Folding free energy function selects native-like protein sequences in the core but not on the surface

Alfonso Jaramillo*, Lorenz Wernisch†‡, Stéphanie Héry§, and Shoshana J. Wodak*¶

*Unité de Conformation de Macromolécules Biologiques, CP160/16, Université Libre de Bruxelles, 50 Avenue F. D. Roosevelt, 1050 Brussels, Belgium; †European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom; ‡School of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, United Kingdom; and §Université Reims Champagne-Ardenne, Laboratoire de Spectroscopies et Structures, Biomoléculaires (LSSBM), Institut Federatif de Recherche 53 Biomolécules, Unité de Formation et de Recherche Sciences–Moulin de la Housse, BP 1039-51687 Reims Cedex 2, France

Edited by Martin Karplus, Harvard University, Cambridge, MA, and approved July 26, 2002 (received for review February 6, 2002)

An automatic protein design procedure is used to select amino acid sequences that optimize the folding free energy function for a given protein. The only information used in designing the sequences is a set of known backbone structures for each protein, a rotamer library, and a well established classical empirical force field, which relies on basic physical chemical principles that underlie molecular interactions and protein stability, and has not been adjusted to yield native-like sequences. Applying the procedure to 7 different known protein folds, representing a total of 45 different native protein structures, yields ensembles of designed sequences displaying remarkable similarity to their natural counterparts in the protein core, but which are distinctly non-native on the protein surface. We show that natural and designed sequences for a given fold score significantly higher than random sequences against profiles derived from both, designed and natural sequence ensembles. Furthermore, we find that designed sequence profiles can be used to retrieve the native sequences for many of the analyzed proteins using standard PSI-BLAST searches in sequence databases. These findings may have important implications for our understanding the selection pressures operating on natural protein sequences and hold promise for improving fold recognition.

Protein sequences are shaped by a complex interplay of different selective pressures, which are poorly understood. The selection pressure for performing the proper function is probably overriding. But in addition, there presumably is selection for maintaining the stability of the 3D structure, for folding or processing efficiency, as well as random drift resulting from neutral mutations. Improving our ability to distinguish between sequence features resulting from these different selection pressures should have a major impact on our ability to predict protein structure and function from sequence (1, 2).

One way of addressing these questions could be by using experimental (3, 4) or computational (5–10) protein design approaches to search sequence space for sequences that satisfy the stability constraints for known protein structures, and comparing these sequences to their naturally occurring counterparts.

Significant progress has been achieved recently in computational methods for protein design (5, 6, 8, 11–13). Notable examples are the *de novo* design of a sequence adopting the zinc finger fold (8) and of a novel α-helix bundle protein with a right-handed superhelical twist (14).

Applying these methods to redesigning the sequences of small proteins was shown to yield sequences displaying significant similarity to the native sequences (5, 6, 11), with overall identity scores relative to the native sequences of 25–30% (10, 15). Kuhlman and Baker (10), who obtained such results for a set of 108 small proteins, deduced that the volume of sequence space optimal for a protein structure is surprisingly restricted to a region around the native sequence. They concluded that stability requirements have played a significant role in shaping natural protein sequences.

Recently we described a new automatic procedure, DESIGNER, for selecting amino acid sequences compatible with a given protein 3D structure (12). Selected sequences minimize a fitness function akin to the free energy of folding, which relies on basic physical chemical principles that underlie molecular interactions and protein stability. This function combines the all-atom force field of CHARMM (16) with a simple empirical surface area-dependent hydration term (17). Unlike all previous studies mentioned above, the parameters of this fitness function have not been adjusted to yield native-like sequences, and no constraints are imposed on the amino acid composition of the designed sequences.

DESIGNER therefore seems to be particularly well suited for investigating, on a more rigorous basis, the relationships between the designed sequences and their natural counterparts, and thereby gaining insight into the factors that shape natural sequences. To this end we apply it to redesign the sequences of seven protein domains comprising the SH3 domain, homeobox domain, DNA binding helix-turn-helix domain, B1 domain of streptococcal protein G, the Ci2 inhibitor, the cold-hock, and antifreeze proteins. To improve exploration of sequence space, backbone flexibility is taken into account by performing the calculations on several backbone templates from different crystal structures of each domain. The procedure generates the family of low free energy sequences for each backbone and these families are combined to yield the global ensemble of designed sequences for the domain. Using these ensembles we examine the relation between the predicted amino acid sequences and the backbone conformations available for a given fold. In addition, we evaluate the similarity between the ensembles of designed and natural sequences. Position-dependent frequencies of the designed and naturally occurring sequences are computed and scored against one another, both for the entire polypeptide, and for core positions only. Lastly, designed sequence profiles are used to test the ability for recognition of native sequences in public sequence databases, using standard sequence alignment procedures.

Materials and Methods

Computing Sequences Compatible with a Given Protein Backbone. To select the amino acid sequences that are compatible with a given protein backbone structure, we use the procedure implemented in the software DESIGNER (12). This software has two main components. A function that measures the fitness of a given sequence for the structure at hand, and the optimization procedure, which selects high-scoring sequences from a very large number of possibilities.

