PROTEINS: Structure, Function, and Genetics 52:225–235 (2003)

Conserved Residue Clustering and Protein Structure Prediction

Ora Schueler-Furman[1] and David Baker[1,2]*
[1]*Department of Biochemistry, University of Washington, Seattle, Washington*
[2]*Howard Hughes Medical Institute, University of Washington, Seattle, Washington*

ABSTRACT **Protein residues that are critical for structure and function are expected to be conserved throughout evolution. Here, we investigate the extent to which these conserved residues are clustered in three-dimensional protein structures. In 92% of the proteins in a data set of 79 proteins, the most conserved positions in multiple sequence alignments are significantly more clustered than randomly selected sets of positions. The comparison to random subsets is not necessarily appropriate, however, because the signal could be the result of differences in the amino acid composition of sets of conserved residues compared to random subsets (hydrophobic residues tend to be close together in the protein core), or differences in sequence separation of the residues in the different sets. In order to overcome these limits, we compare the degree of clustering of the conserved positions on the native structure and on alternative conformations generated by the de novo structure prediction method Rosetta. For 65% of the 79 proteins, the conserved residues are significantly more clustered in the native structure than in the alternative conformations, indicating that the clustering of conserved residues in protein structures goes beyond that expected purely from sequence locality and composition effects. The differences in the spatial distribution of conserved residues can be utilized in de novo protein structure prediction: We find that for 79% of the proteins, selection of the Rosetta generated conformations with the greatest clustering of the conserved residues significantly enriches the fraction of close-to-native structures. Proteins 2003;52:225–235.**

Key words: protein residue conservation; bioinformatics; protein structure prediction; clustering; Rosetta

INTRODUCTION

To what extent do conserved residues in a protein family cluster together in three dimensions? Residues are likely to be conserved in a protein family because they either make critical stabilizing interactions or play important functional roles. Evolutionary pressure for both stability and function could lead to clustering of conserved residues: Residues important for stability are often close together in the hydrophobic core, and functional residues may be close together in enzyme-active sites or protein–protein or protein–ligand binding sites. Mutational experiments indicate that hydrophobic core residues make substantial contributions to stability[1]; in contrast, only some "hot spot" residues in protein–protein interfaces contribute significantly to binding between proteins.[2–4]

The spatial proximity of conserved residues has been analyzed at various levels. Fully conserved residues make more contacts than nonconserved residues,[5] because many of them are located in the hydrophobic core. In addition to fully conserved positions, correlated changes in two residues can hint at conservation at the *pair* level, implying possible physical proximity.[6–8] Finally, positions that are conserved only in specific subfamilies of an alignment may play more family-specific, functional roles and can be clustered in functional patches.[9–11] Functional sites have been identified by searching for patches of conserved residues on the surface of a protein.[11–16] Overall, there is a clear but somewhat weak correlation between clustering of conserved positions and known functional sites, involving mostly conserved polar surface residues.[13,17,18]

The significance of clustering of conserved residues in a protein structure can be assessed by comparison to the clustering of randomly selected residues[19] or of highly variable residues.[20] If, however, a conserved subset contains a substantial number of hydrophobic core residues, or a number of consecutive residues that form a local sequence motif, these will be more clustered than randomly derived subsets that are dispersed among the whole protein structure. In order to eliminate confounding effects of sequence locality and composition, the clustering of conserved residues on a protein structure can be compared to their clustering on alternative protein-like conformations for the sequences. These conformations can be generated in different ways. One possibility is to thread[21–24] the sequence of the protein onto a representative set of known structures. By threading a set of protein sequences each on a pair of structures, namely, their native structure and an

Grant sponsor: Howard Hughes Medical Institute; Grant sponsor: Damon-Runyon Fellowship from the Damon Runyon Cancer Research Foundation accorded to OS-F; Grant number: DRG-1704-02.

*Correspondence to: David Baker, Department of Biochemistry and Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195. E-mail: dabaker@u.washington.edu

Received 14 July 2002; Accepted 25 November 2002

alternative, incorrect structure of same length, Olmea et al.[20] showed that conserved and correlated residues tend to be clustered significantly more in the native structures than in the incorrect models. Alternatively, de novo structure prediction methods can be used to create a large set of structural models (decoys) for the protein. An important difference between the two sets is that models generated by threading a sequence on an incorrect structure will not generally have buried hydrophobic cores, whereas models produced by de novo structure prediction will have mostly buried hydrophobic residues.

