Protein Science (2000), *9*:1106–1119. Cambridge University Press. Printed in the USA.

Prediction of amino acid sequence from structure

KAUSHIK RAHA,[1] ANDREW M. WOLLACOTT,[2] MICHAEL J. ITALIA,[3]
AND JOHN R. DESJARLAIS[2]

[1] Integrative Biosciences Program, Pennsylvania State University, University Park, Pennsylvania 16803
[2] Department of Chemistry, Pennsylvania State University, University Park, Pennsylvania 16803
[3] Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pennsylvania 16803

(RECEIVED February 14, 2000; FINAL REVISION April 7, 2000; ACCEPTED April 14, 2000)

Abstract

We have developed a method for the prediction of an amino acid sequence that is compatible with a three-dimensional backbone structure. Using only a backbone structure of a protein as input, the algorithm is capable of designing sequences that closely resemble natural members of the protein family to which the template structure belongs. In general, the predicted sequences are shown to have multiple sequence profile scores that are dramatically higher than those of random sequences, and sometimes better than some of the natural sequences that make up the superfamily. As anticipated, highly conserved but poorly predicted residues are often those that contribute to the functional rather than structural properties of the protein. Overall, our analysis suggests that statistical profile scores of designed sequences are a novel and valuable figure of merit for assessing and improving protein design algorithms.

Keywords: genetic algorithm; homeodomain; multiple sequence alignment; Pfam; profile; protein design; RRM; SH3

There has been considerable recent success in the development of computational methods for the design of protein sequences, at various degrees of sophistication. Several groups have presented results in which computer algorithms were used to design novel hydrophobic cores for proteins (Hellinga & Richards, 1994; Kono & Doi, 1994; Desjarlais & Handel, 1995; Dahiyat & Mayo, 1996, 1997b; Lazar et al., 1997), in many cases with experimental validation of the proteins by biophysical and/or structural methods (Desjarlais & Handel, 1995; Dahiyat & Mayo, 1996, 1997b; Lazar et al., 1997, 1999; Kono et al., 1998; Johnson et al., 1999). Additional developments in core design have included the incorporation of backbone flexibility in a number of ways (Harbury et al., 1995; Su & Mayo, 1997; Desjarlais & Handel, 1999). In a particularly noteworthy case, this led to the successful design of a novel right-handed coiled-coil motif (Harbury et al., 1998).

Mayo and colleagues have pioneered the development of algorithms for noncore (Dahiyat et al., 1997a) and full sequence design (Dahiyat & Mayo, 1997a; Dahiyat et al., 1997b), using parameterized force fields and sophisticated optimization methods (Desmet et al., 1992; Goldstein, 1994). These methods were used successfully to design a sequence that adopts the zinc finger fold with no requirement for zinc binding (Dahiyat & Mayo, 1997a). The force fields used for these design processes have been parameterized over time by comparison between the calculated and experimentally determined folding stabilities of the designed proteins, a process referred to as the design cycle (Dahiyat & Mayo, 1996; Hellinga, 1997; Gordon et al., 1999; Street & Mayo, 1999). This is a sensible approach that has thus far worked extremely well.

We have developed a sequence prediction algorithm (SPA) for the design of complete protein sequences for moderately sized backbone templates. We have also explored a novel approach to the evaluation and parameterization of this algorithm that is complementary to efforts that rely on feedback from experimental stability data. This approach involves an in-depth analysis of the ability of SPA to design or predict sequences that are similar to those that exist naturally for a given fold. Using four protein motif superfamilies and representative structural templates from each, we demonstrate the ability of SPA to design sequences that look, by statistical profiling methods, as if they belong to the natural superfamilies.

Results

A sequence prediction algorithm (SPA) has been developed to design amino acid sequences that are consistent with a given three-dimensional (3D) backbone structure. The algorithm depends on a combination of filtering, sampling, and optimization procedures, and a relatively straightforward scoring function. This function is a combination of the Amber/OPLS force field (Weiner et al., 1984; Jorgensen & Tirado-Rives, 1988) and additional terms that account implicitly for solvation effects (Eisenberg & McLachlan, 1986). We have also recently derived an important set of terms, referred to as amino acid baseline corrections, which are critical for maintaining reasonable compositions of the designed sequences. The sequence selection process involves a combination of filtering

Reprint requests: John R. Desjarlais, Department of Chemistry, 408 Chandlee Lab, University Park, Pennsylvania 16803; e-mail: jrd@chem.psu.edu.

Table 1. *Motif families and representative structures*

Family	Abbreviation	Aligned sequences	Crystal structure[a]	Protein	Resolution (Å)
SH3	SH3	463	1shg	Spectrin	1.8
RNA recognition motif	RRM	850	1urn	U1A	1.9
Fibronectin type III	FNIII	1923	1ten	Tenacin	1.8
Homeobox	HM	1067	1enh	Engrailed	2.1

[a] Structural references are as follows: 1shg (Musacchio et al., 1992); 1urn (Oubridge et al., 1994); 1ten (Leahy et al., 1992); 1enh (Clarke et al., 1994).

criteria for the choice of input side-chain rotamer possibilities, and a genetic algorithm to perform the combinatorial search for a low scoring sequence/structure. The algorithm and its parameters are described in more detail in Methods.

In this study, we treat the design problem as one of sequence prediction. In other words, rather than attempting to design novel sequences that are different from natural sequences, we assess the ability of a computer algorithm to predict amino acid sequences that are similar to naturally existing sequences that adopt the same tertiary structure as the target. Only the backbone structure is used as input, with no prior knowledge of the native sequence or composition.

Prediction of sequences for superfamily structural motifs

A count of the number of identities between a predicted and native sequence can be used as a simple assessment of the predictive ability of a design algorithm. However, the large amount of sequence degeneracy expected (Bowie et al., 1990) and observed for many protein families suggests that this analysis is limited, and potentially misleading. To fully assess the ability of SPA to predict sequences appropriate for a given structure, we have explored its ability to predict sequences for a small number of structural motifs that belong to protein superfamilies.

We have chosen four protein superfamily motifs for our analysis. These are the SH3 domain, the homeodomain (HM), the fibronectin type III (FNIII) domain, and the RNA recognition motif (RRM). Each of these families are comprised of over 400 evolutionarily related sequences. For convenience, we have used the Pfam alignments (Bateman et al., 2000) of each protein family in our analysis. A single representative high-resolution crystal structure was chosen from each family as a structural template for the design algorithm. Table 1 lists these and the number of sequences contained in edited versions of the Pfam alignments.

The results of sequence prediction on each of the structural motifs are shown in Figure 1. These experiments were performed using SPA with a fixed set of optimized parameters. The agreement between the predicted sequences and the native sequence of each backbone template is significant, with sequence identities ranging from 24–28%. The extent of similarity is remarkable, considering that the only information used for the design process was the backbone structure itself. Koehl and Levitt (1999a, 1999b) recently described similar levels of success for a full-sequence design method. However, their use of the native sequence composition as a constraint provides a significant statistical advantage to the prediction process.

Although these prediction results are very encouraging, it is difficult to fully evaluate the predictive success of the design algorithm by comparing individual sequences, as discussed above.

Profile analysis of predicted sequences

The designed sequences for each of the structural motifs have significant similarity to the native proteins from which the backbone structure was derived. However, a large number of positions are predicted to contain nonnative amino acids. Are these amino

```
KELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFVPAAYVKKLD                                SH3 native
|    |           ||   |      |  | |||   |    | ||
KLKAVAKTNFNASKNEKVTVRQGQQVVIKNPQNPDWWVIQVGGQPGIVPSQILRIRN                                SH3 design 24%

RPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKI                                   HM native
     |   |   |   |             ||  || |    | ||  |
KGKIQFTNWQKDYLNDFFASTSEPSSPELEDLSKKLGINVEELKKWFSEKKEAD                                   HM design  24%

NHTIYINNLNEKIKKDELKKSLHAIFSRFGQILDILVSRSLKMRGQAFVIFKEVSSATNALRSMQGFPFYDKPMRIQYA          RRM native
|  |   | |       | |     |  ||||  | |  |        | |     |  |    ||  |      |
NNVIFVQNINPNVNEPDLQKQIYRKFIQFGQIIEINVNKSNQNIKKVWVVFVRPDAADEAKKAKQGDKFNNTKLVIKKN          RRM design 29%

LDAPSQIEVKDVTDTTALITWFKPLAEIDGIELTYGIKDVPGDRTTIDLTEDENQYSIGNLKPDTEYEVSLISRRGDMSSNPAKETFTT    FNIII native
|  |  |   |       |   ||   ||||    |   |  |   | |  |       || | |          |          |
LIPPQDIKPQDKRPKKVIIWFVKPDTYIDGILVIFGVLWVDNDLKPIWLKKDQLVKPFYNLNPNTVFKIVIVAKKGSDKAPFKEFEFSA    FNIII design 28%
```

Fig. 1. Comparison of designed and native sequences. Protein sequences designed by SPA for the four structural motifs are compared to the native sequence from which the backbone template was derived. Identities are marked with vertical bars. The percentage identity between each designed sequence and its corresponding native sequence is also listed. The region of each sequence that corresponds to its respective Pfam alignment is underlined.