The fitness function is computed as the difference between the free energies of the protein native folded state and a reference state used as a model for the unfolded protein, as described in ref. 12 and *Supporting Methods*, which is published as supporting information on the PNAS web site, www.pnas.org.

To select amino acid sequences with lowest free energies, a simple heuristic procedure is used (12). This procedure yields

This paper was submitted directly (Track II) to the PNAS office.

¶To whom correspondence should be addressed. E-mail: shosh@ucmb.ulb.ac.be.

Table 1. Sequence identities of minimum energy sequences selected by DESIGNER relative to their native counterparts, for seven small protein domains

Domain	Backbone templates	% Identity to *wt*		
		All	Core	Surface
SH3	11	23.9 ± 4.23	54.8 ± 11.73	11.0 ± 7.94
Homeobox	9	15.8 ± 2.95	37.0 ± 10.37	10.2 ± 6.65
HTH	6	24.6 ± 6.69	54.2 ± 18.70	11.6 ± 7.77
Protein G	2	23.6 ± 5.45	48.9 ± 23.86	11.8 ± 0.00
CI2	7	25.2 ± 2.68	60.5 ± 12.70	18.9 ± 5.84
Cold-shock	4	29.2 ± 1.54	71.6 ± 6.27	16.9 ± 1.90
Antifreeze	6	19.8 ± 2.96	43.1 ± 7.93	1.4 ± 3.11

The protein domain is listed in the leftmost column. The number of backbone templates used for each domain is given in the second column. The third through fifth columns list the average identity scores (%) and standard deviations of the designed minimum energy sequences relative to the corresponding native sequence, considering all residues, only core residues, and only surface residues, respectively. Average *P* values corresponding to these scores were between $\approx 10^{-9}$–10^{-3} for the full sequences (all), between $\approx 10^{-13}$ and 10^{-5} for the core residues, and $\approx 10^{-1}$ for the surface residues. The *P* values were computed as described in the legend of Table 3, which is published as supporting information on the PNAS web site. The PDB ID codes of the backbone templates used in the calculations are as follows: SH3, 1cka, 1shg, 1pwt, 1sem, 1shf, 1qcf, 1ckb, 2src, 1bk2, 1abo, and 1fmk; homeobox, 1enh, 1fil, 9ant, 1mnm, and 1au7; HTH domain, 1r69, 2cro, 1lmb, 1lli, 1b0n, and 1 per; protein G B1 domain, 1pgb and 1idg; Ci2, 2ci2, 1cse, 1ypc, 2tec, 2acb, and 1coa; cold-shock, 1csp, 1mjc, 1c9o, and 1csq; antifreeze, 1ops, 1msi, 2jja, 1ame, 1b7i, and 1ekl.

solutions close to the global minimum when the number of iterations is sufficiently large (typically ≈350,000; see supporting information for details). For a full design of a 50-residue polypeptide and typically about 100 rotamer/amino acids at each position, each iteration screens 25,000 possible sequences, leading to a total of $\approx 10^{10}$ sequences screened in the entire procedure. Selected sequences are those with energies between 2 and 6 kcal/mol from the minimum energy sequence.

Combining Designed Sequences for Multiple Backbone Templates. The full design procedure is applied to our set of seven small single-domain proteins. For each domain a number of backbone templates is selected from high-resolution structures deposited in the PDB (refs. 18 and 19; see the legend of Table 1 for a full list). The sequence design procedure is applied to each backbone template individually, yielding its designed sequence family. The families from all of the templates of a given domain are combined on the basis of structural alignments (20), yielding a multiple alignment for the designed sequence family of that domain. Alignments are listed in www.ucmb.ulb.ac.be/~alfonso/supplement/.

Treatment of Natural Sequences and Sequence Analysis. For each domain, the family of natural sequences was obtained from multiple alignments available in PFAM (21). These alignments were pruned of sequences with unusually large or numerous insertions or deletions to ensure that the examined sequences share the same fold and are of similar lengths. Structural alignments were used to improve the sequence alignment in regions. The number of natural sequences used for each domain is given in the legend of Fig. 2. Alignments are available at www.ucmb.ulb.ac.be/~alfonso/supplement/.

To quantify the similarity between the designed and natural sequence families, position-dependent amino acid frequencies (22) were computed from the designed and natural sequence families, respectively. To score a given sequence against a profile, we use the standard score $s = \Sigma_i \Sigma_y f_{iy} S(x_i, y)$, were f_{iy} is the frequency (on a scale of 0–1) of amino acid y in the profile at position i of the sequence, x_i is the amino acid at position i of the target sequence, and $S(x_i, y)$ is the BLOSSUM62 matrix (23).

Results

Properties of Minimum Energy Sequences of a Protein Domain. To illustrate the properties of the minimum energy sequences computed by our sequence design procedure, we first discuss results obtained for the SH3 domain, the protein in our set with the largest number of available high-resolution backbone templates and an appreciable number of known natural sequences.

Thirteen backbones from evolutionarily and structurally related SH3 domains were used in the calculations. Three additional backbones were derived by quenching conformations after, respectively, 20 ps, 40 ps, and 150 ps (1 ps = 10^{-12} s) of a high-temperature (600 K) molecular dynamics simulation performed on the proto-oncogen product c-Crk SH3 domain (PDB ID code 1cka) in presence of explicit solvent. Details about the various natural and simulated backbones are given in Table 2, which is published as supporting information on the PNAS web site.

