{\text{new}} = e_X^{\text{old}} + 0.5 \ln \frac{f_X^{\exp}}{f_X^{\text{calc}}}. \quad (2)$$

With this scheme, if a given type X is too abundant in the designed sequences, eq. (2) leads to an increased stability of the unfolded state when X is present, so that X will be less abundant in the next round. After eight rounds, the frequencies converged to the experimental values; the corrections e_X no longer changed significantly, and the procedure was stopped; the final values are given in Table 2.

Table 2. Empirical Corrections to Unfolded Energy (kcal/mol).

Residue type X	Original contribution	Optimized contribution	Empirical correction e_X
Ala	−2.80	−1.50	−1.30
Asp	−22.17	−15.82	−6.35
Asn	−12.85	−7.81	−5.04
Arg	−20.89	−23.35	2.46
Glu	−22.74	−17.75	−4.99
Gln	−12.02	−8.90	−3.12
His	−13.22	−11.40	−1.82
Ile	−5.63	−5.19	−0.44
Leu	−5.55	−4.63	−0.92
Lys	−16.69	−18.81	2.12
Met	−5.61	−6.99	1.38
Phe	−7.05	−7.16	0.11
Ser	−9.09	−4.14	−4.95
Tyr	−9.77	−9.94	0.17
Thr	−8.64	−4.17	−4.47
Trp	−7.60	−11.37	3.77
Val	−5.50	−2.77	−2.73

Software Implementation

As pointed out by Mayo et al.,[48] the pairwise energy function and discrete conformational space imply that all the relevant energy data can be precomputed and stored in an energy matrix.[10] 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,[46] 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 employed our Proteins @ Home distributed computing platform, which allows us to use the computers of several thousand volunteers in over 70 countries (see the list of participants at biology.polytechnique.fr/proteinsathome). Proteins @ Home is based on the Berkeley Open Infrastructure for Network Computing, BOINC.[49] The Proteins @ Home platform and project will be described in detail elsewhere.[41]

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 Wernisch et al.[10] 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). The CPU requirements of the different steps are given in Table 3.

Table 3. CPU Timing on a 2.6 GHz Intel Xeon Processor.

Protein	Grb2 (1gcqB)
Length	57 amino acids
Number of interacting residue pairs	453
Total CPU time for energy calculation	203 h
CPU time per residue pair	27 min
Number of computed sequences	100,000
CPU time for sequence calculation	27 h
CPU time for 3D structure reconstruction	1 min structure

Natural Sequence Sets

For comparison to the designed sequences, we collected natural sequences from SWISSPROT. For each of the eight SH3 proteins, we retrieved homologues with at least 60% identity, giving about 12 homologous sequences per SH3 protein. A larger set including more distant homologues was then formed by the union of these smaller sets, for a total of 94 natural SH3 sequences. For the other eight test proteins, we proceeded similarly. For each protein (say, lysozyme), we retrieved homologues with at least 60% identity. We did the same for three other proteins of the same SCOP family (C-type lysozyme family, in this example), giving four small sequence sets. We then formed a large sequence set by grouping the small sets. For lysozyme, the large set has 199 homologous sequences.

Molecular Dynamics Simulations

We performed MD simulations using the AMBER force field[50] and a generalized Born (GB) solvent model. Specifically, we used the GB/HCT variant,[43,51] originally proposed by Hawkins et al.[52] The GB parameters were optimized previously for computational sidechain placement and protein mutagenesis.[39] To equilibrate the structures, 50,000 MD steps were done using a timestep of 0.1 fs and the Verlet algorithm. Next, the preequilibrated proteins were gradually heated from 50 to 300 K. After the heating, 250,000 steps of equilibration were done. We then increased the timestep to 0.5 fs and equilibrated the proteins for another 400,000 steps. During the equilibration, we rescaled the atomic velocities every 250 steps. Finally, we ran MD with no velocity scaling for 2 ns.

Results

Comparing the Designed and Native Sequences; Comparison to Earlier Work

The design approach was tested on 16 globular proteins, listed in Table 1 and shown in Figure 1. The test set includes α-helical proteins (protein A, homedomain), β-proteins (lysozyme), and mixed, α, β proteins. Four SH3 proteins (uppermost in Table 1) were used as a "learning set" to optimize the empirical energy corrections, e_X (see Methods). The optimized values (see Table 2) were then used throughout this study.

Figure 2 shows two typical ΔG_{fold} spectra, corresponding to ~450,000 sequences computed for Grb2 and c-Crk. The spectra are very roughly Gaussian, with an energy range of ~60 kcal/mol. The location of the best BLOSUM62 scores (using the native sequence as a target) are indicated. The best BLOSUM scores do not correspond to the lowest energy sequences, but are distributed over the energy spectrum, especially the region of maximal sequence occurrence (the peak of the spectrum). This is not surprising, since natural sequences are usually not optimized by evolution to maximize their stability.[54] Rather, one can often increase the stability of a protein by exchanging hydrophilic core residues for hydrophobic ones.[55,56]