If conserved residues are indeed spatially clustered in space in native protein structures compared to non-native structures created by de novo structure prediction methods, this could be used to select native-like structural models. Characterization of the distribution of conserved residues could thus help to resolve the ambiguities associated with current de novo structure prediction methods. Although there have been significant improvements in the generation of plausible models of the structure for an amino acid sequence over the last few years,[25] selection of the best model from a set of possibilities is a largely unsolved problem because of limitations in current potential functions and sampling methods. It seems plausible that assessment of the clustering of conserved residues in different models could help to resolve some of the ambiguities, because it should be quite orthogonal to the features currently used for model selection that are primarily based on physical energy functions.[26] Indeed, inclusion of tertiary information in form of specific distance constraints can improve the modeling of protein structure both by threading[20,27] and by de novo prediction.[28–31] The study by Olmea et al.[20] shows that such information can enhance both the selection of the correct fold and the alignment quality in threading calculations.

In this study, we investigate the extent to which evolutionary conserved residues are clustered, and whether such potential clustering can be used to select good structural models from a set of decoys created by Rosetta,[32] a relatively successful current ab initio structure prediction procedure. We first analyze a large set of proteins for the degree of clustering of conserved residues and show that they are significantly more clustered than both randomly selected subsets on the same structure and the same subset on a range of different protein decoys. Furthermore, we find that close-to-native conformations can be preferentially selected from sets of alternative conformations based on the extent of clustering of conserved residues.

MATERIALS AND METHODS

Set of Protein Structures

The set of proteins used in this study was compiled from two different sources: from a set of representative proteins, described by Eyrich et al.,[33] that contains primarily small proteins and protein domains, and from the Culled pdb (http://www.fccc.edu/research/labs/dunbrack/pisces/culledpdb.html),[34] a nonredundant set of protein chains with less than 25% sequence identity from the Protein Data Bank (PDB).[35] For each protein, 2000 decoys were created with Rosetta.[36] The final set contains 79 proteins with at least one decoy within 6.0 Å root-mean-square deviation (RMSD) from the native structure, and with a sufficiently deep enough multiple sequence alignment (see below): 1a32, 1a68, 1aa3, 1abo*A*, 1aca, 1acp, 1adr, 1afi, 1aho, 1ap0, 1ark, 1b3a*A*, 1b67*A*, 1bdo, 1bkr*A*, 1bor, 1bq9, 1c5a, 1c9o*A*, 1cc5, 1cc8*A*, 1ccw*A*, 1coo, 1cse*I*, 1csp, 1ctf, 1ctj, 1cyo, 1dol, 1e6i*A*, 1edm*B*, 1ejg*A*, 1elk*A*, 1elw*A*, 1eyv*A*, 1f7l*A*, 1fip*A*, 1fjl*A*, 1fjs*L*, 1fm0*D*, 1fna, 1fqt*A*, 1g6x*A*, 1h4x*A*, 1h75*A*, 1h97*A*, 1hyp, 1icf*I*, 1jbe*A*, 1kjs, 1lkk*A*, 1mzm, 1opd, 1psr*A*, 1ptq, 1qyp, 1r69, 1rb9, 1scj*B*, 1sgp*I*, 1sro, 1stu, 1svy, 1tif, 1tuc, 1ubi, 1vig, 2af8, 2cdx, 2fdn, 2fow, 2gdm, 2pdd, 2trx*A*, 2u1a, 3ebx, 4ubp*A,* 5icb, and 5pti (PDB code with specific chain in *italics*).

Creation of Multiple Sequence Alignments

For each query protein, a set of homologous sequences was collected by an iterative PSI-BLAST search.[37] Based on the output, a multiple sequence alignment was created that includes sequences with less than 90% sequence identity to any other sequence and that span more than 80% of the query sequence. Three different stringency levels were used for the PSI-BLAST runs: (1) *level10*: 10 rounds of PSI-BLAST with an acceptance threshold of $10E^{-10}$; (2) *level7*: 5 rounds with an acceptance threshold of $10E^{-7}$; and (3) *level5*: 5 rounds with an acceptance threshold of $10E^{-5}$. The first level that resulted in a deep-enough multiple sequence alignment (defined as including more than 24 sequences) was retained for further analysis.

Measure of Degree of Conservation

The degree of conservation of a position is defined as its information content (IC),[38]

$$IC_k = \sum_{i=1.20} p_{ik} \ln p_{ik},$$

where p_{ik} is the frequency of amino acid i at position k.

In addition to simply using unit weights for different sequences, we also experienced with information content derived in more sophisticated ways, such as from position-specific weight matrices (PSSM[39]) implemented in PSI-BLAST.[37] Using background amino acid frequencies, weighted sequence profiles can be reconstructed from the PSI-BLAST PSSM matrix and used to calculate the absolute information content. For simplicity, however, in the calculations reported here we used unit weights, since both approaches yielded similar results (see Results section).