The different backbones display root mean square deviations (rmsd) ranging between 1 and 2 Å, except for the unrelated HIV integrase DNA binding domain where the rmsd are larger (3–4 Å on average). For each of the considered backbones our procedure selected from among all possible sequences those with energies of 2–6 kcal/mol above the minimum energy (see supporting information). The number of sequences selected on this basis ranged between 77 and 1,000, depending on the backbone. Given the uncertainty associated with the energy function used in the selection procedure, all these sequences were considered to be compatible with the considered fold (12).

The top-ranking lowest energy sequences computed for individual SH3 backbones display on average 23.9 ± 4.2% identity to their native counterpart. A significantly higher identity level of 54.8 ± 11.7% is displayed when only buried positions are scored. These positions are defined as residues exposing <10% of their solvent-accessible surface area to solvent; they number between 10 and 15 in SH3. All these identity scores are highly significant, with *P* values in the range of 10^{-16}–10^{-4} (see Tables 2 and 3).

A direct consequence of these observations is that the identity scores of the designed sequences are particularly low for surface residues. These are defined here as residues exposing more than 50% of their solvent-accessible surface area to solvent. Their identity scores are 11 ± 7.9%, with some individual designed sequences displaying zero identity to the native sequence.

We find, furthermore, that the amino acid composition of surface residues is markedly different from native. It tends to be enriched in Arg and Gln amino acids. Detailed analysis of a few structures computed with the minimum energy sequences for the C-crk SH3 (PDB ID code 1cka) and cold-shock (1c9o) proteins shows that the designed Gln and Arg side chains form stabilizing H bonds with surrounding polar and charge residues, or otherwise stick into the surrounding solvent. The number of H bonds in these structures is on average somewhat higher (≈9) than in the corresponding native proteins (≈6).

Sequences designed using backbones derived from the high-temperature c-Crk SH3 molecular dynamics trajectory featured somewhat different properties. Interestingly, although they had an appreciable level of identity to the native sequence (29–32%), their core residues were less well conserved (28–57%) than in the other designs. We were able to verify that the limits of secondary structures in these backbones (24) differed from those in the native crystal structure, suggesting that the corresponding structures have undergone some local unfolding. This interpretation is supported by the observation that the identity scores of the core residues in the MD conformations decrease as the simulation time at elevated temperature, and hence the degree of unfolding, increases (see Table 2).

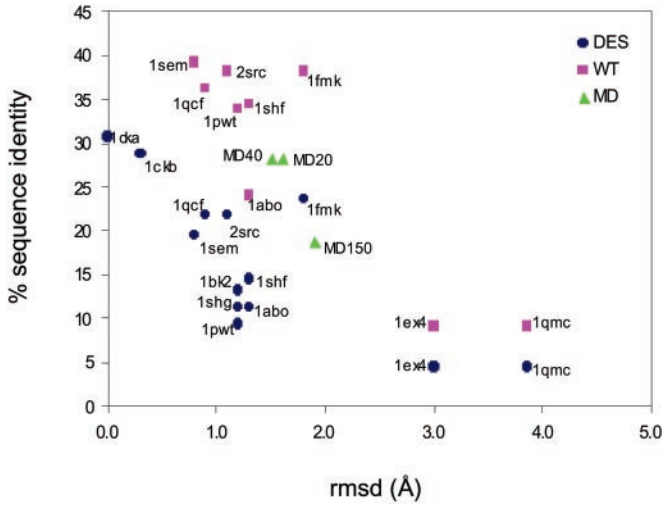

Fig. 1. Sequence identity of native and designed sequences as a function of the root mean square deviation (rmsd) of backbone atoms of SH3 domains. Sequence identity levels (%; vertical axis) and the rmsd (Å) of backbone atoms (horizontal axis) are measured relative to those of the c-Crk SH3 domain (1cka). Points represent the various sequence/template combinations analyzed. They are denoted by the PDB ID code of the corresponding template, with designed sequences (blue circles), native sequences (purple squares), and designed sequences using backbone templates from MD simulations (green triangles; see text). These templates are denoted 1ckaMD20, 1ckaMD40, and 1ckaMD150, corresponding to conformations from high-temperature MD simulations of respectively, 20 ps, 40 ps, and 150 ps (1 ps = 10^{-12} s). Structural alignments were performed using CE software (39), and the identity scores were computed over the entire sequence, without corrections for the alignment lengths, as those are essentially constant across the examined SH3 proteins.

Influence of the Backbone Conformation on the Designed Sequences. To further characterize the influence of the backbone conformation on the selected sequences, it is helpful to examine the relation between sequence and backbone similarities, and to compare it with previous analyses on natural sequences (25).

Fig. 1 displays the percent identity of the minimum energy sequences relative to the sequence of c-Crk (1cka) computed over the entire sequence after structural alignment, versus the rmsd of the corresponding templates relative to the 1cka backbone. We see clearly that the sequence identity level decreases as the rmsd of the template relative to the reference backbone increases. As expected, the lowest sequence similarity and largest rmsd, relative to 1cka, is displayed by the sequence of the unrelated HIV integrase DNA binding domain (1qmc).