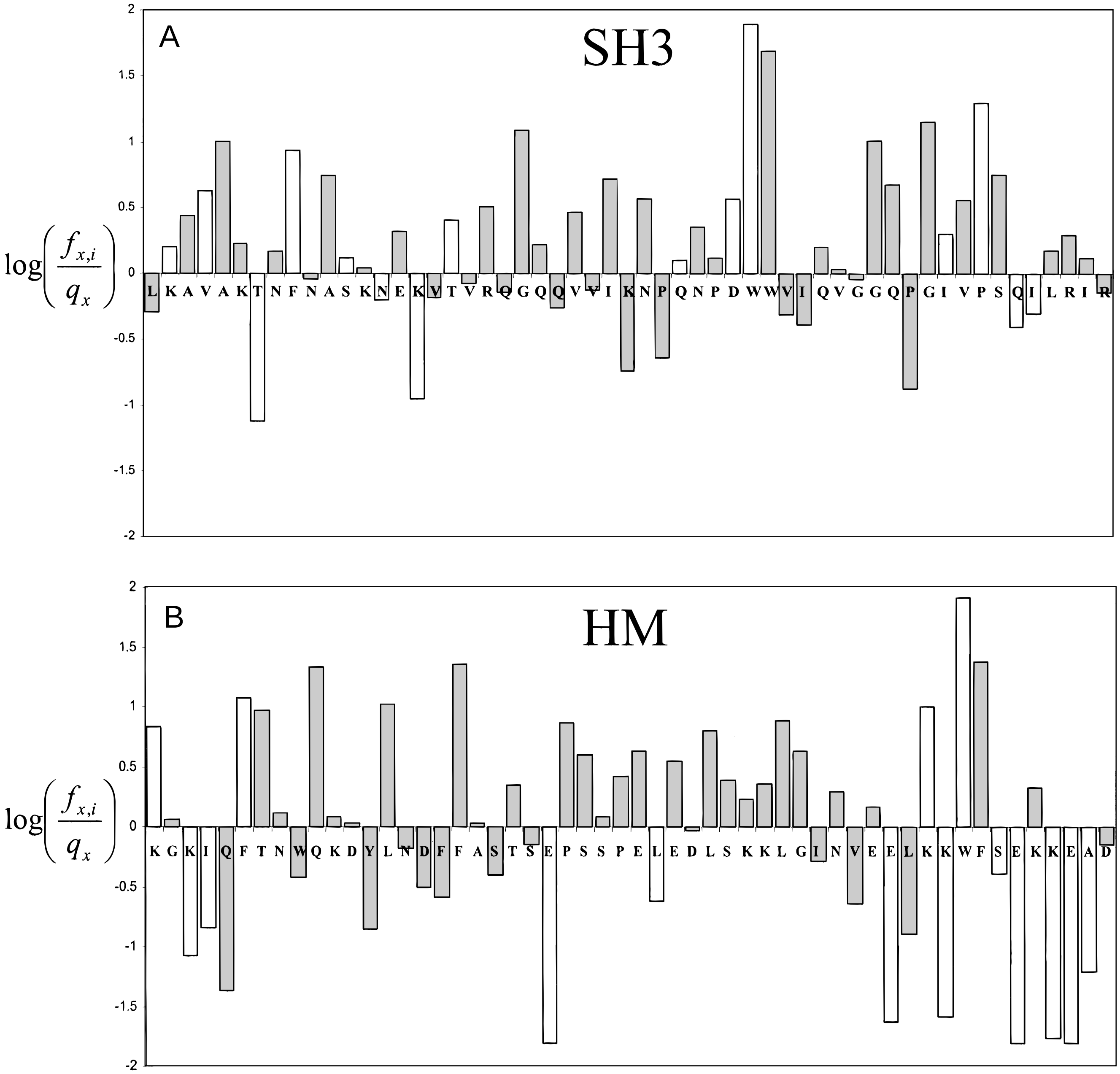


Fig. 2. Profiles of designed sequences compared to multiple sequence alignments. Log-odds ratios were calculated for each position of each designed sequence, using the Pfam alignments listed in Table 1. Positive values represent positions for which the designed amino acid occurs more frequently than random in the natural sequences. Negative values represent positions for which the designed amino acid has been selected against. Positions for which the native or designed side chain makes a close contact with ligand (if any atoms of the side chain are within 4 Å of a ligand atom) are designated with open bars. Note the strong correspondence between positions with negative values and those participating in functional contacts. (*Figure continues on facing page.*)

acids reasonable alternatives? This question can be addressed by comparing the designed sequences to a large family of sequences expected to adopt the same 3D structure.

For each of the superfamily motifs used in our prediction experiments, over 400 nonredundant, evolutionarily related, sequences can be aligned. With the reasonable assumption that all of these sequences fold to approximately the same structure, the large number of sequences provides us with a statistical evaluation of the suitability of a designed amino acid at any position. Designed residues that have been selected against in the natural sequences will be found rarely in the aligned sequences, whereas those that contribute favorably to the folding or function of the protein will be found frequently. We draw from established methods (Durbin et al., 1998) for estimating the statistical likelihood that a newly determined protein sequence belongs to an existing family of related proteins, using a profile derived from a multiple sequence alignment of the family (Gribskov et al., 1987). Here, we use the method to determine if a designed (predicted) protein sequence

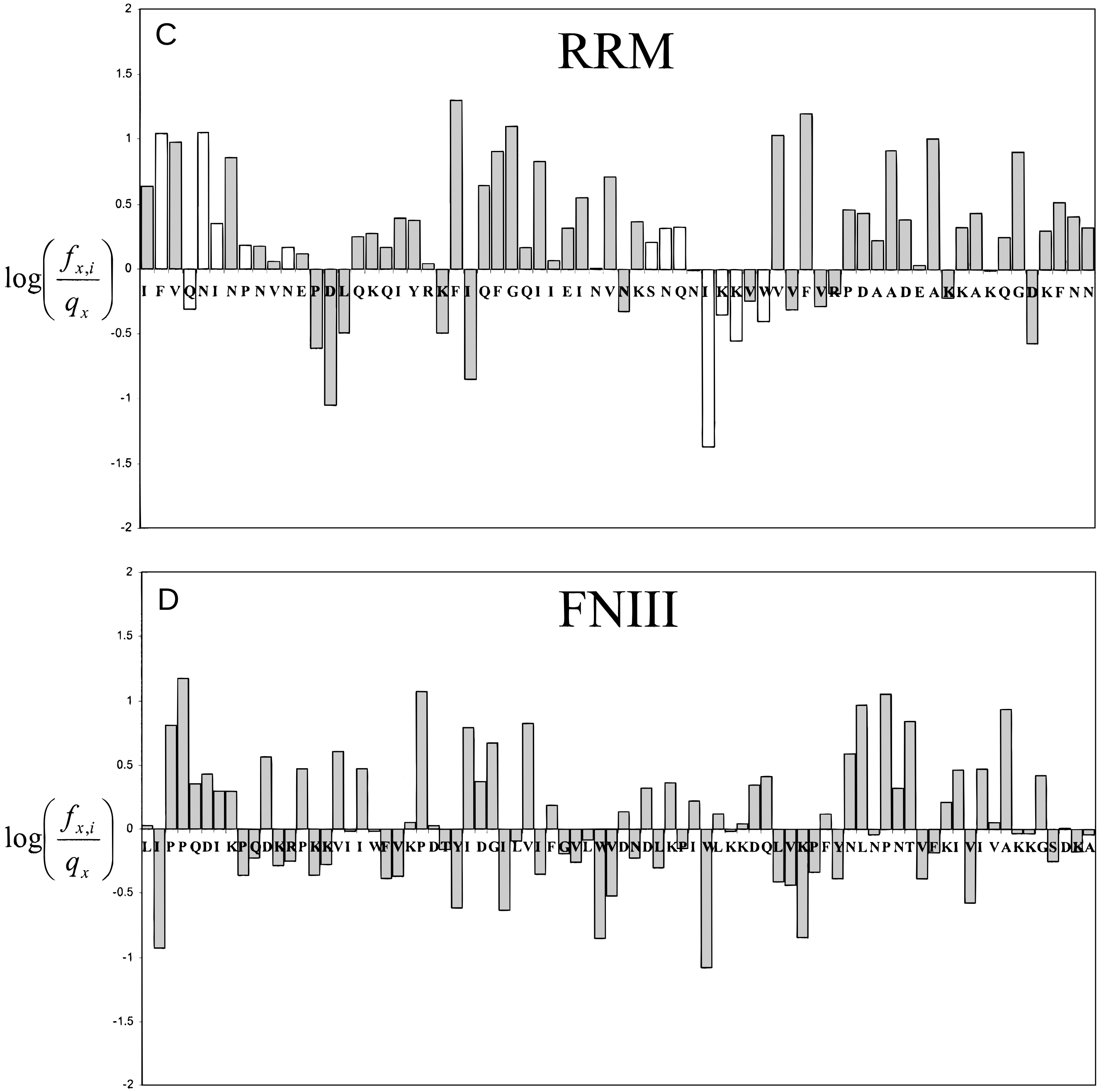


Fig. 2. *Continued.*

would be categorized as an evolutionary member of the family from which its structural template was derived (if it was not known that the sequence was indeed designed).