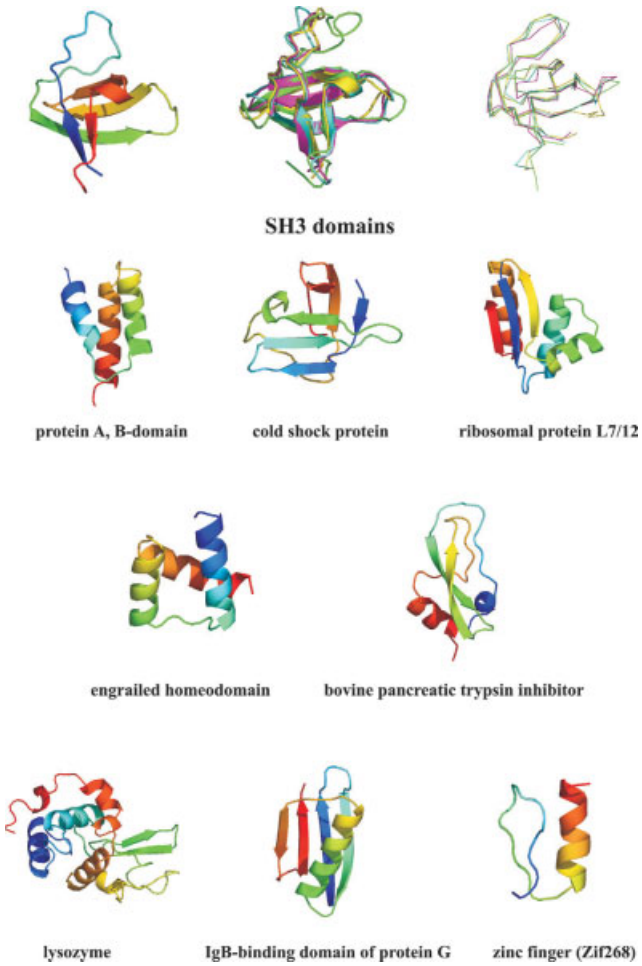

Figure 1. The test set of proteins. The eight SH3 domains (upper line) include Grb2 (left), four proteins (including Grb2) used as a learning set for the unfolded energy correction (center), and four other SH3 domains (right; shown as a structural alignment that also includes Grb2). The other eight proteins are below. Figure produced with Pymol.[53] [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

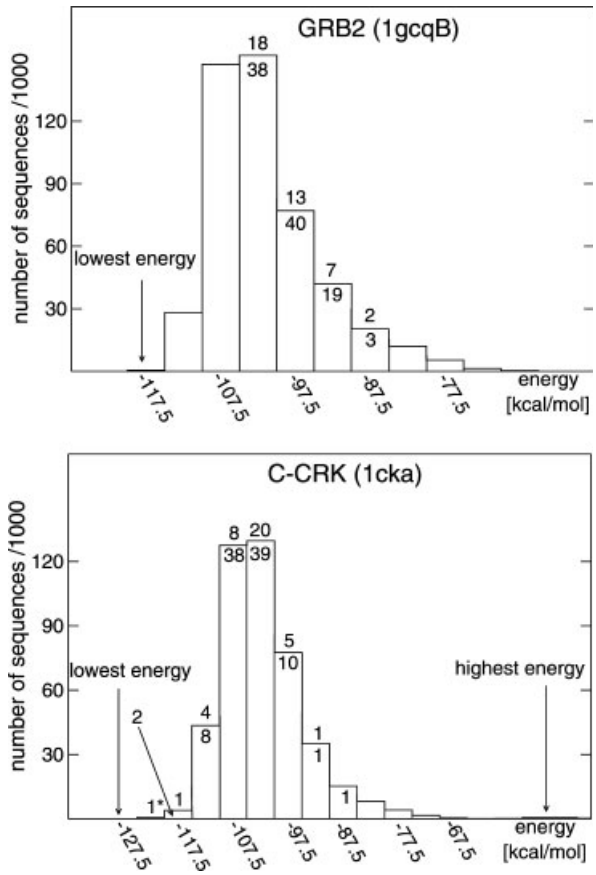

Figure 2. Histograms of Grb2 and c-Crk folding energies (ΔG_{fold}) for the computed sequences. Numbers above (within) bins indicate the number of sequences that are in the top 40 (top 100) scoring sequences (using the Blosum62 similarity matrix, with the native sequence as the target).

Table 4 gives the identity scores for the 16 proteins. "Reduced" identities are also given; they are computed with the nonmutating residues excluded (Cys, Gly, and Pro), leading to lower identities. If we consider the complete set of sequences for the eight SH3 proteins (~450,000 each), the highest average identity is for Vav (1gcqC), with 42.5% and the lowest is for the HCK kinase, with 23.5%. The four proteins of the "learning set" give an average identity of 34.5%; the remaining SH3 domains give 30.7%. The other eight proteins give an average identity of 30.4%. The averages over the low energy sequences (sequences with the highest folding free energy) are slightly higher: 35.0% for the SH3 domains and 31.8% for the other eight proteins (Table 4). If we consider only the best scoring sequences within each data set, we obviously

Table 4. Identity Scores of Computed Sequences.

PDB code	Protein name	Length (reduced)	Full identity			Reduced identity		
			Mean[a]	Low energy[b]	Best[c]	Mean	Low energy	Best
1gcqB	Grb2	57 (46)	37.3	37.2	49.5	22.4	22.1	37.5
1gcqC	Vav	69 (52)	42.5	45.6	55.0	23.7	27.8	40.2
1cka	c-Crk	56 (48)	34.5	39.8	52.1	23.6	29.8	44.2
1shg	Alpha-spectrin	57 (53)	23.6	26.9	37.6	17.8	21.4	32.9
1abo	ABL kinase	58 (49)	35.9	37.7	48.5	24.1	26.3	39.1
1ad5	HCK kinase	58 (54)	23.5	22.6	38.4	17.8	16.9	33.8
1csk	Csk	56 (46)	37.1	37.3	48.5	23.5	23.6	37.4
1fmk	c-Src	60 (54)	26.3	32.4	43.5	18.2	24.9	37.2
1lz1	Lysozyme	130 (109)	33.2	33.8	43.5	20.3	21.0	32.6
4pti	BPTI	58 (42)	38.5	36.5	50.7	15.1	12.3	31.9
1ctf	L7/12	68 (61)	32.5	31.7	44.3	24.7	23.9	41.5
1enh	Homeodomain	54 (52)	18.2	18.9	32.1	15.0	15.8	29.5
1pgb	Protein G	56 (52)	29.2	31.3	43.4	23.7	26.0	39.0
1zaa	Zif268	28 (24)	37.3	43.6	51.7	26.8	34.2	43.6
1c9o	Cold shock	66 (55)	32.4	33.0	46.1	18.8	19.6	35.4
1bdd	Protein A	53 (49)	22.1	25.8	35.0	15.7	19.7	29.6
	Average		33.2	33.4	45.0	20.7	22.8	36.6

Identity (%) with respect to the native sequence. Reduced identity takes into account only residues allowed to mutate (all except Cys, Gly, and Pro).