Interestingly, the minimum energy sequence designed for a particular template resembles more the native sequence of that template than the native c-Crk sequence, or the native sequence of any of the other SH3 domains. It is noteworthy that lowest energy sequences designed using backbones derived from the molecular dynamics trajectory of 1cka display ≈5–10% higher sequence identity to 1cka than those using native backbones of other SH3 domains with equivalent rmsd.

Thus, information on the native sequence appears to be in some way "encoded" in its backbone. This can be explained by considering that proteins with somewhat different sequences may be viewed as stabilizing slightly different conformations of the same fold. These conformations would correspond to distinct local minima of the energy landscape for that fold (26, 27). Conversely, somewhat different polypeptide backbones would "stabilize" different amino acid sequences, each defining a distinct local minimum in sequence space. This, in turn, points out that the sequence space accessible to a given 3D structure cannot be adequately explored by computational procedures such as DESIGNER, unless the conformational variability of the polypeptide backbone is taken into account.

Minimum Energy Sequences for a Set of Small Globular Proteins. Table 1 summarizes the identity scores relative to native for the designed sequences of lowest energy computed for all of the seven domains of different folds and secondary structures considered here. The listed scores represent averages and standard deviations of the percent identity of the minimum energy sequence relative to its native counterpart, computed for the entire polypeptide and taken over the all of the structural templates of each domain.

These scores are between 15.8 and 29.2%, and hence in the same range as those obtained for the SH3 domain. The homeobox and antifreeze proteins display the lowest scores (15.6–19.8%; Table 1), corresponding to the highest P values (-10^{-5}–10^{-3}). The most prominent differences are once more displayed between the core and surface positions. The designed core amino acids are on average 52.8% identical to their native counterparts, whereas surface residues display significantly lower identity levels of 11.7%.

To further investigate the differences between the designed and natural sequences, we extend our analysis to the ensemble of known natural sequences associated with each of the seven protein domains. This ensemble comprises between 113 and 1,225 sequences for four of the domains (cold-shock, HTH, SH3, and homeobox), and between 20 and 35 sequences for the three remaining domains (Ci2, Protein G, and antifreeze).

Among the interesting questions to address is how the diversity of the designed sequence ensembles compares with those of the natural sequences, and to what extent the diversity of the two types of ensembles differs in the core versus surface regions. A very rough estimate can be obtained by comparing the average identity scores computed, respectively, between the designed sequences for each domain and between their natural counterparts. When the full polypeptide is considered, the average identity scores of the designed sequences for the seven domains range between 33.0 and 57.8%, with a rather uniform standard deviations of ≈10%. For the natural sequences, the same quantity spans a wider range of 22.3–73.5%, with standard deviations of 7–18%. These differences are probably due to biases resulting from the very small number of natural sequences available for some of the domains and do not represent real differences in sequence diversity.

Per domain, the diversity of the designed and natural sequences is also comparable, but clearer differences appear when core or surface regions are considered (see Tables 4 and 5, which are published as supporting information on the PNAS web site). As expected, all identity scores are generally much higher in the core than in surface regions or than for the full polypeptide. But the average identity scores of designed sequences are systematically lower than those of the natural ones in the core, whereas the opposite is true in surface regions. Thus, the designed sequences appear to be more diverse in the core and less diverse in surface positions in comparison with the natural sequences, at least in the set of protein domains examined here. But this result clearly needs further confirmation by a proper statistical analysis (28) of a larger protein sample (F. Sirota-Leite, A.J., and S.J.W., unpublished work).

Scoring Designed Sequences Against Their Natural Counterparts. To that end, we compute for each domain the positional frequency matrix (profile) (22, 29) from the multiple alignment of its natural sequences. The natural and designed sequences for each fold are then individually scored against the natural sequence profile of that fold, using a standard scoring function (see *Materials and Methods* and Fig. 2 legend). In addition, these scores are compared with those of completely random sequences, and of random H/P sequences. The latter are random sequences required to have hydrophobic amino acids in buried positions and polar ones on the surface (see *Materials and Methods*).