For each designed sequence, we determined the frequency of occurrence of each of its amino acids in the Pfam (Bateman et al., 2000) sequence alignment for the corresponding superfamily. The alignment of the designed sequence to the natural family is set to be identical to the alignment of the natural sequence from which the structural design template was derived. A log-odds ratio relative to a random model was defined for each designed amino acid as follows:

$$\log_{10}(f_{x,i}/q_x)$$

where $f_{x,i}$ is the frequency of the designed amino acid type x at position i in the alignment, and q_x is the overall frequency of occurrence of amino acid type x in all known proteins. Positive log-odds values represent positions for which the designed amino acid occurs more frequently than random, or has been selected for. Negative values represent positions for which the designed amino acid has been selected against. Plots of these values for the designed sequence of each motif are shown in Figure 2. A corresponding structural map of the log-odds ratio values is shown in

Figure 3 for the SH3 motif design: the map reveals that in this case conserved amino acids predicted by SPA are dispersed throughout the structure.

There are several important trends in the data that indicate that the design algorithm has a strong predictive ability. First, one observes that many of the most highly conserved positions are, in fact, correctly predicted by SPA. Furthermore, many of the positions for which the predicted amino acid differs from the native contain amino acids that occur at significant frequencies in the alignment. These substitutions might represent neutral substitutions to the sequence. At most of these positions, the frequency of occurrence of the designed amino acid significantly exceeds that expected from random occurrence, suggesting that their selection is based on structural considerations. Interestingly, there are a small number of positions for which a designed but nonnative amino acid occurs at a frequency higher than that of the native amino acid.

The general trend in the results is that amino acids in the designed sequences are found at a lower frequency than those of the native. This is, of course, not surprising and could result from the combination of several effects. First, we assume that the design algorithm is not perfect, particularly with respect to the accuracy of the potential energy function. Second, because only structure is considered by SPA, amino acids conserved for functional reasons alone will not be predicted. Indeed, for several of the motifs, positions for which highly conserved amino acids are not predicted by SPA are found to be close to the site of functional interaction with the cognate ligands of these proteins, as discussed below. Third, our design algorithm does not consider determinants of folding kinetics, so amino acids conserved for those reasons may constitute an additional subset for which SPA performs poorly. Finally, there is some possibility that the native sequences are not fully optimized, and that the designed amino acids represent potential improvements to the protein.

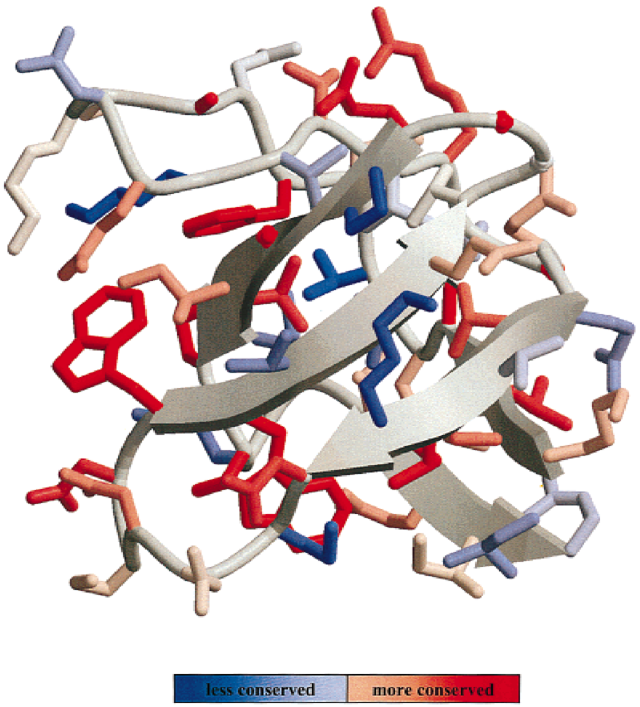


Fig. 3. Structural distribution of designed amino acid conservation in an SH3 domain. Amino acids designed by the program SPA are color coded according to their extent of occurrence at the same position in natural SH3 proteins, calculated as a log-odds ratio compared to random occurrence frequencies (Fig. 2). This figure was prepared using MOLSCRIPT (Kraulis, 1991).

Table 2. *Average log-odds values for subsets of amino acid and structural types*

	Helix	Sheet	Coil	Turn	N-cap	Functional[a]	Total nonfunctional
A	0.53	0.79	0.35	—	—	−1.22	0.57
D	−0.13	0.22	−0.12	0.24	—	0.57	0.05
E	−0.05	0.32	—	0.32	—	—	0.04
F	0.88	0.24	—	—	—	1.01	0.54
G	—	0.55	0.06	0.67	—	—	0.62
I	0.39	0.23	−0.18	—	—	−0.56	0.14
K	0.11	−0.04	0.02	0.05	—	−0.51	0.01
L	0.26	0.07	0.00	−0.08	—	−0.62	0.10
N	−0.04	0.56	0.12	0.17	0.28	0.22	0.20
P	0.08	−0.43	0.66	1.06	—	0.74	0.23
Q	0.58	0.15	−0.25	—	—	−0.07	0.18
R	0.04	0.08	0.13	—	—	—	0.04
S	0.05	—	—	−0.25	0.61	−0.02	0.09
T	—	—	0.59	—	0.96	−0.36	0.34
V	−0.67	0.15	−0.35	—	—	—	0.05
W	−0.45	0.20	—	−0.85	—	0.53	−0.14
Y	−0.25	—	−0.50	—	—	—	−0.37
Total	0.16	0.16	0.09	0.28	0.62	−0.12	

[a] Defined as in Figure 2. These positions were not included in the statistics for the other classes.

Profile scores

The concept of multiple sequence profiles is clearly useful for evaluating the performance of a sequence prediction algorithm. The analysis can be carried further with the designation of a single profile score for each of the designed sequences, taken as the sum of the position-specific log-odds ratios defined above. This single numerical value can be used to report on the merit of a designed sequence. More importantly, this figure can be used to evaluate the performance of SPA for various combinations of parameters, as discussed below.

Although the designed sequences are shown above to have amino acids with generally lower frequencies than those associated with their structural templates, each of the families contain a large number of sequences, with a wide range of individual profile scores. In Figure 4, we show the distribution of calculated profile scores for each member of the pertinent Pfam alignment, compared to a distribution of profile scores for a set of randomly generated sequences. As an additional control, we generate random sequences constrained by a simple HP model for each template structure, using a contact score definition of buried versus exterior positions in the structure (Micheletti et al., 1998). The score of each sequence predicted by SPA is also highlighted. In all cases, the profile score of the predicted sequence is several standard deviations above both random sequence distributions. In some cases, the sequence predicted by SPA scores better than a significant fraction of the native sequences contained in the Pfam alignment.

These results convincingly demonstrate the ability of SPA to predict sequences that look like evolutionary relatives of natural protein families, using nothing more than the coordinates of the backbone structure from a single member of the family. The designed sequences, however, generally score lower than a majority of the natural sequences. In the sections that follow, we examine more closely the contributions of individual designed amino acids to the overall profile score.

Prediction of conserved interaction patterns

Inspection of the designed structures and their corresponding log-odds plots reveals the structural location of positive and negative contributions to the profile score. Many of the strongly positive contributions to the profile score come from de novo prediction of highly conserved hydrophobic core residues. This is not surprising, given the considerable success reported for computational design of hydrophobic cores (Hellinga & Richards, 1994; Kono & Doi, 1994; Desjarlais & Handel, 1995; Dahiyat & Mayo, 1996, 1997b; Lazar et al., 1997; Kono et al., 1998). A structural view of hydrophobic core prediction for the RRM motif is shown in Figure 5A.

The hydrophobic core residues constitute only a subset of positions for which SPA predicts conserved amino acids. Others include conserved hydrogen bonding interactions, an example of which is shown in Figure 5B. Our potential function is defined to explicitly penalize the burial of polar atoms unless their hydrogen bonding potential is satisfied. Many of the conserved interactions predicted by SPA involve a polar backbone atom that would be buried in the absence of a complementary polar side chain. The fact that these positions are highly conserved in the native proteins underscores their importance in maintaining structure and stability.

A third type of predicted interaction is shown in Figure 5C, where a designed Lys-Asp salt-bridge interaction at positions 22 and 64 of the RRM motif is displayed (numbered according to Fig. 2C). Although each of these amino acids occurs infrequently at those positions in the RRM alignment, their occurrence is highly correlated, suggesting that the interaction is favorable and specific.

There are many other positions for which SPA predicts a residue that contributes favorably to the profile score, but in which the selective pressure, as gauged from the natural sequences, is less dramatic. At these positions, SPA appears to be sensitive to subtle combinations of influence from steric, solvation, and electrostatic effects.

As shown in Table 2, the ability of SPA to predict conserved amino acids is not particular to structural class. However, some amino acid types, such as Tyr and Trp, are found to contribute unfavorably to the profile score more frequently, suggesting that our potential function can be further refined.