[a] Average over all 450,000 computed sequences.

[b] Average over the 100 lowest energy sequences.

[c] Average over the 40 highest-identity sequences.

obtain much higher identity scores: 46% for the eight SH3 proteins and 43.4% for the other eight proteins. If the native sequence is known ahead of time (the most common situation by far), these very high-scoring sequences can be identified and used. Collections of designed sequences obtained for each protein are available as Supplementary Material.

The above identity rates compare favorably to several previous studies of whole protein redesign. Larson et al. recently applied an expensive, flexible backbone approach to 253 different protein families.[24] Average identity scores between 26% and 33% were obtained, several points below our typical scores. For the Kunitz_BPTI family, an average identity of 26% was given[24]; we obtain a mean value of 38.5% for 4pti (Table 4).

In 2000, Baker and coworkers obtained an average sequence identity between computed and experimental sequences of 27% for a test set of 108 proteins.[9] To obtain this result, they optimized their energy function so that the lowest energy sequences gave the highest identity scores. Further optimization of the energy function, more elaborate rotamer libraries and the use of a flexible protein backbone enabled them to improve the designed sequences considerably, so that in 2003 they obtained an average identity of 35% for nine redesigned globular proteins.[33] In 2005, they reported an average sequence identity of 37% for the redesign of 42 small globular proteins.[15] This last result is about 4% higher than our low-energy sequences, but a bit lower than our best sequences. Their method is designed to yield a small number of highly optimized sequences. In contrast, future applications to fold recognition will require sequence sets that are both realistic and sufficiently diverse.[15,22–24]

In a more restrictive design procedure, Koehl and Levitt mutated residues by exchanging existing pairs, thus constraining the overall amino acid composition to the native one.[18] Their designed sequences of 1ctf and 1tim had average identities of 36 and 16%, respectively. Here, we report an average identity of 31.7% for our lowest-energy sequences (and 44.3% for the best-scoring sequences). In 1997, Dahiyat et al. described the first completely redesigned protein, the zinc finger Zif268.[48] Their best sequence had an identity of 36% with the native protein. Here, our highest identity score is 51.7%.

In 2005, Pokala and Handel carried out full protein design for seven proteins.[25] They reported identity values for two somewhat different approaches. The best approach differs by an optimization of the van der Waals energy term and by the incorporation of a negative design criterion, which restricts the amino acid composition at the protein surface. This approach gave identity scores between 33.5 and 46.7%. The only protein common to our study is protein G. Their best method gave an average identity of 41.1%, while their other method gave 31.1%. It is not clear how much of the difference is due to the van der Waals energy and how much is due to the surface composition restraint. Our method, which does not restrain surface composition, gave 31.3% for the low energy sequences. The good results of Pokala and Handel with either method are also due to their more detailed, all-atom force field (OPLS-AA) and their expensive, generalized Born solvation model.

The method most similar to ours is that of Wodak and coworkers, applied in 2002 to seven different protein folds.[36] Four of them are represented in our data set: SH3 domains, the homeobox, protein G, and the cold shock protein. Despite their using

a more complex, all-atom force field and a more complex rotamer library,[57] our results are significantly improved. Thus, for 11 SH3 domains, they reported an average identity score of 23.9%, compared to 34.9% obtained here for eight SH3 domains. We average over our 40 lowest-energy sequences, which is comparable to the averaging method employed in.[36] We also have notably improved scores for the cold shock protein, the homeodomain, and protein G (Table 4).

The Computed Sequences are Similar to Experimental Sequences

To compare the designed sequences to natural sequences in more detail, we created two sets of natural SH3 sequences. First, for each of our eight SH3 proteins, we retrieved from SWISSPROT sequences that share at least 60% identity with the query sequence. The second set of natural sequences includes 94 entries and is simply the union of the smaller SH3 sets collected above. Figure 3 compares the identity and BLOSUM scores obtained with the natural and computed sequences; see also Table 5 and Supplementary Material. We applied the following weighting scheme to the sequences. The variability of each amino acid within the larger set of natural sequences was computed. If the frequency of the prevailing residue exceeded 80%, a weighting factor of 1 was assigned. A frequency between 50 and 80% led to a weight of 0.75, and a frequency of less than 50% led to a weight of 0.5. In most cases, the computed scores (Fig. 3) overlap with the range of the large set of natural sequences. For four of the SH3 domains (1gcqC,1cka, 1abo, and 1csk), the weighted average BLOSUM62 score of the computed sequences actually exceeds or is very close to the corresponding value of the natural set. Thus, in most cases, the computed sequences behave like moderately-distant homologues of the native

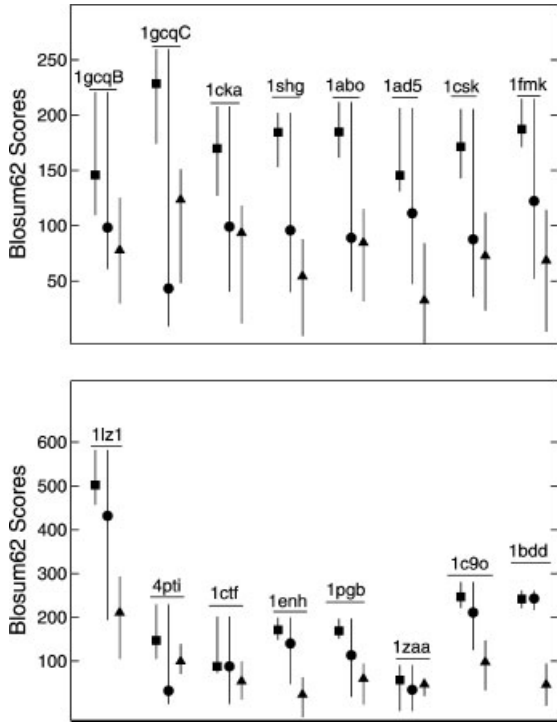

Figure 3. Comparison of native and computed BLOSUM62 scores. Proteins are identified by their PDB code. For each protein, three vertical lines represent the range of scores within the small natural set (left), the large natural set (middle), and the computed set (right). The average score in each set is also shown (as a square, a dot, a triangle).