Definition of Subset of Evolutionary Conserved Residues

A subset S was defined by the most conserved residues within a protein, selected based on their IC values. Several parameters were varied and evaluated for performance: (1) Subset *size*: 5 (c_5), 10 (c_{10}), and 15 (c_{15}); (2) Subset *amino acids*: Positions for which >50% of the sequences contain a certain type of amino acid were excluded: no exclusion (all); single exclusion: glycine (−G), proline (-P), alanine (-A), valine (-V), leucine (-L), isoleucine (-I), cys-

teine (-C); exclusion of combinations of amino acids: nonhydrophobic subset (-VLIMF) and nonpolar subset (-DEQNRKST). In cases where several positions showed the same IC values, all were included, resulting in increased subset size.

Measure of Degree of Clustering of Subset (M_S)

The degree of clustering of a subset s was measured as the average inverse distance between all pairs of positions in the subset:

$$M_s = \langle 1/r \rangle = 1/N_{\text{pairs}} \sum_{i=1,}^{Ns-1} \sum_{j=i+1,}^{Ns} (1/r_{ij}), \quad (1)$$

where N_s is the size of the subset; N_{pairs} is the number of different position pairs in the subset: $N_{pairs} = (N_s - 1)N_S/2$; and r_{ij} is the distance between the average sidechain centroid of residues i and j.[32] A high value for M_S indicates that most residues in the subset are near to each other, clustered within one or a few clusters. Single outlier positions should not influence the value of M_s significantly.

Assessment of Significance of Clustering of Evolutionary Conserved Residues

Comparison to random subsets of residues

The degree of clustering of conserved residues was compared to values obtained from 2000 random, same-size subsets from the same structure. We evaluated for how many proteins the extent of clustering of the conserved residues was in the top 5% of the cluster scores (M_S) of the random set. In addition, the clustering was compared to all possible sets of the same connectivity. These sets were created by shifting the selected subset along the protein sequence (circular permutation).

Comparison to the conserved subset in decoys

The degree of clustering of conserved residues was determined by comparing the M_S value in the experimentally determined structure to values obtained for the same subset in 2000 decoys. We evaluated the number of proteins for which the rank of the native structure was within the first five percentiles of all the decoy structures. Decoys with less than 4 Å RMSD from the native structure were not included in the calculation.

Use of Clustering Measure in Structure Prediction

M_S values were used to rank all of the decoys. To assess whether greater clustering of conserved residues occurred in more native-like decoys, the enrichment of low RMSD decoys by score was computed:

$$E_x = \frac{\%\ \textit{of low RMSD decoys}\ \text{within the set of}\ \textit{high score decoys}}{\%\ \text{of}\ \textit{low RMSD decoys}\ \text{within the set of all decoys}} \quad (2)$$

where *high score decoys* and *low RMSD decoys* are defined as those decoys that belong to the subset of decoys with the *X*% best score and *X*% lowest RMSD values, respectively. We used a cutoff of 15% ($p_x = 0.15$). Thus, if we obtain m low-RMSD decoys in a draw of N total structures from a population with a fraction of p_x low-RMSD decoys, the enrichment would be:

$$E_x = (m/N)/p_x.$$

If the N structures are drawn randomly, the probability distribution for m is the standard binomial distribution

$$\text{Prob}(m \mid N,p_x) = N!/[m!(N-m)!] \cdot p_x^m (1-p_x)^{N-m},$$

and the probability of an enrichment of E'_x or better is

$$p_{\text{val}} = \text{Prob}(E_x \geq E'_x) = \sum_{m=m'}^{N} \text{Prob}(m \mid N,p_x), \quad (3)$$

where

$$m = p_x N E_x.$$

The more successful a scoring scheme, the larger the enrichment. To evaluate a scoring method we count the number of decoy sets with significant enrichment (e.g. $p_{\text{val}} \leq 0.05$).

Evaluation of Clustering of Conserved Positions on the Protein Surface and in the Protein Core

For the definition of the conserved subset, we included either all positions or only positions at the protein surface/core, determined by excluding 10%, 30%, or 50% of the most buried/exposed residues in the protein structure. The degree of burial of a certain residue was based on the number of residue centroids within 6.0 Å of its own average centroid.

RESULTS

Are Conserved Residues Clustered in Proteins?

The spatial clustering of conserved residues in protein structures is investigated in several steps. After showing the clustering of conserved residues in a specific example, we proceed to a general assessment on a set of proteins. The actual clustering is compared to alternative subsets of residues on the same structure, as well as to the same subset of residues on alternative structures.