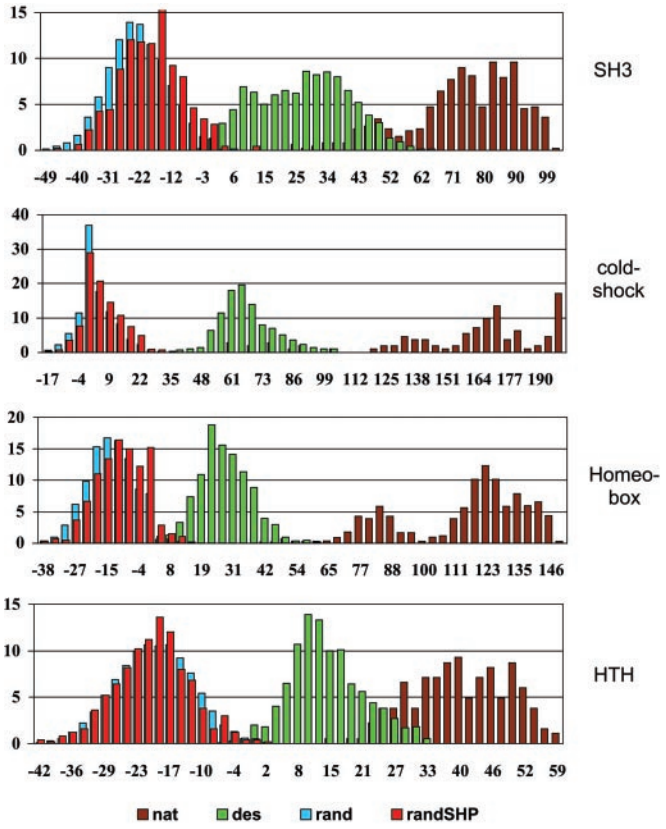

Fig. 2. Assessing the similarity between designed and natural sequences of four small protein domains. Displayed are the histograms of the similarity scores s, computed using the following formula: $s = \Sigma_i \Sigma_y f_{iy} S(x_i, y)$, where f_{iy} is the frequency (on the scale of 0–1) of amino acid y at position i in the natural sequences of each protein, x_i is the amino acid at position i in the target sequence, and $S(x_i, y)$ is the BLOSSUM62 matrix (23). The values of s are given on the horizontal axis; the number of scored sequences is given on the vertical axis. The names of the different protein domains are listed on the right. Brown, natural sequences; green, designed sequences; blue, random sequences; red, random H/P sequences. Random H/P sequences are random sequences subject to the restriction that buried positions harbor hydrophobic amino acids and solvent exposed positions polar amino acids. The natural sequences, from which position specific amino acid frequencies were derived, were taken from PFAM (21) after pruning (see text). The number of sequences in each profile was: 534/2700 for SH3, 1225/1377 for the homeobox proteins, 184/887 for HTH, and 113/324 for the cold-shock proteins. The smaller of the two numbers is that of the natural sequences and the larger is of the designed sequences. The number of natural sequences available for the remaining three domains analyzed in this study was too small to permit a similar analysis (see text).

Fig. 2, displays the score distributions obtained for the SH3, HTH, and homeobox domains, and the cold-shock protein, the four domains with the largest set of available natural sequences. Very similar albeit noisier results were obtained for the three remaining domains, for which many fewer native sequences are available (data not shown).

A first important observation is that the designed sequences score significantly better than completely random sequences, and better than random H/P sequences. For all of the domains, the score distributions of the random and random H/P sequence ensembles overlap nearly perfectly on the low end of the scale of Fig. 2, and are well separated not only from the scores of natural sequences, but also from those of the designed ones. This behavior indicates that our calculations introduce constraints that go well beyond the requirement of simply burying nonpolar amino acids and exposing polar amino acids.

A second key observation is that the scores of the designed sequences span a rather wide range. In general, these scores are lower than those of the natural sequences, but for several domains (SH3 domain, HTH, and cold-shock in Fig. 2), some overlap between the two distributions is nonetheless observed.

Scoring Sequences Against Profiles of Core Positions. In view of the results presented above, it seemed reasonable to assume that the differences between score distributions of designed and natural sequences (Fig. 2) might be significantly reduced if only core positions were considered.

To check this hypothesis we now analyze how individual sequences from different ensembles (designed, natural, random, and random HP sequences, respectively) score against core profiles. The latter are the position specific amino acid frequencies of buried positions only. They tend to be noisier than profiles of the complete polypeptide, because the number of scored positions is low, ranging from 6 for the homeobox, to 14 for the SH3 domain. We therefore computed them from the ensembles of designed sequences for each domain, because those are larger than their natural counterparts (comprising between 300 and 2,700 sequences, compared with 113–1,225 for the natural sequences), and should hence suffer less from noise problems. The analysis could therefore be carried out for all seven domains, including the three domains for which the number of natural sequence was particularly low.

The results, shown in Fig. 3, lead to several very interesting observations. For most domains, the scores of random and random HP sequences are well separated from those of the designed sequences, as in the profiles of the entire protein (Fig. 2), but the separation between the scores of random and natural sequences is narrower. Most strikingly still, we see that the overlap between the scores of the designed and native sequences, computed here against the core profile of designed sequences, is significantly better than for the scores computed for the full protein (Fig. 2). With the possible exception of the HTH domain, a surprisingly good overlap is observed for all of the domains, including protein G and Ci2, for which only a limited number of natural sequences is available.

This, together with the observations on the similarities between the designed and natural core profiles, confirms that the resemblance between the designed and natural sequence ensembles is clearly much greater in core positions than for the sequence over all.

Are Designed Sequence Profiles Effective in Recognizing the Native Sequence? Having shown that that there is a significant overlap between the designed and natural sequence ensembles, we tested whether the profiles derived from the ensembles of designed sequences, computed for the different templates of the seven protein domains, can be used to retrieve the corresponding native sequences from SWISS-PROT (30), using one iteration of the PSI-BLAST sequence alignment procedures (31) with standard settings.