Structure vs. function

While in general each designed sequence contains a large number of amino acids found frequently in the natural alignments, there are in all cases a significant number of positions for which SPA predicts an amino acid that has been selected against in the natural sequences. Inspection of the structural location of these positions indicates that a large fraction of them cluster to the functional interaction sites of the molecules (Table 2). Examples of such positions are highlighted in Figure 2. Structural representations of this effect are shown in Figure 6 for the SH3 and RRM domains. As shown in Figure 6A, designed amino acids Thr7, Lys16, Gln47, and Ile48 (numbered according to Fig. 2), all of which have negative contributions to the designed SH3 profile score, are located in the peptide binding groove of the natural SH3 domains. Because peptide binding is not included as a constraint in our design procedure, SPA selects amino acids that are consistent with the structure alone, but have been selected against for functional reasons—they are apparently incompatible with peptide binding. A similar example is observed in the design of a larger set of amino acids within the RRM structure. These amino acids, all of which have negative contributions to the designed RRM profile score, cluster on the RNA binding face of the U1A structure, according the structure of the complex of U1A with an RNA hairpin (Oubridge et al., 1994), as shown in Figure 5B. Interestingly, in both cases there are a number of nearby residues that are highly conserved and correctly predicted by SPA, suggesting some overlap of functional and structural conservation. Finally, 11 out of 54 residues in the design of a homeodomain sequence appear to have negative contributions to the profile score because of their proximity to the DNA ligand in the functional interface (Kissinger et al., 1990; Fraenkel et al., 1998). This substantial fraction of designed and nonconserved residues largely accounts for the lower profile score of this designed sequence relative to those designed for the other motifs. A similar analysis for the FNIII domains is not straightforward, as the interaction sites of these modules is varied and difficult to clearly define.

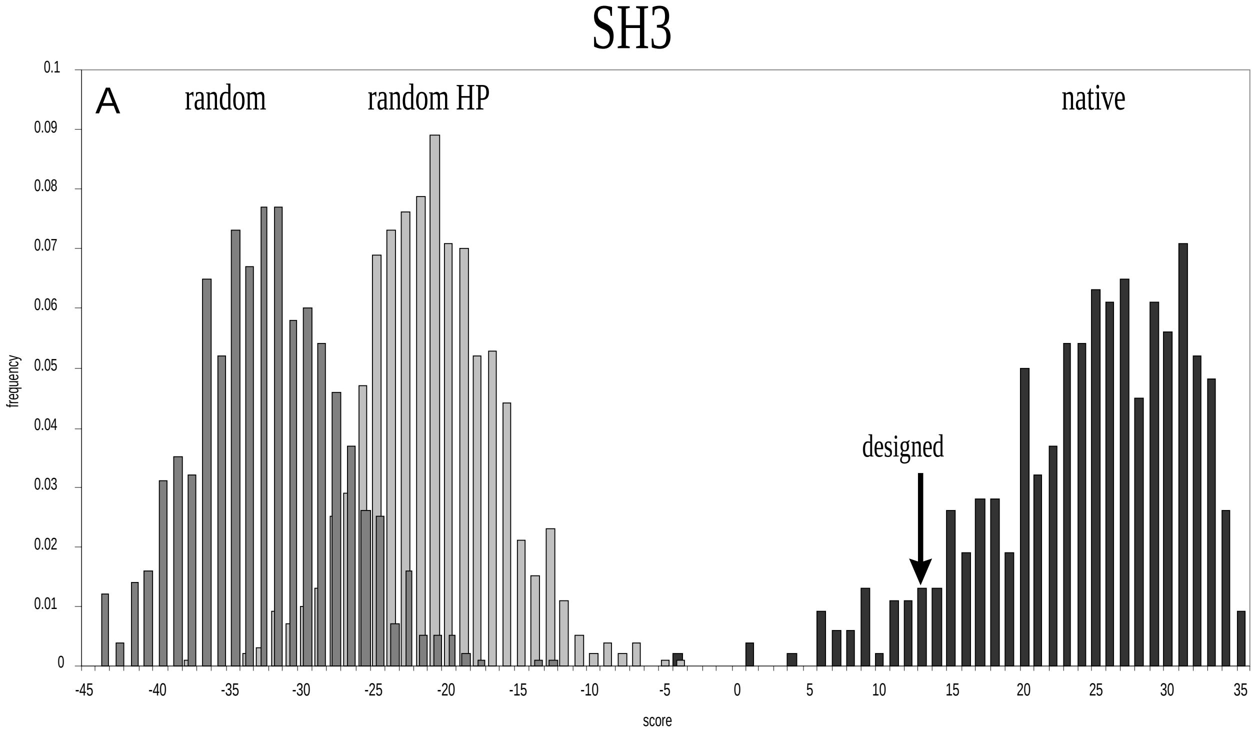
Structure vs. folding

Our design algorithm only calculates the compatibility of a sequence with a backbone structure. In no manner does it consider the folding process itself. Interestingly, for all of the designed sequences except the homeodomain, at least one designed proline has a strong negative contribution to the profile. Because prolines frequently contribute to slow isomerization folding phases, it is possible that the presence of excessive prolines in proteins is selected against. Although there may be other predicted amino acids that compromise the ability of the protein to fold quickly, they are not obvious in the analysis performed here. However, future experimental characterization of such designed sequences will provide interesting insight into this issue.

Prediction of structurally precise sequences

The statistical agreement between the SPA-predicted sequences and those of each corresponding superfamily is noteworthy and highly encouraging. These results suggest that SPA might have an ability to design proteins that fold stably and uniquely to the input structures. Another important feature of natural proteins is their ability to fold to a precise and ordered structure. Such precision has been shown in various studies to be closely linked to the functional integrity of the protein. For example, Sauer and colleagues, studying a hydrophobic core variant of λ-repressor, showed that a difference in backbone structure of 0.3 Å root-mean-square deviation can result in a significant loss in DNA binding affinity (Lim et al., 1994).

We have assessed the ability of SPA to predict structurally precise sequences by comparing the number of identities between the designed sequence and the native sequence of the structural tem-



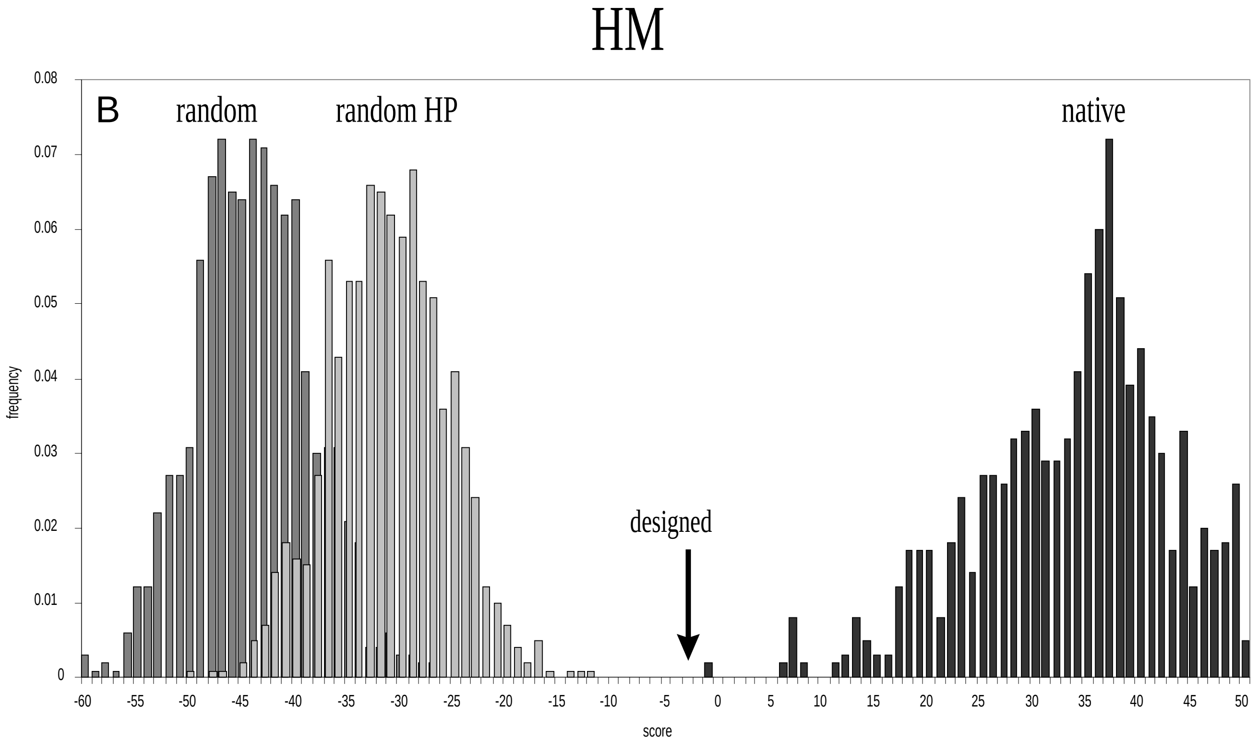


Fig. 4. Distributions of profile scores for random and natural sequences. (**A**) SH3; (**B**) homeodomain; (**C**) RNA recognition motif; and (**D**) fibronectin type III. Log-odds profile scores based on the Pfam alignments were calculated using random, designed, or natural sequences. For each structure, a simple hydrophobic-polar (HP) model was constructed and used as a constraint for generating the random HP sequences (see Methods). (*Figure continues on facing page.*)

plate and between the designed sequence and all other native sequences in the alignment. The results of this analysis, shown in Figure 7, are intriguing. For the RRM and FNIII motifs, the designed sequences are significantly more similar to the native sequence of the backbone template than to most other members of the family, demonstrating that SPA is sensitive to the idiosyncra-