Table 5. Comparison of BLOSUM62 Scores Between Natural and Computed Sequences.

PDB code	Unweighted scores				Weighted scores			
	Natural	Total[a]	Low energy[b]	Best[c]	Natural	Total	Low energy	Best
1gcqB	114.1	112.8	102.1	161.1	98.2	84.8	77.9	119.4
1gcqC	29.7	149.8	163.2	204.5	43.2	111.8	123.5	145.5
1cka	113.0	105.1	129.5	159.4	99.1	77.6	93.6	111.8
1shg	111.6	51.1	64.4	101.2	95.9	41.3	54.3	79.3
1abo	104.0	103.5	109.8	145.4	89.0	80.4	84.7	106.2
1ad5	134.0	45.0	33.7	89.5	111.3	39.8	32.5	69.3
1csk	98.0	99.8	99.5	147.5	87.7	70.6	72.7	105.2
1fmk	151.3	61.9	82.5	127.2	122.2	52.6	68.6	101.2
1lz1	487.9	236.6	244.8	324.3	431.1	205.2	211.0	278.2
4pti	37.0	138.8	131.1	181.2	31.6	104.1	99.4	131.1
1ctf	71.5	91.5	86.0	142.3	65.1	58.4	53.9	92.35
1enh	156.6	25.3	28.7	58.6	139.6	19.0	23.7	44.62
1pgb	154.1	61.6	73.9	115.8	113.5	50.4	59.8	86.24
1zaa	24.8	61.14	73.0	82.0	22.7	39.9	47.1	52.1
1c9o	232.1	97.1	98.6	152.3	160.2	93.7	96.7	139.7
1bdd	246.2	46.5	45.5	81.9	238.5	47.3	46.4	81.2

BLOSUM62 scores with respect to the native sequence. Weighting scheme is described in the main text.
[a] Average over all 450,000 computed sequences.
[b] Average over the 100 lowest energy sequences.
[c] Average over the 100 highest-identity sequences.

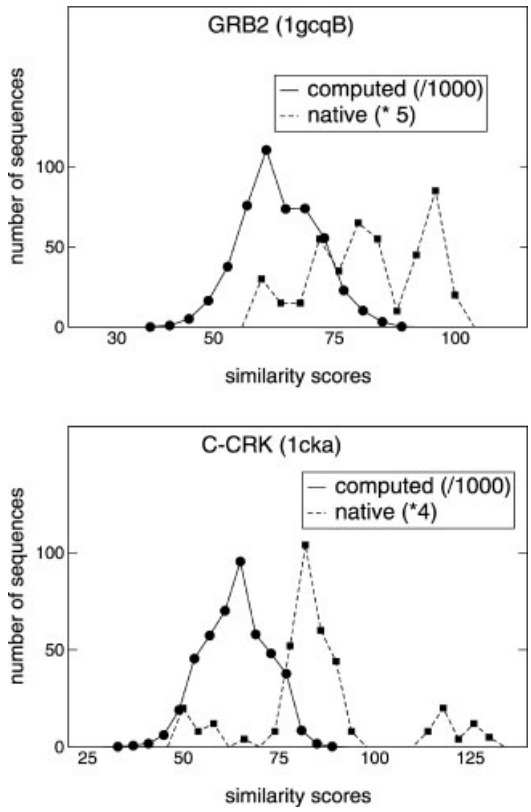

Figure 4. Experimental and computed similarity scores for two SH3 proteins: Grb2 and c-Crk. Dashed: Blosum62 scores for 94 natural SH3 domains, compared to the native sequence. Black: scores for the computed sequences.

sequence. The same trend is mostly seen for the other eight proteins. Only for 1bdd are the computed sequences distinctly below the natural set.

Figure 4 further illustrates the designed sequences for the four SH3 proteins of our learning set. The computed sequences are compared to a sequence profile obtained from the large set of 94 natural SH3 sequences. Similarity scores were computed using the formula:[36]

$$s = \sum_i \sum_a f_{ia} S(x_i, a). \tag{3}$$

Here, x_i is the amino acid type at position i in the native sequence; a is one of the 20 amino acid types, f_{ia} is the frequency (between 0 and 1) of amino acid type a at position i in the set of natural sequences; $S(x_i, a)$ is the BLOSUM62 similarity score; the first sum is over the native sequence; the second sum is over the amino acid types. The values of s are given on the horizontal axis of Figure 4 and the number of scores are given on the vertical axis. The similarity scores of the designed sequences fall within the range of the native scores in all four cases. The similarity scores of the designed sequences span the range 30–90, whereas the native scores span the range 50–125. In three of the four cases (Grb2, c-Crk, and alpha-spectrin), the designed sequences overlap with the lower part of the spectrum of natural sequences. For Vav, the designed sequences overlap with the upper part of the spectrum. Overall, the similarity scores are consistent with results presented in Figure 3. As with the BLOSUM scores, we find that the similarity scores of the designed sequences overlap with the scores of natural sequences of other SH3 proteins.

Finally, BLAST searches of the SWISSPROT databank were done using the best-scoring BLOSUM62 sequences of the eight SH3 proteins. In all eight cases, a series of native SH3 proteins were retrieved, which are shown in Figures 5 and 6 for Grb2 and c-crk, respectively. Encouragingly, the native sequence was retrieved in all eight cases. No high-scoring false-positives were obtained; e.g., all the sequences with an E-value below 0.001 were indeed SH3 domains. Conversely, when an unrelated protein of a comparable size, BPTI, is used as a query, no SH3 sequences are retrieved among the top 1000 sequences returned by BLAST; only sequences from the BPTI_Kunitz SCOP family are among the top 1000 sequences. The same is true if a high-scoring designed BPTI sequence is used as a query.