Results obtained show that the designed sequence profiles have a quite good native recognition performance. Of the 45 tested profiles, 33 (72%) were able to retrieve the native sequence of the corresponding structural template with acceptable to good significance (see Table 5). For 18 of these profiles, the native sequence was retrieved with E values between 2.10^{-8} and 0.1. For a further 10 profiles, it was retrieved with E values between 0.1 and 5, and for the remaining 5, the native sequences had E values in the range of 5–9.4. We could, furthermore, verify that in the majority of the cases, unrelated sequences scored much more poorly. It should be mentioned, however, that the results obtained for proteins such as the antifreeze and Ci2, which contain as many as six prolines might be biased by the fact that these residues were not redesigned here, although there is good indication that many of these prolines tend to reappear when designed (unpublished results).

These rather encouraging results can therefore be taken to indicate that sequences computed solely on the basis of structural constraints and physical chemical principles more often than not contain the necessary information to enable native sequence rec-

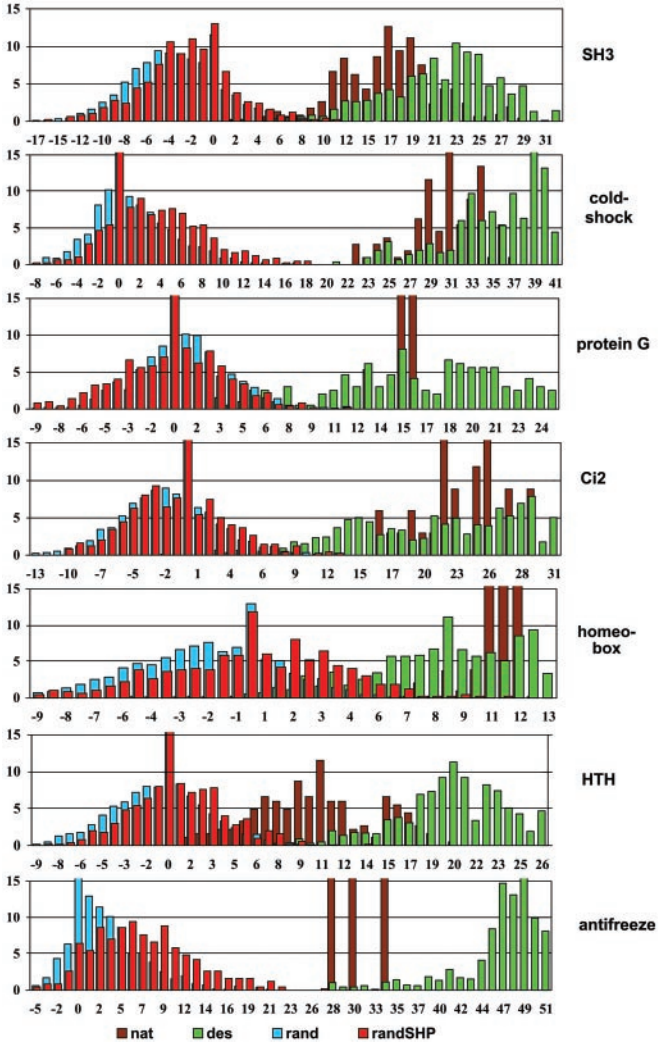

Fig. 3. Similarity between designed and natural sequences in core residues of seven small protein domains. Similarity scores were computed as in Fig. 2, but using the positions specific amino acid frequency matrices computed from the low energy sequences selected by DESIGNER, instead of those of natural sequences. In addition, only core residues (burying ≥90% of their solvent-accessible surface area) were scored. The names of the protein domains are indicated on the right hand side. Horizontal axis, score values; vertical axis, number of scored sequences. Green, scores of designed sequences; brown, scores of natural sequences; blue, scores of random sequences; red, scores of random H/P sequences. (See legend of Fig. 2 for the meaning of H/P sequences.) For protein G B1 domain, Ci2, and the antifreeze protein, the number of sequences used in the analysis are 20/300, 35/889, and 23/723, respectively. The first of the two numbers is that of the natural sequences; the second is of the designed sequences.

ognition from amongst a very large number of possibilities. The fact that this is obtained with standard sequence alignment procedures and default parameters suggests that performance could probably improve by using more sophisticated profile-based HMM methods (32). But if fold recognition is the main goal, profiles derived from designed sequences computed while constraining the amino acid composition to be native-like, as done by other authors (5, 6), should be even more effective than those designed here, because such sequences display on average a 5–10% higher identity levels to their native counterpart (see Tables 6 and 7, which are published as supporting information on the PNAS web site).

Discussion

The main finding of this study is that there is significant overlap between the ensembles of low free energy and natural sequences for a set of small proteins. This overlap, furthermore, enables a good level of native sequence recognition using the low free energy sequence profiles and standard sequence alignment methods without any attempt to optimize recognition performance. Moreover, evidence is provided that this is essentially due to a remarkable resemblance between the designed and natural sequences in the protein core, whereas the resemblance between the two sequence ensembles is particularly poor on the protein surface.

Two recent studies in which protein sequences, designed using analogous procedures, were compared with their natural counterparts (10, 15), also reported a higher similarity between the designed and natural sequences in the core versus the surface regions. However, just as all other design procedures to date, they too were in one way or another fine-tuned to yield natural sequences. In particular, the scoring function of Kuhlman and Baker (10) was "trained" to produce native-like sequences for a set of reference proteins using as many as 26 adjustable parameters.