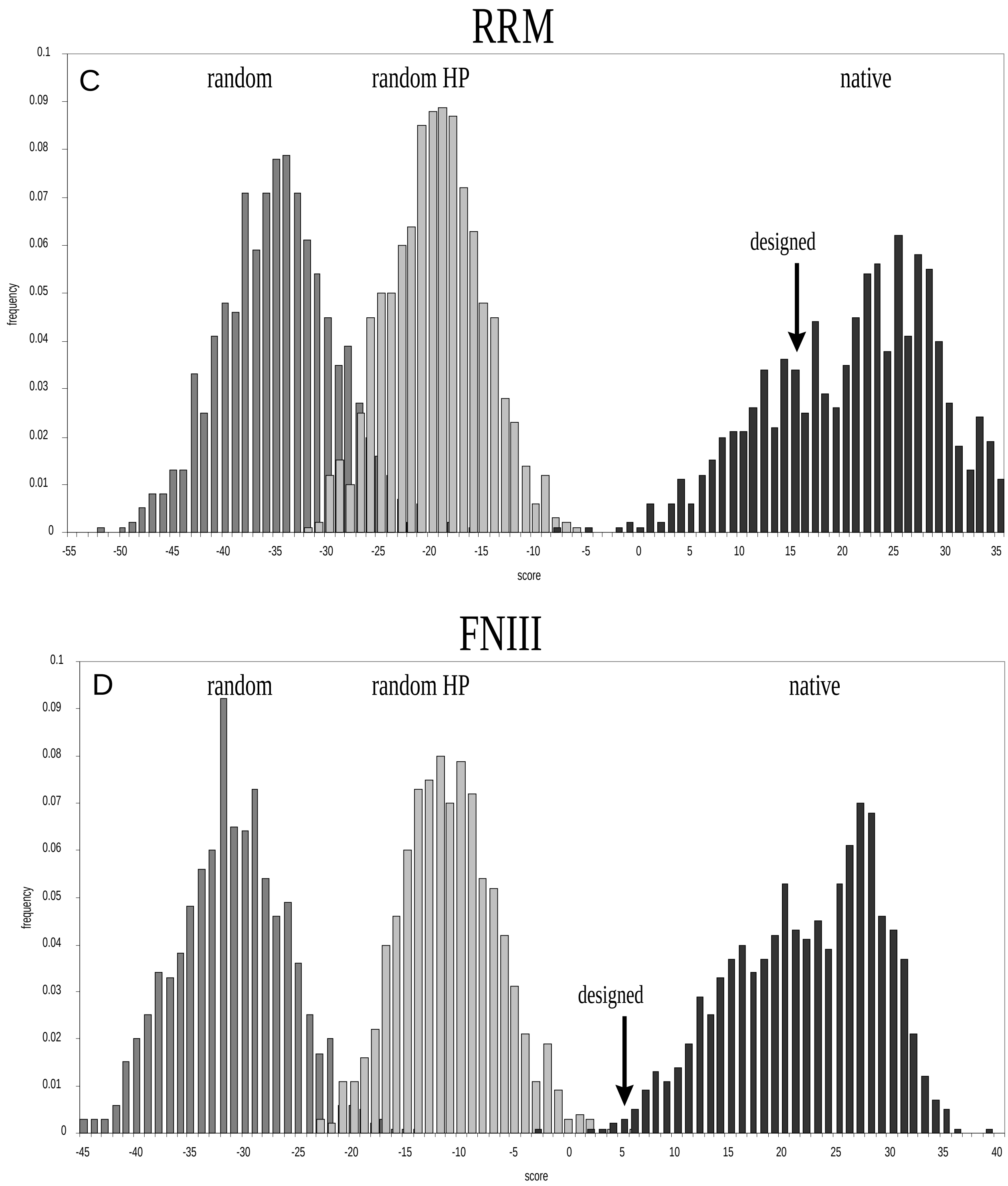


Fig. 4. *Continued.*

sies of the template structure. In contrast, for the HM and SH3 families, the designed sequence shares similar levels of identity with many members of the family. One possible explanation for this is that the natural members of these families have generally more similar structures, leading to less distinctive pressure toward particular sequences.

Parameterization of design algorithms using profile scores

A high profile score is generally taken as evidence that a protein sequence belongs to the protein family from which the profile was derived. In general, high profile scores for designed sequences are thus desirable. Pursuing this notion, we can evaluate the influence

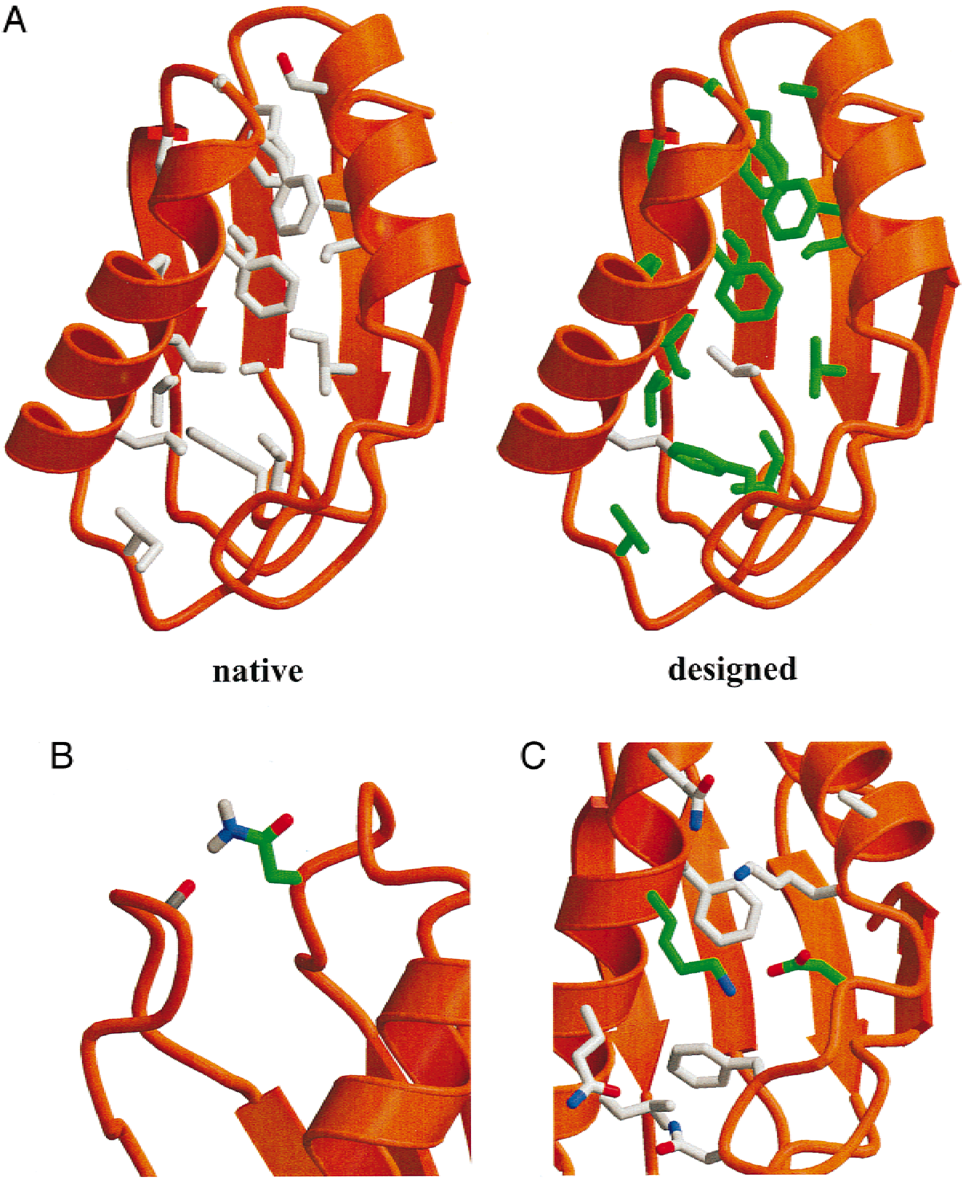


Fig. 5. Representative amino acids in a designed RRM sequence. **A:** Hydrophobic core amino acids predicted by SPA and found frequently in natural RRM sequences are highlighted in green. The core residues of the native protein are shown for comparison. **B:** Prediction of the highly conserved amino acid Asn7. The orientation is similar to that observed in the native structure. **C:** Prediction of an infrequent but highly correlated amino acid pair of Lys22 and Asp64, suggestive of a potential salt-bridge interaction.

of scoring parameters on the performance of SPA by calculating profile scores of sequences designed using different parameter conditions. In fact, the solvation parameters used for the simulations shown here were derived in part by a coarse search for the combination of parameters that gave the best overall profile scores for the four motifs. Although the analysis was incomplete due to time constraints on the simulations, the removal of the current set of solvation parameters strongly affects the profile scores of the resulting designed sequences (Table 3). With the exception of the FNIII design, removal of the solvation parameters results in a significant decrease in profile score.