The degree of spatial clustering of conserved residues has been traditionally assessed by clustering the conserved residues and recording the cluster size and the number of clusters. Although well suited for patch identification, this is not optimal for overall assessment of the degree of clustering, because how to combine the two numbers in a single measure is not obvious, and only a small range of integer values can be obtained. Here, we assess the degree of spatial clustering of conserved residues by a measure that combines the features of the traditional measures: $M_S = \langle 1/r \rangle$—the average inverse distance between conserved residue pairs. It is primarily influenced by residues close in space and not sensitive to single outliers, and the selection of well-clustered sets of residues on a protein is facilitated by the continuous distribution of values that are obtained from alternative residue subsets. M_S is similar to the measure of the number of contacting pairs of conserved or correlated residues,[20] but is continuous and not dependent on a contact definition, and is thus well suited for ranking.

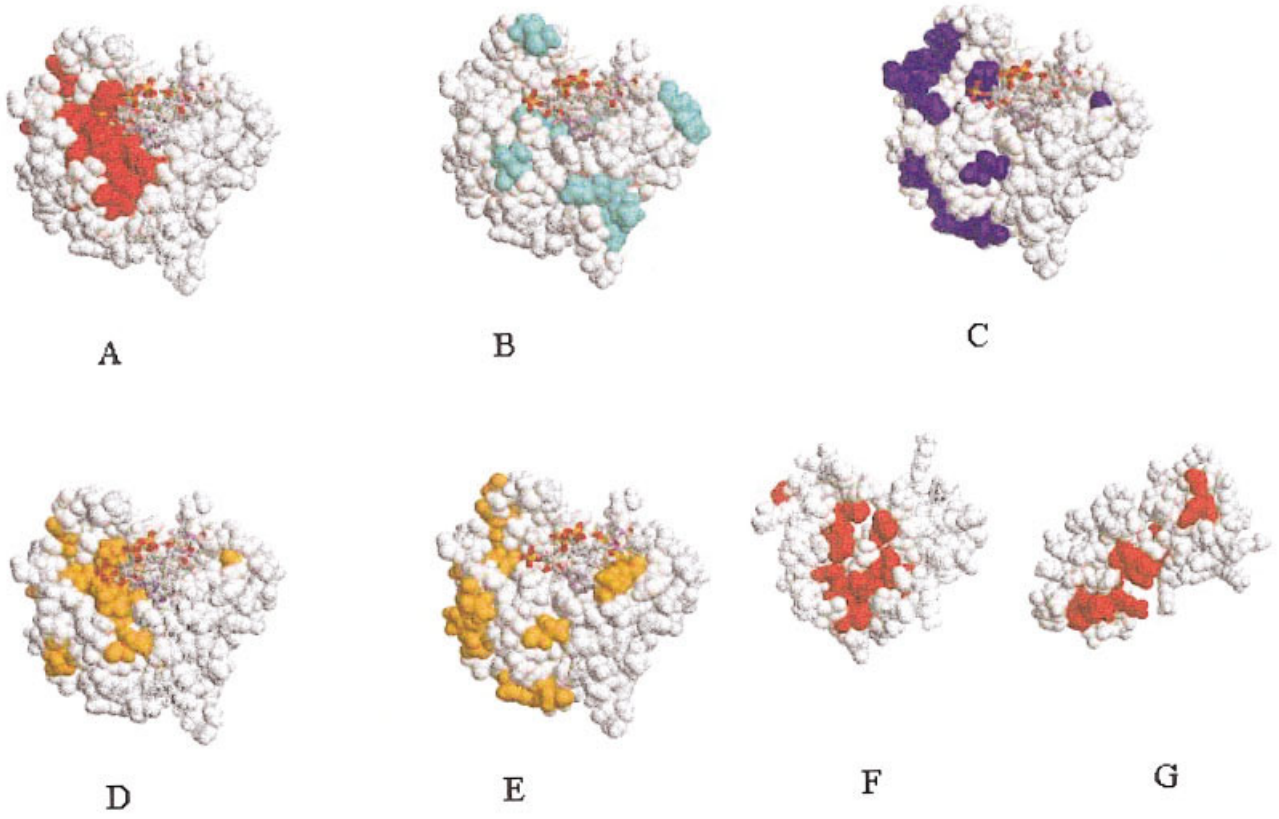

Fig. 1. Example of clustering of evolutionary conserved residues in proximity of the ligand, in acyl-coenzyme A binding protein (ACBP, PDB code 1aca). The selected subset is colored, and the ligand is shown as a stick model. (**A**) Subset of conserved residues, selected according to parameters c_{15} -G (subset size 15, excluding conserved glycines). (**B**) Example for a random subset of same size. (**C**) Shifted subset ($\delta = 10$). (**D**) Subset of conserved *surface* residues: 10% most buried residues not considered. (**E**) As in (D), but 30% most buried residues excluded. (**F, G**) Subset of conserved residues highlighted on two decoys with 3.6 Å (**F**) and 15 Å RMSD (**G**) to 1aca, respectively. Figure 1 was created using the program RASMOL.[43]