In contrast, our folding free energy comprises a widely used classical molecular mechanics force field and an implicit hydration term, previously optimized to yield experimental values for the vacuum to water dissolution free energies of amino acids. It contains only three adjustable parameters (see supporting information for details), whose values were derived from physical chemical considerations, and no constraints were imposed on the amino acid composition, as done elsewhere (5, 6). Another key difference is that we sampled sequence space more than 3 orders of magnitude more extensively than in Kuhlman and Baker (10). We run ≈350,000 iterations of our heuristic algorithm and use several different backbones for each protein, whereas they use very short Monte Carlo runs, equivalent in total to only ≈50 iterations of our algorithm, clearly not enough for adequate sampling of sequence space (12).

The results presented here are therefore particularly significant because they were obtained without any "memory" of the expected characteristics of the natural sequences. If we believe that our folding free energy adequately embodies the basic physical chemical principles that underlie protein stability, then our findings lead to the important conclusion that stability requirements represent a significant evolutionary selection pressure on the amino acid sequence of core residues, but probably not on that of surface residues.

What are the selection pressures operating on surface residues? It is of some significance that all of the proteins used in our calculations have known interaction partners *in vivo*. The SH3 domains engage in interactions with cognate peptides. The homeobox and HTH domains interact with DNA. The cold-shock proteins interact with DNA and RNA, the protein G B1-domain and Ci2 protein bind, respectively, to immunoglobulins and chymotrypsin, and the antifreeze protein binds to ice. Most of these proteins also form dimers, trimers, or higher-order complexes. The natural sequences at surface positions may thus have been selected, at least in part, for mediating these different interactions, probably at the expense of protein stability.

The fact that designed sequences are very different from their natural counterparts on the protein surface may reflect just that. Indeed, the designed sequences have an increased proportion of Arg and Gln side chains, whereas the ratio of polar to nonpolar amino acids remains native-like (data not shown). The total number of H bonds is also increased by two to three per protein, but a fraction of the additional (positive) charges remain unpaired and pointing into the solvent.

One interpretation of these findings might be that our free energy function is inadequate, and unable to handle the delicate balance between electrostatics and solvent interactions in surface regions. To address this issue, we tested the CHARMM implementation of a recent continuum solvent model (ACE) of Schaeffer and colleagues (33, 34). This model is believed to approximate very well (up to

94.6% for BPTI; ref. 33) the much more costly Poisson Boltzman (PB) calculations (35), increasingly regarded as a standard for evaluating free energies in protein conformational searches and protein–ligand interactions (36).

Using a modified free energy, which incorporates ACE, to simply re-rank the sequences computed by DESIGNER for several of the folds analyzed here did, however, not result in a significant change in the surface amino acid composition of the highest-ranking sequences. Low energy sequences were still significantly non-native (30% identity to the native sequence, on average) and had similar proportions of Arg and Gln residues in surface positions. Moreover, a typical result, obtained for the SH3 domain, was that with ACE, the native sequence ranked even lower on the energy scale (160.7 kcal/mol above the minimum) than with the simple surface area-dependent term (46.5 kcal/mol; see Table 8, which is published as supporting information on the PNAS web site).

One cannot rule out that other approximations, such as the crude model used for the unfolded or reference state, may be at fault. It is difficult to ignore, however, that features of the designed sequences—namely, the larger proportion of positively charged residues (Arg in particular) and the increased number of H bonds—exhibit interesting parallels with those of proteins from organisms growing at very high temperatures (80–97°C; hyperthermophiles).

The amino acid composition of proteins from 7 complete hyperthermophile genomes was recently compared with that of 22 mesophiles (37). This study showed that the hyperthermophiles display an excess of Lys, Arg, and Glu amino acids relative to their mesophilic counterparts, suggesting an increased formation of charge–charge interactions in their proteins. Interestingly, Gln and Asn residues, whose proportion also increases in our designed sequences, are much less frequent in hyperthermophiles. The latter finding, however, seems to be due to a mechanism that these organisms have evolved against the synthesis of these temperature-sensitive amino acids (37).

In another recent study, the cold-shock protein from a mesophile (*Bacillus subtilis*) was stabilized to a similar degree as a thermophilic variant from *Bacillus caldolyticus* by mutating only two Glu residues at positions 3 and 66 to Arg and Leu, respectively. The sequences designed here, using templates from the corresponding organisms, never contain Glu amino acids at the two positions, but often contain Gln and sometimes Arg (in position 66), together with numerous substitutions elsewhere.

Hence, a reasonable interpretation of our results may be that in natural sequences, amino acids in surface positions have not been optimized for protein stability, but selected primarily for functional reasons. This interpretation is very much in line with the recent proposal that amino acids directly involved in ligand recognition or catalysis can often be identified in a protein 3D structure as surface residues located in a particularly destabilizing environment (38).

This may imply that in general, evolution has compromised less on stability in favor of other requirements in core residues, whereas for surface residues the compromises that were made with regard to stability may be more substantial than has hitherto been realized.

This might explain at least in part why proteins tend to be so large in comparison with the size of their business portions. A large core in proteins, whose sequence would be selected for optimizing the folding free energy, would allow more flexibility in the functional adaptation process in surface positions. In particular, it would allow these positions to tolerate a few particularly destabilizing residue constellations that may be required for function, or a larger number of mildly destabilizing residue constellations.