Surprisingly, the profile score appears to be relatively insensitive to the removal of the amino acid baseline correction factors. Although this is true, compositional analysis indicates that the inclusion of these factors is important for generating reasonable sequence compositions. For instance, although the profile score of the SH3 sequence designed in the absence of baseline corrections factors is similar to that which includes them, it contains six Trp residues as opposed to the two found in the normal design and the native sequence. The values of the baseline parameters were derived in a separate study (K. Raha & J.R. Desjarlais, unpubl. data). Perhaps the use of profile scores will be a constructive strategy for further refinement of those parameters.

Future work will focus on a higher precision exploration of the effect of various parameter types and weights on the designed sequence profile scores. While the results shown here suggest that

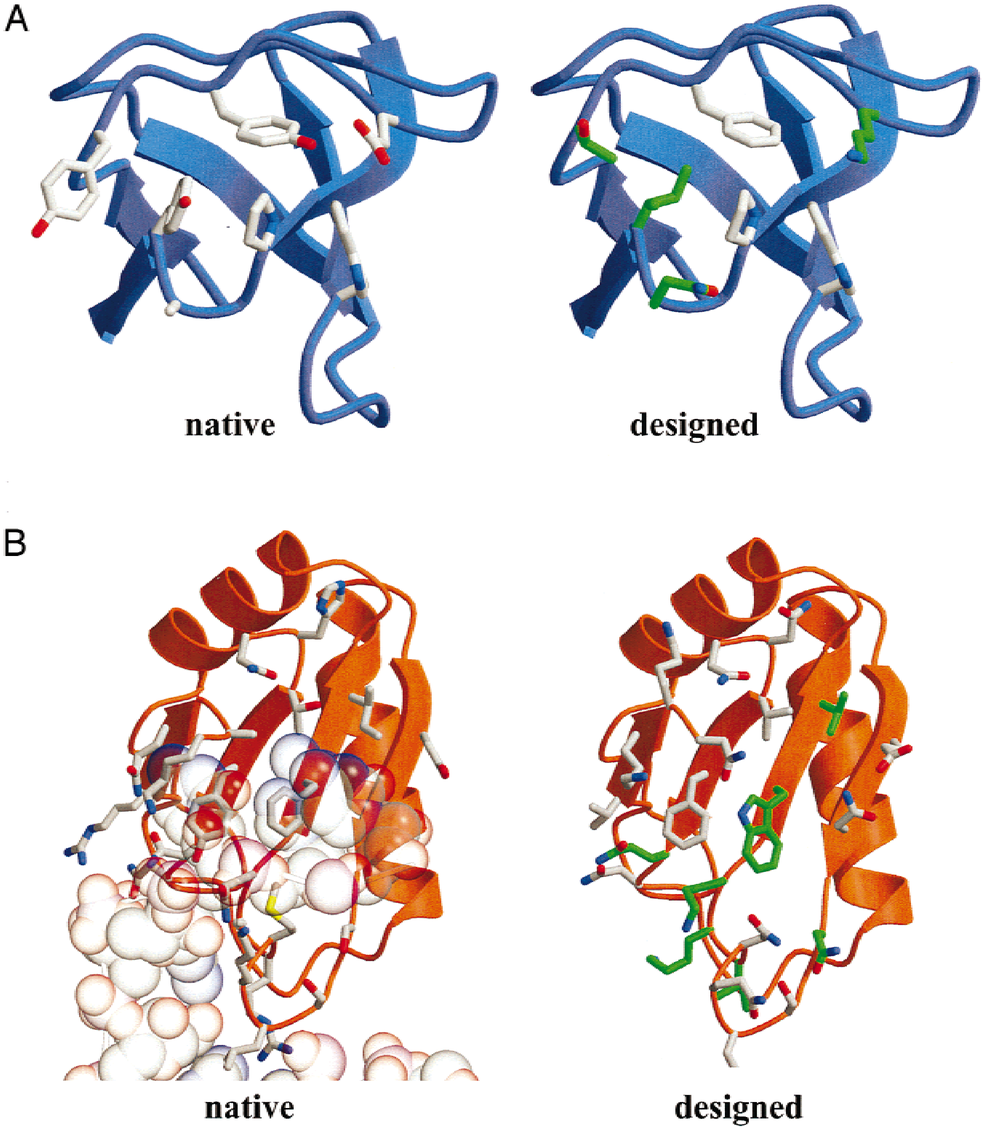


Fig. 6. Structure vs. function. Models of the replaced functional sites in designed proteins. **A:** View looking down into the peptide binding site of the spectrin SH3 domain and the designed SH3 domain. **B:** Comparison of the complex of U1A–RNA and the designed RRM protein. Residues shown in green are selected against in the natural sequences, corresponding to the highlighted bars in Figure 2 with negative log-odds values.

the parameters are reasonably well defined, some of our results suggest there is room for improvement.

Discussion

We have demonstrated the ability of SPA to design amino acid sequences that resemble natural members of protein families, using four representative superfamily motifs. For the SH3 and RRM families, the designed protein sequences score better than a significant fraction of the native sequences making up the alignment. Although this is highly suggestive of an accurate design algorithm, a more complete evaluation of the ability of SPA to design appropriate sequences for a structural motif will require experimental production and characterization of the designed proteins. We note that while a single deleterious mutation would have a relatively minor effect on the profile score of the designed sequence, it could completely preclude formation of the target structure. Nevertheless, the use of the profile score strategy described here has greatly facilitated our initial search for a well-balanced scoring function for computational protein design.

One significant potential advantage of the use of profile scores for preliminary evaluation of designed sequences relates to our supposition that a profile score, in contrast to experimental stability measurements, is a comprehensive measure of the compatibility of a sequence with a structure. Profile scores are likely to reflect a combination of protein traits, including stability, structural specificity, solubility, and perhaps even foldability. Of course, from a purely structural perspective, they also contain extraneous information: amino acids are often conserved because of functional constraints to which SPA is insensitive. In the cases studied here,

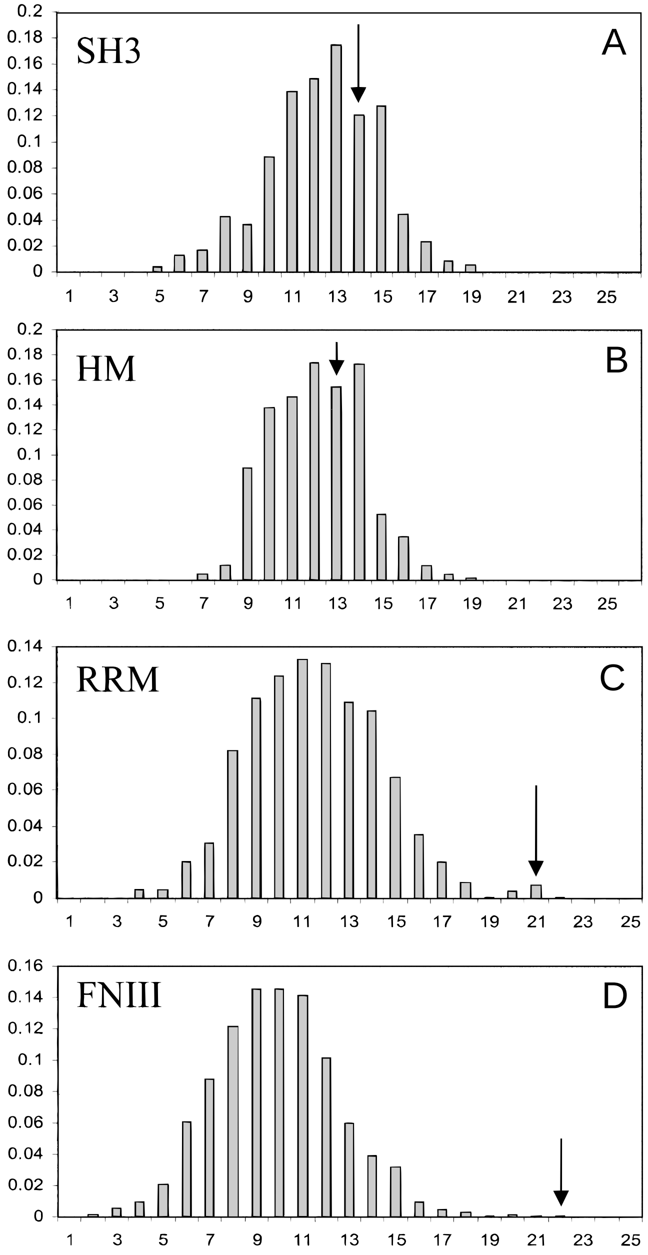


Fig. 7. Distribution of identities between designed and natural sequences. For each structural motif, the distribution of identities between the designed and natural sequences is plotted. The position of the sequence of the template structure is shown with an arrow.