A conservation score for a residue in a protein sequence can be derived from multiple-sequence alignments in various ways (see review[40]). The information content of a position can be calculated based on Shannon's entropy,[38] provided the different sequences in the alignment are properly weighted. Here, we use a simple measure to approximate the degree of conservation of a residue by calculating the information content from a multiple-sequence alignment of sequences with less than 90% identity that are given unit weights. As an alternative to this scheme, the information content can also be derived from position-specific weight matrices (PSSM[39]), such as those created in a PSI-BLAST run.[37] Comparison of the conserved subsets created by these two approaches showed that they mostly contain the same residues (on average, 11 out of 15). Furthermore, inspection of a number of proteins revealed that the performance was similar for both subsets, although sometimes the PSSM-derived subset performance was slightly weaker. We therefore chose to use unit weights. For the detection of remote homologs by a profile, however, appropriate sequence weighting is important for the creation of a successful PSSM.[37]

We have varied the definition of the conserved residue subset to identify the combination of parameters that yields the most pronounced clustering. Different subset sizes, as well as exclusion of different types of conserved amino acid residues, have been assessed for their influence on the performance. This provides information about the individual contribution of different amino acid types to clustering, and indicates how best to measure clustering for protein structure prediction.

An Example: Acyl-Coenzyme A Binding Protein (ACBP; PDB code 1aca[41])

In a ligand-binding protein, we would expect that the residues involved in binding the ligand are conserved and clustered in vicinity of the ligand. Figure 1(A) shows that the conserved residues in ACBP form a patch next to the ligand acyl-coenzyme A. How can the significance of such a clustering be assessed, and how general is this phenomenon?

Subsets of Evolutionary Conserved Positions Are Significantly More Clustered Than Randomly Selected Same-Size Subsets

The significance of clustering of evolutionary conserved residues in space can be assessed by comparison to the clustering of randomly selected, same-size subsets on the same protein. For ACBP, random sets of residues are less clustered [Fig. 1(B)]. Analysis of a set of proteins shows that for most proteins the clustering of the conserved