Finally, our findings on the significant rate of native sequence retrieval, using the designed sequence profiles, bodes well for the application of these profiles to fold recognition. In addition, the conclusions reached on the significant influence of functional requirements on surface residues suggest that recognition of distantly related proteins with the same fold but different functions might be improved by down-weighting the influence of these residues in the scoring scheme.

Although, clearly, further analyses using both theoretical and experimental approaches are needed to verify our findings, we believe that that they offer new insights into the sequence–structure relations in proteins.

We are grateful to Jean Richelle and Ricardo Valente for assistance with the computer systems at the Service de Conformation de Macromolécules Biologiques in Brussels and at the European Bioinformatics Institute in Cambridge. This work was supported by European Commission EU Grant BIO4 CT97-2086 and the Action de Recherches Concertées de la Communauté Française de Belgique.

1. Baker, D. & Sali, A. (2001) *Science* **294,** 93–96.
2. Baxter, S. M. & Fetrow, J. S. (2001) *Curr. Opin. Drug Disc. Dev.* **4,** 291–295.
3. Sauer, R. T. & Harrison, S. C. (1996) *Curr. Opin. Struct. Biol.* **6,** 51–52.
4. Plaxco, K. W., Riddle, D. S., Grantcharova, V. & Baker, D. (1998) *Curr. Opin. Struct. Biol.* **8,** 80–85.
5. Koehl, P. & Levitt, M. (1999) *J. Mol. Biol.* **293,** 1161–1181.
6. Koehl, P. & Levitt, M. (1999) *J. Mol. Biol.* **293,** 1183–1193.
7. Ponder, J. W. & Richards, F. M. (1987) *J. Mol. Biol.* **193,** 775–791.
8. Dahiyat, B. I. & Mayo, S. L. (1997) *Science* **278,** 82–87.
9. Desjarlais, J. R. & Handel, T. M. (1995) *Protein Sci.* **4,** 2006–2018.
10. Kuhlman, B. & Baker, D. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 10383–10388.
11. Desjarlais, J. R. & Handel, T. M. (1999) *J. Mol. Biol.* **289,** 305–318.
12. Wernisch, L., Héry, S. & Wodak, S. J. (2000) *J. Mol. Biol.* **301,** 713–736.
13. Gordon, D. B. & Mayo, S. L. (1999) *Structure Fold. Des.* **7,** 1089–1098.
14. Harbury, P., Plecs, J., Tidor, B., Alber, T. & Kim, P. (1998) *Science* **282,** 1462–1467.
15. Raha, K., Wollacott, A. M., Italia, M. J. & Desjarlais, J. R. (2000) *Protein Sci.* **9,** 1106–1119.
16. MacKerell, J. A. D., Bashford, D., Bellott, M., Dunbrack, J. R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., *et al.* (1998) *J. Phys. Chem.* **102,** 3586–3616.
17. Ooi, T., Oobatake, M., Némethy, G. & Scheraga, H. A. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 3086–3090.
18. Bernstein, F., Koetzle, T., Williams, G., Meyer, E., Brice, M., Rodgers, J., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112,** 535–542.
19. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000) *Nucleic Acids Res.* **28,** 235–242.
20. Boutonnet, N. S., Rooman, M. J., Ochagavia, M. E., Richelle, J. & Wodak, S. J. (1995) *Protein Eng.* **8,** 647–662.
21. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. (2000) *Nucleic Acids Res.* **28,** 263–266.
22. Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 4355–4358.
23. Henikoff, S. & Henikoff, J. G. (2000) *Adv. Protein Chem.* **54,** 73–97.
24. Kabsch, W. & Sander, C. (1983) *Biopolymers* **22,** 2577–2637.
25. Chothia, C. & Lesk, A. (1986) *EMBO J.* **5,** 823–826.
26. Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. (1991) *Science* **254,** 1598–1603.
27. Shortle, D., Simons, K. T. & Baker, D. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 11158–11162.
28. Koehl, P. & Levitt, M. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 1280–1285.
29. Gribskov, M., Luthy, R. & Eisenberg, D. (1990) *Methods Enzymol.* **183,** 146–159.
30. Bairoch, A. & Apweiler, R. (1997) *Nucleic Acids Res.* **25,** 31–36.
31. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
32. Eddy, S. R. (1996) *Curr. Opin. Struct. Biol.* **6,** 361–365.
33. Schaefer, M. & Karplus, M. (1996) *J. Phys. Chem.* **100,** 1578–1599.
34. Schaefer, M., Bartels, C. & Karplus, M. (1998) *J. Mol. Biol.* **284,** 835–848.
35. Gilson, M. K. & Honig, B. (1988) *Proteins* **4,** 7–18.
36. Elcock, A. H., Sept, D. & McCammon, J. A. (2001) *J. Phys. Chem.* **105,** 1504–1518.
37. Cambillau, C. & Claverie, J. M. (2000) *J. Biol. Chem.* **275,** 32383–32386.
38. Elcock, A. H. (2001) *J. Mol. Biol.* **312,** 885–896.
39. Shindyalov, I. N. & Bourne, P. E. (1998) *Protein Eng.* **11,** 739–747.