Table 3. *Influence of parameter removal on log-odds profile scores*

Motif	Design score	No solvation	No baselines
SH3	12.9	−5.2	13.6
HM	−2.2	−9.9	−6.9
RRM	15.5	2.3	15.3
FNIII	4.7	5.4	12.3

however, positions with a dominant role in the functional activities of proteins are clearly demarked by their positions in structures of each template in complex with a cognate ligand.

The use of four distinct structural fold families in the evaluation of SPA leads us to the preliminary conclusion that the ability of SPA to properly design sequences for a structure is general. Given the historical difficulty of designing β-sheet versus α-helical proteins (Hecht, 1994), the level of apparent success on the β-rich SH3 and RRM motifs is noteworthy. Interestingly, the success does not depend on the use of any explicit secondary structure propensity term.

The sizes of the designed motifs studied here, while modest compared to many natural proteins, are significant when compared to previously reported computationally designed proteins. This was achieved using fairly modest computer resources, similar to those available to many modern laboratories, made possible by the use of a finely tuned genetic algorithm for the combinatorial optimization. Although the merits of various deterministic vs. stochastic search methods such as a GA are often contrasted (Desjarlais & Clarke, 1998), the results shown here suggest that a genetic algorithm approach is sufficient for arriving at reasonable sequences for a defined target structure.

Applications of the technology described here include its eventual use for the design of novel proteins, or the modification of existing proteins for improved properties. As recently discussed by Koehl and Levitt (1999b), such algorithms can also be used to generate diverse sets of virtual sequences that would be useful for protein fold recognition. Finally, application of SPA combined with multiple sequence analysis might eventually be used to predict functional regions of proteins by virtue of its tendency to predict nonconserved amino acids at functional positions.

Methods

Description of the sequence prediction algorithm (SPA)

Potential function and geometries

The Amber potential function (Weiner et al., 1984) with the OPLS nonbonded parameters (Jorgensen & Tirado-Rives, 1988) is used as a basis for evaluation of the energies of protein models with different sequences. Our form of the potential includes most of the terms of the Amber potential: nonbonded, electrostatic, and torsional energies. Fixed bond lengths and angles (set at the equilibrium values described for the Amber force field) are used for side-chain geometries, eliminating the need for bond stretching and angle bending terms. The energy of a model is therefore calculated as follows:

$$E = \sum_{torsions} \frac{V_n}{2}[1+\cos(n\chi)] + \sum_i \sum_{j>i} 4\epsilon \left[\left(\frac{\sigma}{R_{i,j}} \right)^{12} - \left(\frac{\sigma}{R_{i,j}} \right)^{6} \right] + \frac{q_i q_j}{DR_{i,j}} + \sum_i S_i \, \Delta A_i + \sum_{x=1}^{20} n_x B_x$$

where R_{ij} is the distance between atoms i and j; σ and ϵ are the Lennard–Jones parameters related to the radii and well depth, respectively. The first term is a sum over side-chain dihedral angles; the second term is a sum of nonbonded (Lennard–Jones) interactions over all atom pairs (side chain–side chain and side chain–

Table 4. *Atomic solvation parameters*

Atom types	Solvation parameter (cal/mol/Å^2)
Buried C	+20
Buried N,O	−120
Exposed C	+60

backbone); the third term is a sum of electrostatic interactions summed over all charged atom pairs. Scaling factors for the nonbonded and electrostatic terms, and combining rules for σ and ϵ, are those defined for use of the OPLS parameter set. In the current version of the algorithm, backbone geometries are fixed, so backbone self-energy terms are not evaluated.

The fourth term ($\Sigma_i S_i \Delta A_i$) is used to represent the solvation energetics of the system (Eisenberg & McLachlan, 1986). The solvation free energy of a model structure is determined by summing the products of the atomic solvation parameter and the estimated change in solvent accessible surface area for each atom in the model structure, where the change is relative to an estimate of the average exposure of that atom type in the unfolded state of the protein. The use of atomic solvation parameters is expected to provide an approximation of the true solvation free energy and has been used effectively for protein design (Gordon et al., 1999). Furthermore, recent theoretical results indicate that, despite its simplicity, it can largely reproduce the energetics calculated using more sophisticated methods (Hendsch & Tidor, 1999). Here we use only three solvation parameters, corresponding to the burial of polar atoms (N,O), the burial of nonpolar atoms (C), and the exposure of nonpolar atoms (C). The first two terms represent conventional use of atomic solvation parameters, relating to the free energy cost of desolvation of polar groups and the strength of the hydrophobic effect, respectively. The desolvation penalty for the burial of polar atoms is only calculated when the atom is not participating in a hydrogen bond. This is assessed using the condition that the distance between the hydrogen atom and the acceptor atom is <2.5 Å, and if the following function has a value less than −0.3:

$$f(\theta,\phi) = \cos^2(\theta_{D,H,A})\cdot\cos(\phi_{H,A,AA})$$

where D, H, and A refer to the donor, hydrogen, and acceptor atoms, respectively, and the AA refers to the acceptor antecedent atom. If a polar atom is indeed involved in a hydrogen bond, we use the approximation that there is no desolvation penalty associated with that atom. The final term, a penalty factor for exposure of nonpolar surface, has been applied successfully for designing proteins by Mayo and colleagues (Dahiyat et al., 1997b; Gordon et al., 1999), and may be considered to be both an implicit fold-specificity constraint and a solubility constraint.

The strengths of the three parameters for the simulations performed herein were derived from a coarse grid search over combinations of the parameters. The values that appeared to give the best overall results in terms of designed sequence profile scores are shown in Table 4. The strengths of these parameters are within the range of similar parameters derived from other studies (Juffer et al., 1995).

Amino acid baseline corrections

We have recently generated a set of correction factors to account for changes in amino acid sequence within the design process (K. Raha & J.R. Desjarlais, unpubl. obs.). These factors account for the absence of an explicit reference state in the calculation of the energy of a designed sequence. We refer to the factors as amino acid baseline corrections. Because the terms depend only on the identity of amino acid at each position, the correction factors depend on composition only. The application of these 20 factors is straightforward and is of the following form:

$$F = \sum_{x=1}^{20} n_x B_x$$

where F is the total compositional correction term, n_x is the number of times amino acid type x occurs in the designed sequence, and B_x is the baseline correction factor for amino acid x, given in Table 5.

Side-chain sampling and optimization

A rotamer library of statistically prevalent combinations of side-chain dihedral angles (Dunbrack & Cohen, 1997) is used to guide sampling of side-chain identities and orientations in the combinatorial search for low energy structures (all amino acids except Cys, His, and Met were used for sequence prediction). Additional flexibility is incorporated by adding discrete increments of ±15° to each dihedral angle of each library rotamer.

A genetic algorithm (Holland, 1992) is used for performing the combinatorial search. An initial population of 300 members is generated by creating models with side chains at each position sampled randomly from the rotamer library. This sampling is biased according to a Boltzmann probability of the rotamer, calculated from its energy of interaction with the backbone structure and a temperature of 2,000 K. The energy of each model in the population is calculated according to the scoring function described above. Based on these energies, selective recombination between models is performed using a uniform crossover scheme. Parent models are selected from a roulette wheel weighted according to the Boltzmann probability of the model, calculated from its energy and a temperature that is set at each round according to a predefined diversity value. This value, defined as the informational entropy of the population, is set to decay linearly from 5.5 to 3.0 throughout the simulation. Finally a small amount of random mutation at a frequency of 0.04 is used to modify the population generated by crossover of parent models. This cycle of energy

Table 5. *Amino acid baseline correction factors*

A	−1.791	M	0.335
C	−0.402	N	−0.118
D	−0.251	P	−2.18
E	−0.02	Q	0.175
F	1.005	R	0.914
G	0	S	−1.001
H	0.771	T	−0.971
I	−0.864	V	−2.385
K	−0.002	W	2.823
L	0.066	Y	0.975

evaluation, selective recombination, and mutagenesis is repeated 200 times. The designed sequences reported here were derived from a two-stage process where five separate GA simulations were performed and the output of these runs was used to seed a final simulation.

Rotamer filtering

Because of the enormous combinatorial complexity involved in protein sequence optimization, we pre-filter the rotamer library for a given structural template. Filtering is based on steric and solvent effects. The steric filter is straightforward. For a given position, any rotamer that results in an energy of interaction with the backbone structure $>$20 kcal/mol is rejected. The second filter is designed to prevent the burial of polar groups or the hyperexposure of nonpolar groups. This filtering stage is performed as follows. Each possible side-chain rotamer is placed into a position on the backbone structure. The extent of burial of each of its atoms is then assessed relative to a set of generic side-chain centroid coordinates at all other positions, defined at 2.9 Å from the C_α atom along a standard geometry C_α–C_β bond vector. A contact score for each rotamer atom is defined as (Micheletti et al., 1998)

$$C_a = \sum_{i=1}^{chainlength} \frac{1}{1 + e^{d_{a,i} - 6.5}}$$

where C_a is the contact score for atom a, and $d_{a,i}$ is the distance between atom a and the side-chain centroid at position i. Rotamers of side chains containing polar atoms {D,E,K,N,Q,R,S,T,Y,W} are eliminated when any of their polar atoms have a contact score $>$5.5 and are incapable of forming hydrogen bonds with the backbone. Rotamers of nonpolar side chains {F,I,L,V,P,W} are eliminated when any of their atoms have a contact score $<$2.0. These criteria are defined conservatively because of the coarse nature of the definition of burial. Trp side chains are subject to both criteria. Ala and Gly residues are not subject to filtering.