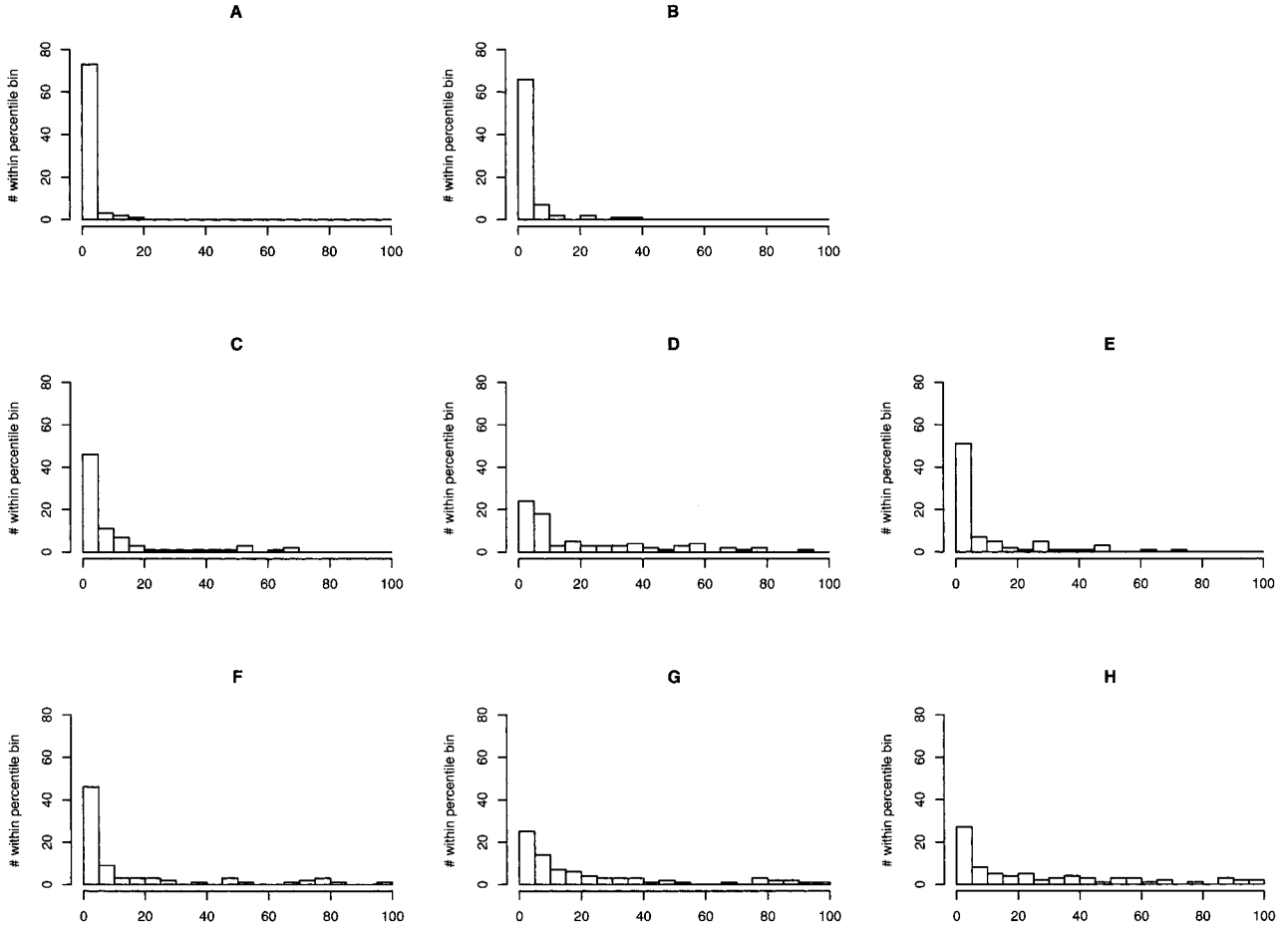

Fig. 2. Conserved residues tend to be clustered in space. For the 79 proteins, the evolutionary conserved subset was ranked among alternative, same-size subsets based on M_S (eq. 1). Histograms of the rank (in percentiles) of the conserved subset are given for different comparisons. (**A–E**): Comparison to 2000 alternative sets of residues on the native structure: (**A**) Random subsets. (**B**) Subsets with same connectivity, generated by sliding the conserved subset over the protein sequence. (**C**) Random subsets, excluding the 30% most buried positions. (**D**) Random subsets, excluding the 50% most buried positions. (**E**) Random subsets, excluding the 50% most exposed positions. (**F–H**) Comparison of the same sequence subset on 2000 alternative decoy structures. (**F**) Subset selected from all residues. (**G**) 30% most buried residues excluded from the subset. (**H**) 50% most exposed residues excluded from the subset.

subset is ranked very high, within the first 5% of random subsets [Fig. 2(A)]. The most significant signal is obtained by choosing a conserved subset of 15 residues (c_{15}; see Methods section). From an analysis of subsets created by excluding single types of conserved amino acids, it appears that exclusion of conserved glycines (c_{15} -G) results in best discrimination of the conserved subset from random sets. This can be explained by the known role of conserved glycines in turn formation that may locate them far from the protein core and conserved functional sites. Combination c_{15} -G ranks 92% of the proteins within the first 5% of random subsets. Similar results were obtained with the exclusion of other single amino acid types. Table I(a) shows the performance of some selected combinations. The worst performance is obtained when conserved leucines, valines, or alanines are excluded from the subset, consistent with their abundance in the protein core. The results obtained for nonhydrophobic and nonpolar subsets demonstrate that conserved hydrophobic amino acids contribute more to the clustering of the subset than do conserved polar amino acids. Exclusion of the latter increases the signal to 96%.

The Clustering of Conserved Positions Is Not Due to Sequence Connectivity Effects

Conserved positions tend to form motifs that are local in sequence, and will therefore be located near each other in space. This could result in more significant clustering compared to randomly selected subsets. In order to assess whether this is the case, we compared the degree of clustering of the conserved residues to subsets with the same sequence connectivity. In ACBP, sliding the subset of conserved residues by 10 residue positions significantly reduces clustering [Fig. 1(C)]. Different subsets were created on a large scale by sliding the original subset of residues on the sequence with all possible shifts ($\delta_1, \ldots \delta_{n-1}$; n = length of protein), and their degree of clustering was evaluated. Figure 2(B) shows that the ranks obtained from this comparison are only slightly lower than the ranks compared to random subsets. This indicates that

TABLE I. Extent of Spatial Clustering of Different Residue Subsets[†]