Analysis of Grb2 and c-Crk Sequences

Figure 5 illustrates the designed sequences for Grb2. Three blocks of sequences are shown: the 25 best-scoring designed sequences (top); 25 homologues retrieved from SWISSPROT with the native sequence as a query (middle); 25 homologues retrieved from SWISSPROT with the best-scoring designed sequence as a query (bottom). The 3D structures shown above the sequences depict the original protein and the best BLOSUM62-scoring sequence. Sequences are color-coded as follows: blue, DENQ; red, HKR; yellow, FWY; pink, ST; orange, ILMV; green, ACGP.

A long loop, typical of SH3 domains, extends from Ala5 to Gly22. In all the natural sequences, the L_6F_7 motif at the beginning of the loop is either maintained or mutated to one of the following, homologous motifs: I_6Y_7, I_6Y_7, or V_6Y_7. L_6 and F_7 are surface residues, whose sidechains point towards solvent. In the designed sequences more hydrophilic residues are found: K_6A_7 or K_6N_7. The next stretch of residues, predominantly DF(D/E), is closely mimicked by the designed motif, EFN, which occurs in nine out of 25 computed sequences. For the other 16 sequences, we find EFA or KFE. The remaining part of the loop shows a similar level of conservation. At positions 16 and 17, the conserved motif $(E/D)_{16}$ L_{17} is usually changed to $Y_{16}L_{17}$ and rarely to $A_{16}L_{17}$ in the designed sequences. Inspection of the native sequences of Vav (1gcqC) shows an aromatic residue in the same position as in the designed sequences. This aromatic residue interacts with other hydrophobic residues at the end of β strands 3 and 4. Therefore, one can assume that this mutation will not affect the overall stability of the protein.

The core of the Grb2 structure is a five-stranded beta sheet. In this region, designed and natural sequences agree well. The first strand includes either $Y_2V_3Q_4$, $K_2V_3Q_4$, or $K_2V_3R_4$. The designed sequences start with an ionic amino acid followed by either V_2 or an aromatic residue $(Y/F)_3$, then either an ionized residue (E/K) or a hydrophobic one (W/V). In one case, the natural $K_2V_3Q_4$ motif is reproduced. The second strand goes from G_{22} to D_{29}. The natural $(D/E)_{23}$ is usually maintained. The weakly conserved native $F_{24}I_{25}$ is mostly changed to $H_{24}I_{25}$. The solvent-exposed H_{26} is replaced in the designed sequences by a charged E_{26} or K_{26}. Finally, the

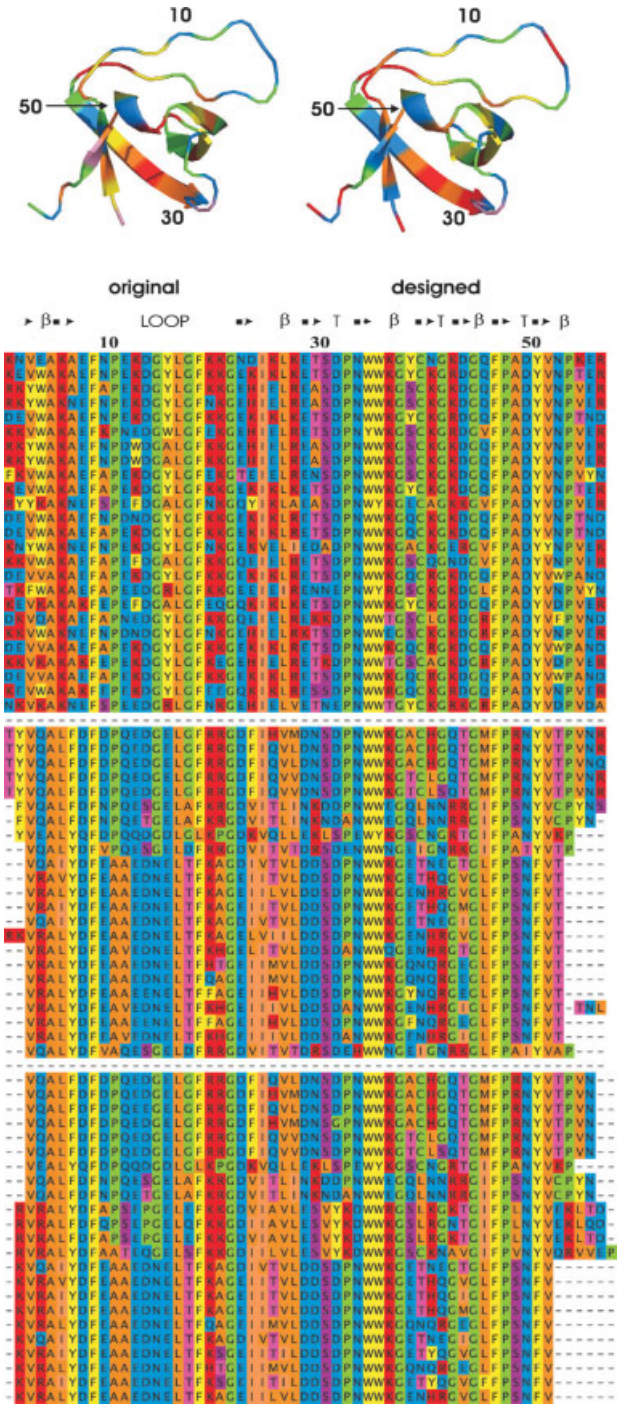

Figure 5. Grb2 sequences. Upper group: 25 high-scoring computed sequences. Middle group: 23 natural sequences obtained from SWISSPROT, with the native sequence as query. Bottom group: 23 natural sequences from SWISSPROT using a computed sequence as query. The colors distinguish six amino acid groups: {DENQ}, {HKR}, {FYW}, {ACGP}, {ST}, and {ILMV}. Secondary structure and residue numbers are shown, along with the 3D structure (colored according to the native and a designed sequence).