Other groups have used definitions of surface, buried, and boundary positions to generate position-specific subsets of amino acid types. The approach described here obviates the need for explicit definition of burial class, in principle allowing appropriate subsets of rotamers from all amino acid types at some positions.

Profile analysis of designed sequences

Pfam alignments were downloaded from the Pfam web site (http://pfam.wustl.edu/index.html). All redundant entries were eliminated. Sequences with large gaps relative to the rest of the alignment were also eliminated so that designed sequences were compared only to proteins of similar structure. Only positions that align to the native sequence of the design template were used in the analysis. In other words, no gap or insertion penalties were accrued for either the designed or natural sequences when calculating profile scores. This was based on the philosophy that the designed sequence should not be unduly penalized or rewarded by virtue of the number of insertions or deletions contained in its template structure. As a control, we also collected profile statistics with sequence weighting incorporated to correct for biased sampling in the aligned sequences (Henikoff & Henikoff, 1994). The results were not significantly different from those reported here, presumably because of the large number of sequences included in each alignment.

HP models for random sequence generation

Hydrophobic-polar models for each structural template were constructed using a procedure similar to that described by Micheletti et al. (1998), with a cutoff value for the contact score of 5.5. Any generic side-chain centroid with a contact score higher than 5.5 was assigned an H. Lower scoring centroids were assigned a P. Subsets of amino acids were defined as $H = \{A,F,G,I,L,P,V,W,Y\}$ and $P = \{A,D,E,K,N,Q,R,S,T\}$.

Acknowledgments

This work was supported by an NSF CAREER Award No. CHE 9876234.

References

Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. 2000. The Pfam protein families database. *Nucleic Acids Res 28*:263–266.

Bowie JU, Reidhaar-Olson JF, Lim WA, Sauer RT. 1990. Deciphering the message in protein sequences: Tolerance to amino acid substitutions. *Science 247*:1306–1310.

Clarke ND, Kissinger CR, Desjarlais J, Gilliland GL, Pabo CO. 1994. Structural studies of the engrailed homeodomain. *Protein Sci 3*:1779–1787.

Dahiyat BI, Gordon DB, Mayo SL. 1997a. Automated design of the surface positions of protein helices. *Protein Sci 6*:1333–1337.

Dahiyat BI, Mayo SL. 1997a. De novo protein design: Fully automated sequence selection. *Science 278*:82–87.

Dahiyat BI, Mayo SL. 1997b. Probing the role of packing specificity in protein design. *Proc Natl Acad Sci USA 94*:10172–10177.

Dahiyat BI, Mayo SL. 1996. Protein design automation. *Protein Sci 5*:895–903.

Dahiyat BI, Sarisky CA, Mayo SL. 1997b. De novo protein design: Towards fully automated sequence selection. *J Mol Biol 273*:789–796.

Desjarlais JR, Clarke ND. 1998. Computer search algorithms in protein modification and design. *Curr Opin Struct Biol 8*:471–475.

Desjarlais JR, Handel TM. 1995. De novo design of the hydrophobic cores of proteins. *Protein Sci 4*:2006–2018.

Desjarlais JR, Handel TM. 1999. Side-chain and backbone flexibility in protein core design. *J Mol Biol 290*:305–318.

Desmet J, De Maeyer M, Hazes B, Lasters I. 1992. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature 356*:539–542.

Dunbrack RL Jr, Cohen FE. 1997. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci 6*:1661–1681.

Durbin R, Eddy S, Krogh A, Mitchison G. 1998. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge, UK: Cambridge University Press.

Eisenberg D, McLachlan AD. 1986. Solvation energy in protein folding and binding. *Nature 319*:199–203.

Fraenkel E, Rould MA, Chambers KA, Pabo CO. 1998. Engrailed homeodomain-DNA complex at 2.2 A resolution: A detailed view of the interface and comparison with other engrailed structures. *J Mol Biol 284*:351–361.

Goldstein RF. 1994. Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys J 66*:1335–1340.

Gordon DB, Marshall SA, Mayo SL. 1999. Energy functions for protein design. *Curr Opin Struct Biol 9*:509–513.

Gribskov M, McLachlan AD, Eisenberg D. 1987. Profile analysis: Detection of distantly related proteins. *Proc Natl Acad Sci USA 84*:4355–4338.

Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. 1998. High-resolution protein design with backbone freedom. *Science 282*:1462–1467.

Harbury PB, Tidor B, Kim PS. 1995. Repacking protein cores with backbone freedom: Structure prediction for coiled coils. *Proc Natl Acad Sci USA 92*:8408–8412.

Hecht MH. 1994. De novo design of beta-sheet proteins. *Proc Natl Acad Sci USA 91*:8729–8730.

Hellinga HW. 1997. Rational protein design: Combining theory and experiment. *Proc Natl Acad Sci USA 94*:10015–10017.

Hellinga HW, Richards FM. 1994. Optimal sequence selection in proteins of known structure by simulated evolution. *Proc Natl Acad Sci USA 91*:5803–5807.

Hendsch ZS, Tidor B. 1999. Electrostatic interactions in the GCN4 leucine zipper: Substantial contributions arise from intramolecular interactions enhanced on binding. *Protein Sci 8*:1381–1392.

Henikoff S, Henikoff JG. 1994. Position-based sequence weights. *J Mol Biol 243*:574–578.

Holland JH. 1992. *Adaptation in natural and artificial systems*. Cambridge, Massachusetts: The MIT Press.

Johnson EC, Lazar GA, Desjarlais JR, Handel TM. 1999. Solution structure and dynamics of a designed hydrophobic core variant of ubiquitin. *Struct Fold Des 7*:967–976.

Jorgensen WL, Tirado-Rives J. 1988. The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *J Am Chem Soc 110*:1657–1666.

Juffer AH, Eisenhaber F, Hubbard SJ, Walther D, Argos P. 1995. Comparison of atomic solvation parametric sets: Applicability and limitations in protein folding and binding. *Protein Sci 4*:2499–2509. [Also Erratum. 1996. *Protein Sci 5*(8):1748–1749.]

Kissinger CR, Liu BS, Martin-Blanco E, Kornberg TB, Pabo CO. 1990. Crystal structure of an engrailed homeodomain–DNA complex at 2.8 A resolution: A framework for understanding homeodomain–DNA interactions. *Cell 63*:579–590.

Koehl P, Levitt M. 1999a. De novo protein design. I. In search of stability and specificity. *J Mol Biol 293*:1161–1181.

Koehl P, Levitt M. 1999b. De novo protein design. II. Plasticity in sequence space. *J Mol Biol 293*:1183–1193.

Kono H, Doi J. 1994. Energy minimization method using automata network for sequence and side-chain conformation prediction from given backbone geometry. *Proteins 19*:244–255.

Kono H, Nishiyama M, Tanokura M, Doi J. 1998. Designing the hydrophobic core of *Thermus flavus* malate dehydrogenase based on side-chain packing. *Protein Eng 11*:47–52.

Kraulis PJ. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J Appl Crystallogr 24*:946–950.

Lazar GA, Desjarlais JR, Handel TM. 1997. De novo design of the hydrophobic core of ubiquitin. *Protein Sci 6*:1167–1178.

Lazar GA, Johnson EC, Desjarlais JR, Handel TM. 1999. Rotamer strain as a determinant of protein structural specificity. *Protein Sci 8*:2598–2610.

Leahy DJ, Hendrickson WA, Aukhil I, Erickson HP. 1992. Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. *Science 258*:987–991.

Lim WA, Hodel A, Sauer RT, Richards FM. 1994. The crystal structure of a mutant protein with altered but improved hydrophobic core packing. *Proc Natl Acad Sci USA 91*:423–427.

Micheletti C, Seno F, Maritan A, Banavar JR. 1998. Design of proteins with hydrophobic and polar amino acids. *Proteins 32*:80–87.

Musacchio A, Noble M, Pauptit R, Wierenga R, Saraste M. 1992. Crystal structure of a Src-homology 3 (SH3) domain. *Nature 359*:851–855.

Oubridge C, Ito N, Evans PR, Teo CH, Nagai K. 1994. Crystal structure at 1.92 A resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature 372*:432–438.

Street AG, Mayo SL. 1999. Computational protein design. *Struct Fold Des 7*:R105–R109.

Su A, Mayo SL. 1997. Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci 6*:1701–1707.

Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta S, Weiner P. 1984. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J Am Chem Soc 106*:765–784.