(a) Comparison of conserved subset to alternative subsets on native structure	Random subsets	Shifted subsets	Random subsets on protein surface[a]	Random subsets in protein core[b]
All amino acids	89%	84%	56%	65%
Best condition for single exclusion	−G: 92%	−P: 85%	−P: 61%	−P: 68%
Worst condition for single exclusion	−A,−L,−V: 86%	−L,−V,−I: 81%	−A: 48%	−L: 39%
Nonhydrophobic[c]	61%[e]	53%	42%	32%
Nonpolar[d]	96%	92%	43%	71%

(b) Comparison of conserved subset on native to alternative structures	Whole protein	Protein surface	Protein core
All amino acids	53%	38%	35%
Best condition for single exclusion	−C: 65%	−L: 42%	−C: 37%
Worst condition for single exclusion	−A: 52%	−I: 32%	−A: 23%
Nonhydrophobic	49%	30%	17%
Nonpolar	51%	28%	32%

[†]The percentage of proteins with M_S values within the top 5% of a distribution of M_S values in alternative subsets is given for size c_{15}.
[a]30% most buried residues excluded
[b]50% most exposed residues excluded
[c]no phenylalanine, isoleucine, leucine, methionine and valine
[d]no aspartate, asparagine, glutamate, glutamine, arginine, lysine, serine and threonine
[e]underlined numbers: subset size c_{10}

clustering in space is not simply a consequence of clustering in sequence. Results for specific combinations of parameters for the creation of the subset of conserved residues (i.e., variation of subset size and exclusion of different conserved amino acid types from the subset) show similar trends, as in the comparison to random subsets [Table I(a)]. The best combination here (c_{15}-P) ranks 85% of the proteins within the first 5% of randomly shifted subsets. As for glycine, proline also is frequently conserved in turns and helix ends. Moreover, again, the contributions of hydrophobic amino acids (leucine, isoleucine, and valine) are important. Finally, again, more proteins are significantly ranked with the use of the nonpolar subset (e.g., 92% compared to 85%).

Clustering of Evolutionary Conserved Positions Is Mostly, but Not Solely Determined by Buried Core Residues

Often the conserved residues are located in the protein core and are therefore more clustered than randomly selected positions. In order to evaluate whether the degree of clustering is also significant for conserved *surface* residues, we repeated the same analysis for exposed protein residues only. These were defined by excluding 10%, 30%, or 50% of the most buried positions (see Methods section for definition of burial). The exposed conserved residues were then compared to randomly selected same-size subsets of exposed residues. Although excluding only 10% of the most buried residues did not change the number of significantly ranked proteins, excluding more buried residues resulted in a significant drop in performance [Fig. 2(C and D) and Table I(a)]. For example, (c_{15} -P) ranked 90%, 85%, 61%, and 35%, respectively, of the proteins within the first five percentiles when 0%, 10%, 30%, and 50% of the most buried positions are excluded. Thus, although evolutionary conserved subsets are significantly more clustered than randomly selected subsets when the whole protein is considered, the same analysis on the protein surface shows only a smaller degree of clustering of evolutionary conserved surface subsets compared to random surface subsets. This indicates that the buried core residues make an important contribution to the clustering. Combined with the finding that conserved hydrophobic residues are important for obtaining significantly clustered subsets, whereas exclusion of conserved polar residues from the subset increase the signal, this suggests that hydrophobic core positions are the main contributors to the signal of clustering of conserved residues. As expected, exclusion of conserved polar residues from exposed subsets reduces the extent of clustering, indicating that those are important on the protein surface.

For comparison, we assessed the degree of clustering of evolutionary conserved *core* positions by excluding 50% of the most exposed residues [Fig. 2(E)]. Comparison of the evolutionary conserved subset of core residues to same-size random core subsets results in better performance than that observed for surface residues, but worse performance than for the whole protein [c_{15} -P: 68% compared to 35% and 90%, respectively; compare columns in Table I(a)]. Thus, the degree of clustering of evolutionary conserved residues is more extensive for core positions than for surface positions, but less than for all positions together, when measured against randomly derived, same-size subsets of core, surface, and overall positions, respectively. As expected for hydrophobic cores, the contribution of hydrophobic residues is important (e.g., decreased signal for the nonhydrophobic subset), and exclusion of polar residues improves the signal (increased signal for the nonpolar subset). The general picture emerges that both

core positions (mostly hydrophobic) and surface positions (mostly polar) are important for the signal observed.

Exclusion of more residues, either by considering only buried or surface residues, or by excluding amino acid types, leads to a smaller range of positions from which the subset is selected. For some proteins, this leads to the selection of positions that do not show a high degree of evolutionary conservation and, as a consequence, show a smaller degree of clustering, resulting in a less significant signal when compared to random subsets. This could be one reason that a more stringent definition of surface residues by excluding the 50% most buried residues (instead of 30%) significantly reduces the clustering signal. In such cases, smaller subsets, such as the combination (c_{10} -L), give the best result for distinguishing conserved surface subsets from randomly derived surface subsets.