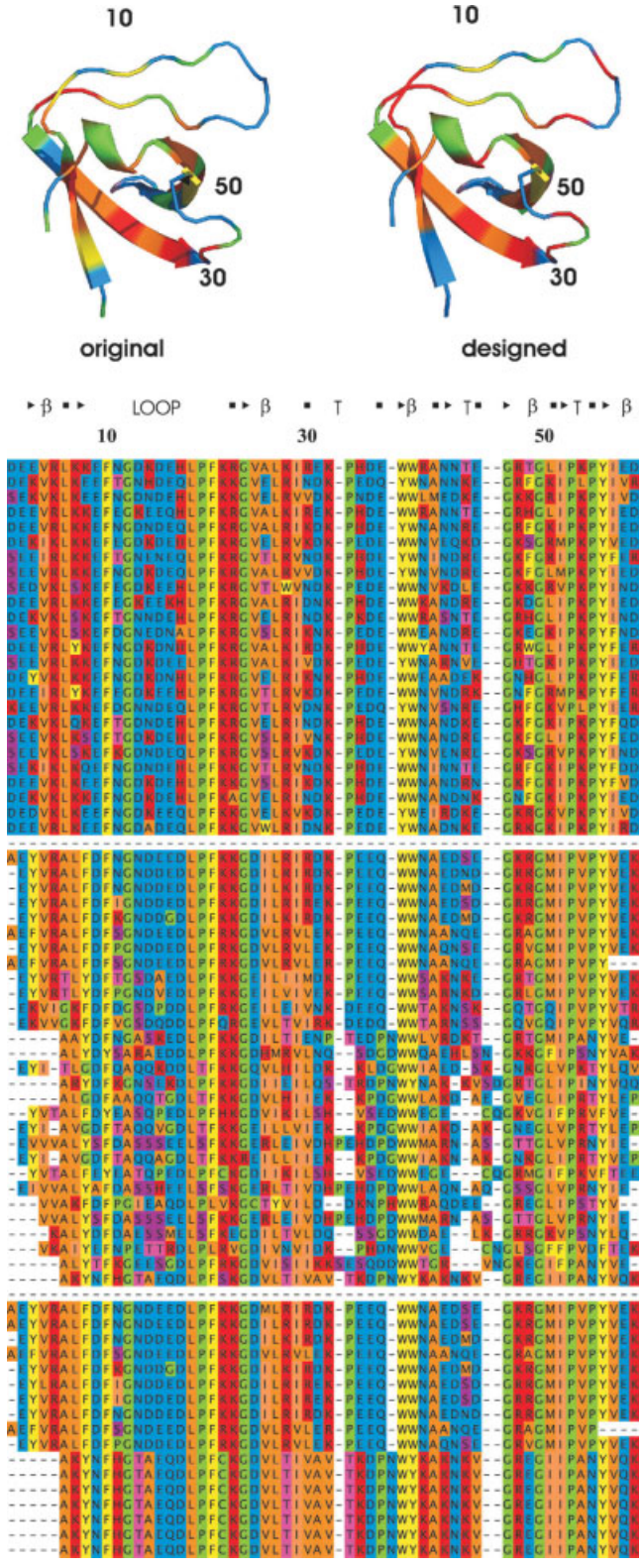

Figure 6. 1CKA sequences. Same representation as in Figure 5.

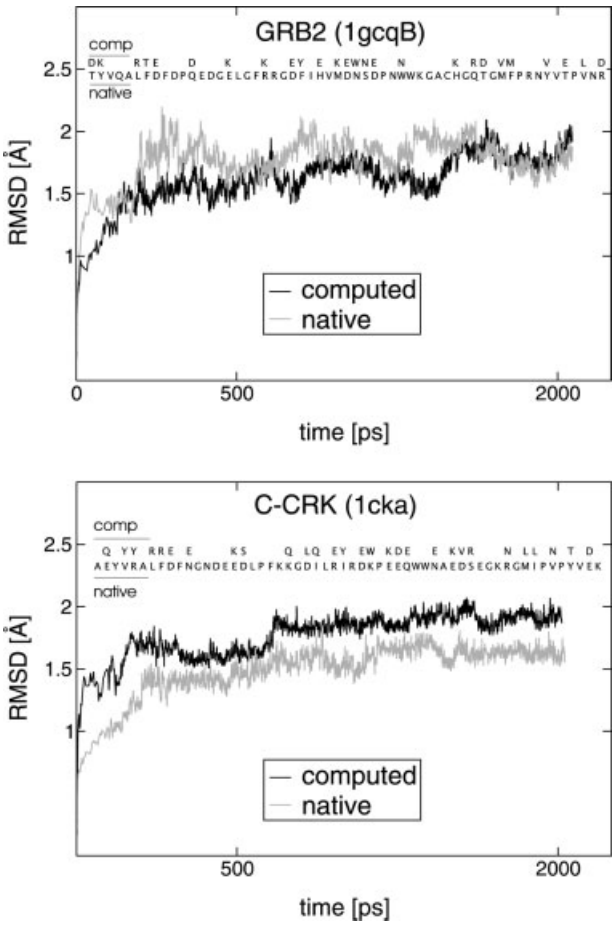

Figure 7. Molecular dynamics simulations of two native and designed proteins. RMS deviation (Å) from the initial, experimental, native structure. For each protein, the native protein and one designed variant were simulated. The native and computed sequences are shown as insets (only mutated positions in the computed sequence are shown). The AMBER force field and a generalized Born solvent were used.

native, hydrophobic-hydrophobic-ionic motif is usually changed to a hydrophobic-ionic-ionic pattern. The loop that connects strands two and three is accurately reproduced. At the beginning of the third strand, the native $W_{35}W_{36}K_{37}$ pattern is maintained. After the third strand, the prevalence of $K_{43}D_{44}$ in the designed sequences is due to a new salt bridge. Finally, the rest of the Grb2 sequence is fairly well-reproduced by the design.

C-Crk (Fig. 6) has a longer first beta strand than Grb2. The predominant native motif, $E_2Y_3V_4R_5$, is reproduced in three out of four residues ($E_2V_4R(/K)_5$), whereas in postion three we find mostly E or D. In one case, the complete native sequence is recapitulated. The following, long loop (A_2 to G_{23}) shows a strong overlap between designed and native sequences, but the L_7F_8 pattern at the beginning is not reproduced. The subsequent beta strand, from G_{23} to D_{30}, is a mixture of hydrophilic and hydrophobic amino acids in both sets of sequences. The native $W_{38}W_{39}$ motif is conserved in the designed sequences.

The Stability of the Designed Proteins is Supported by Molecular Dynamics

To further test the stability of the designed sequences we performed molecular dynamics simulations (MD) for the four proteins of the "learning set:" Grb2, Vav, c-Crk and spectrin. In each case, we considered both the native protein and a high-scoring sequence. For reasons of efficiency, we used an implicit solvent model, as for the design calculations. However, both the force field and the solvent model used here were of a higher quality than the ones used for the design calculations. Indeed, we used the all-atom, AMBER force field,[50] instead of the "polar hydrogen," Charmm19 force field[38] used above. Instead of the CASA solvent model, we used a recent, high-quality, generalized Born model (GB), which is known to yield good quality protein structures in MD simulations.[39,43,58] MD simulations were run for 2 ns in each case.

The deviations of the Grb2 and c-Crk structures from the experimental, crystal structures are shown in Figure 7 as a function of time. Data for Vav and spectrin are similar. For the four native sequences, the MD structures agree very well with the crystal structures, with rms deviations of about 1.5 or 2 Å after 2 ns. This is comparable to other proteins studied with GB,[58,59] and with explicit solvent treatments. For the designed sequences, the quality of the MD structures is comparable (Grb2, Vav) or very slightly worse (c-Crk, spectrin), with rms deviations of about 1.7–2.2 Å after 2 ns of MD. These low values suggest that the designed sequences are stable, but prefer a slightly different backbone conformation.

Conclusions

The overall design procedure implemented here combines several well-established ingredients: a molecular mechanics energy function, implicit solvent, a fixed backbone, and sidechain rotamers.[9,31,36,48] Sequences are selected based on their folding free energy, using a tripeptide model of the unfolded state. At a more detailed level, compared to previous studies of whole protein redesign, there are significant differences in one or more of the ingredients in each case. In general, our procedure uses the simplest ingredients: a "polar hydrogen" force field, a surface area solvent model, a fixed protein backbone, a heuristic method[10] for exploring sequence and rotamer space. Nevertheless, our results are comparable to other recent studies. The good results obtained here can be attributed to our earlier, complete reparameterization of the CASA solvent model[39] and a careful optimization of other model parameters. In particular, an empirical correction to the unfolded state energy was parameterized.

Thus, the present study extends our knowledge of computational protein design and its robustness or sensitivity to model details. We also tested the use of large-scale volunteer computing for this application, with our BOINC-based, Proteins @ Home platform. For eight SH3 domains and a heterogeneous set of eight other proteins, we obtained designed sequences with a native-like character. In all cases tested, the best designed sequences allowed us to retrieve the native sequence from the SWISSPROT database (with no false positives), suggesting that the method has the potential to become a fold recognition method.

The method has several obvious limitations, some of which are shared by competing implementations. By selecting for low folding energies, we could over-optimize stability, compared to natural evolution. By focussing on sequences with a high similarity to native, we could miss sequences that have a low sequence homology but a high structural homology. Nevertheless, the overall performance of the method appears encouraging. Many additional questions remain open. The amount of diversity within the computed sequences is promising but needs more investigation, for example, as well as their sensitivity to the fixed backbone assumption. In the future, we plan to implement a new, residue-pairwise, generalized Born solvent,[60] and to apply the method to problems of fold recognition.[22]

Acknowledgments

The authors thank the many volunteers who have participated in the Proteins at Home project and contributed computer cycles used in this work. We thank the BOINC development community for testing the alpha version of the Proteins at Home platform. They thank Christine Bathelt, Alexey Aleksandrov, and Najette Amara for discussions.

References

1. Drexler, K. Proc Natl Acad Sci USA 1981, 78, 5275.
2. Eisenberg, D. Nature 1982, 295, 99.
3. Pabo, C. Nature 1983, 301, 200.
4. Ponder, J.; Richards, F. M. J Mol Biol 1988, 193, 775.
5. Hellinga, H.; Richards, F. Proc Natl Acad Sci USA 1994, 91, 5803.
6. Dahiyat, B.; Mayo, S. Prot Sci 1996, 5, 895.
7. Harbury, P. B.; Plecs, J. J.; Tidor, B.; Alber, T.; Kim, P. S. Science 1998, 1998, 1462.
8. Desjarlais, J.; Handel, T. J Mol Biol 1999, 289, 305.
9. Kuhlman, B.; Baker, D. Proc Natl Acad Sci USA 2000, 97, 10383.
10. Wernisch, L.; Héry, S.; Wodak, S. J Mol Biol 2000, 301, 713.
11. Kuhlman, B.; Dantas, G.; Ireton, G.; Varani, G.; Stoddard, B.; Baker, D. Science 2003, 302, 1364.
12. Dwyer, M.; Looger, L.; Hellinga, H. Science 2004, 304, 1967.
13. Havranek, J.; Harbury, P. Nat Struct Biol 2003, 10, 45.
14. Ventura, S.; Serrano, L. Proteins 2004, 56, 1.
15. Saunders, C.; Baker, D. J Mol Biol 2005, 346, 631.
16. Wollacott, A. M.; Zanghellini, A.; Murphy, P.; Baker, D. Prot Sci 2007, 16, 165.
17. Koehl, P.; Levitt, M. J Mol Biol 1999, 293, 1161.
18. Koehl, P.; Levitt, M. J Mol Biol 1999, 293, 1183.
19. Koehl, P.; Levitt, M. Proc Natl Acad Sci USA 1999, 96, 12524.
20. Pokala, N.; Handel, T. Prot Sci 2004, 13, 925.
21. Chowdry, A. B.; Reynolds, K. A.; Hanes, M. S.; Voorhies, M.; Pokala, N.; Handel, T. J Comput Chem 2007, 28, 2378.
22. Larson, S.; Garg, A.; Desjarlais, J.; Pande, V. Proteins 2003, 51, 390.
23. Larson, S.; Pande, V. J Mol Biol 2003, 332, 275.
24. Larson, S.; England, J. E.; Desjarlais, J.; Pande, V. Prot Sci 2002, 11, 2804.
25. Pokala, N.; Handel, T. M. J Mol Biol 2005, 347, 203.
26. Raha, K.; Wollacott, A. M.; Italia, M. J.; Desjarlais, J. R. Prot Sci 2000, 9, 1106.
27. Cochran, F. V.; Wu, S. P.; Wang, W.; Nanda, V.; Saven, J. G.; Therien, M. J.; DeGrado, W. F. J Am Chem Soc 2005, 127, 1346.
28. Swift, J.; Wehbi, W. A.; Kelly, B. D.; Stowell, X. F.; Saven, J. G.; Dmochowski, I. J. J Am Chem Soc 2006, 128, 6611.
29. Kang, S. G.; Saven, J. G. Curr Opin Chem Biol 2007, 11, 329.
30. Baker, D. Phil Trans R Soc Lond 2006, 361, 459.
31. Butterfoss, G.; Kuhlman, B. Ann Rev Biophys Biomolec Struct 2006, 35, 49.
32. Guérois, R.; Lopez de la Paz, M. Eds. Protein Design: Methods And Applications; Humana Press: New Jersey 2007.
33. Dantas, G.; Kuhlman, B.; Callender, D.; Wong, M.; Baker, D. J Mol Biol 2003, 332, 449.
34. Koehl, P.; Levitt, M. Proc Natl Acad Sci USA 2002, 99, 1280.
35. Koehl, P.; Levitt, M. Proc Natl Acad Sci USA 2002, 99, 691.
36. Jaramillo, A.; Wernisch, L.; Héry, S.; Wodak, S. Proc Natl Acad Sci USA 2002, 99, 13554.
37. Tuffery, P.; Etchebest, C.; Hazout, S.; Lavery, R. J Biomol Struct Dyn 1991, 8, 1267.
38. Brooks, B.; Bruccoleri, R.; Olafson, B.; States, D.; Swaminathan, S.; Karplus, M. J Comp Chem 1983, 4, 187.
39. Lopes, A.; Aleksandrov, A.; Bathelt, C.; Archontis, G.; Simonson, T. Proteins 2007, 67, 853.
40. Liang, S.; Grishin, N. Proteins 2004, 54, 271.
41. Simonson, T.; Mignon, D.; Schmidt am Busch, M.; Lopes, A.; Bathelt, C. In Distributed and Grid Computing—Science Made Transparent for Everyone. Principles, Applications and Supporting Communities; Tektum Publishers: Berlin, 2007.
42. Andreeva, A.; Howorth, D.; Brenner, S. E.; Hubbard, J. J.; Chothia, C.; Murzin, A. G. Nucl Acids Res 2004, 32, D226.
43. Onufriev, A.; Bashford, D.; Case, D. J Phys Chem B 2000, 104, 3712.
44. Fraternali, F.; van Gunsteren, W. J Mol Biol 1996, 256, 939.
45. Lee, B.; Richards, F. J Mol Biol 1971, 55, 379.
46. Brünger, A. T. X-plor version 3.1, A System for X-Ray Crystallography and NMR; Yale University Press: New Haven, 1992.
47. Street, A.; Mayo, S. Fold Desig 1998, 3, 253.
48. Dahiyat, B. I.; Gordon, D. B.; Mayo, S. L. Prot Sci 1997, 6, 1333.
49. Anderson, D. P. Boinc: A System for Public-Resource Computing and Storage. In 5th IEEE/ACM International Workshop on Grid Computing; IEEE Computer Society Press; USA; 2004.
50. Cornell, W.; Cieplak, P.; Bayly, C.; Gould, I.; Merz, K.; Ferguson, D.; Spellmeyer, D.; Fox, T.; Caldwell, J.; Kollman, P. J Am Chem Soc 1995, 117, 5179.
51. Moulinier, L.; Case, D.; Simonson, T. Acta Cryst D 2003, 59, 2094.
52. Hawkins, G.; Cramer, C.; Truhlar, D. Chem Phys Lett 1995, 246, 122.
53. DeLano, W. L. The PyMOL Molecular Graphics System; DeLano Scientific: San Carlos, CA, USA, 2002.
54. DePristo, M. A.; Weinreich, D. M.; Hartl, D. L. Nature Rev Genet 2005, 6, 678.
55. Stites, W.; Gittis, A.; Lattman, E.; Shortle, D. J Mol Biol 1991, 221, 7.
56. Shortle, D. Curr Opin Struct Biol 1993, 3, 66.
57. Dunbrack, R.; Karplus, M. J Mol Biol 1993, 230, 543.
58. Simonson, T.; Carlsson, J.; Case, D. A. J Am Chem Soc 2004, 126, 4167.
59. Feig, M.; Brooks, C. L., III. Curr Opin Struct Biol 2004, 14, 217.
60. Archontis, G.; Simonson, T. J Phys Chem B 2005, 109, 22667.