For ACBP, the subset of exposed conserved residues is much less clustered than the conserved residues taken from the whole protein. Buried, conserved positions that connect conserved surface positions are no longer selected; instead, remote positions are selected, resulting in a noncontiguous patch [Fig. 1(D and E)]. This results in less significant ranking among random subsets than what was observed for subsets derived from all positions.

Subsets of Evolutionary Conserved Positions Are Significantly More Clustered in Experimentally Derived Structures Than in Decoys

These results show that conserved residues are significantly more clustered in protein structures than randomly chosen residues in the same structure. To what extent does this derive from the nearly ubiquitous clustering of conserved hydrophobic residues in protein cores? To address this issue, we need to generate structures with different core-packing arrangements. This can be done using the Rosetta de novo structure prediction method. This controls for both connectivity and amino acid composition: The sequence connectivity and composition of the conserved subset on Rosetta-generated conformations is identical to that of native structures, and the hydrophobic residues tend to be buried.

Indeed, the conserved residues are less clustered on decoy structures of ACBP than on the experimentally derived structure 1aca [Fig. 1(F and G)]. In general, the rank of the native structure compared to decoy structures is high [Table I(b)]: For the best combination (c_{15} -C) for 65% of the proteins, the native structure ranks within the first 5% in the decoy population. This signal is somewhat less significant than that for the comparison to random subsets [Fig. 2 (cf. E and A)]. In contrast to the comparison of the conserved subset to random subsets on the native structure, for the comparison to decoys, exclusion of conserved cysteines results in improved ranking. This is probably a result of the tendency of Rosetta to bring cysteines together more than observed in native structures, which will increase their weight inappropriately in the clustering measure M_S. The nonhydrophobic and nonpolar subsets both resulted in weaker signals, indicating that for distinction of the native structure from decoys based on degree of clustering, both hydrophobic and polar conserved residues are important. The dominant contribution of hydrophobic residues to the spatial clustering of conserved residues compared to random subsets reflects in part differences in amino acid composition (the randomization will exchange nonpolar buried residues with polar surface residues). In the comparisons to alternative conformations, the amino acid composition of the subset is fixed, and the contribution of the polar residues to the clustering is more evident.

As for the comparison of native to random subsets, the clustering of evolutionary conserved positions on native structures compared to decoys is less pronounced when only exposed or only buried residues are considered [Table I(b), Fig. 2 (cf. F, G, and H, and A, C, and E)].

In summary, a consistent picture of clustered, conserved residues emerges: Evolutionary conserved residues are not only significantly more clustered on the native structure than randomly selected sets on the same structure, but also when compared to the same sets on decoy structures. In both comparisons, the most significant signal is obtained for a large subset (e.g., c_{15}) and when considering the whole protein. Hydrophobic conserved residues contribute to the signal on the whole protein, as well as on the protein core, whereas polar conserved residues contribute to the signal on the surface of the protein. In the comparison to decoys, the polar residues also contribute to the signal on the whole protein.

These results suggest that assessment of the clustering of conserved residues could be used for recognition of low-RMSD decoys.

Can Low-RMSD Models Be Selected Based on Clustering of Conserved Residues? *Low-RMSD Decoys of ACBP Have Good M_S* Scores

The clustering of evolutionary conserved residues on an example of a low- and high-RMSD decoy of ACBP is shown in Figure 1(F). The degree of clustering is much more pronounced for the better model. A plot of the degree of clustering of conserved residues measured by M_S versus the RMSD of decoys to the native structure of ACBP (1aca) shows a negative correlation between the two [Fig. 3(A)]. Therefore, M_S can be used to select good decoys. Indeed, the set of decoys with *high scores* (defined as the 15% best-scoring decoys; see Methods section) is enriched 2.6 times with *low-RMSD decoys* (defined as the 15% lowest RMSD decoys). This means that 40% (2.6* 15%) of the decoys with *high scores* are *low-RMSD decoys*.

Because native structures tend to be more compact than modeled decoys, a simple screening based on the radius of gyration of a decoy could be applied as well. M_S describes also features that are independent of compactness, as demonstrated by scoring only a subset of compact structures. The enrichment obtained by scoring among a subset of the decoys of ACBP with the 50% lowest R_{gyr} is still significant [E = 2.16; see Fig. 3(B)]. However, particularly for elongated native structures with high R_{gyr}, the selection of decoys based on the clustered conserved residues

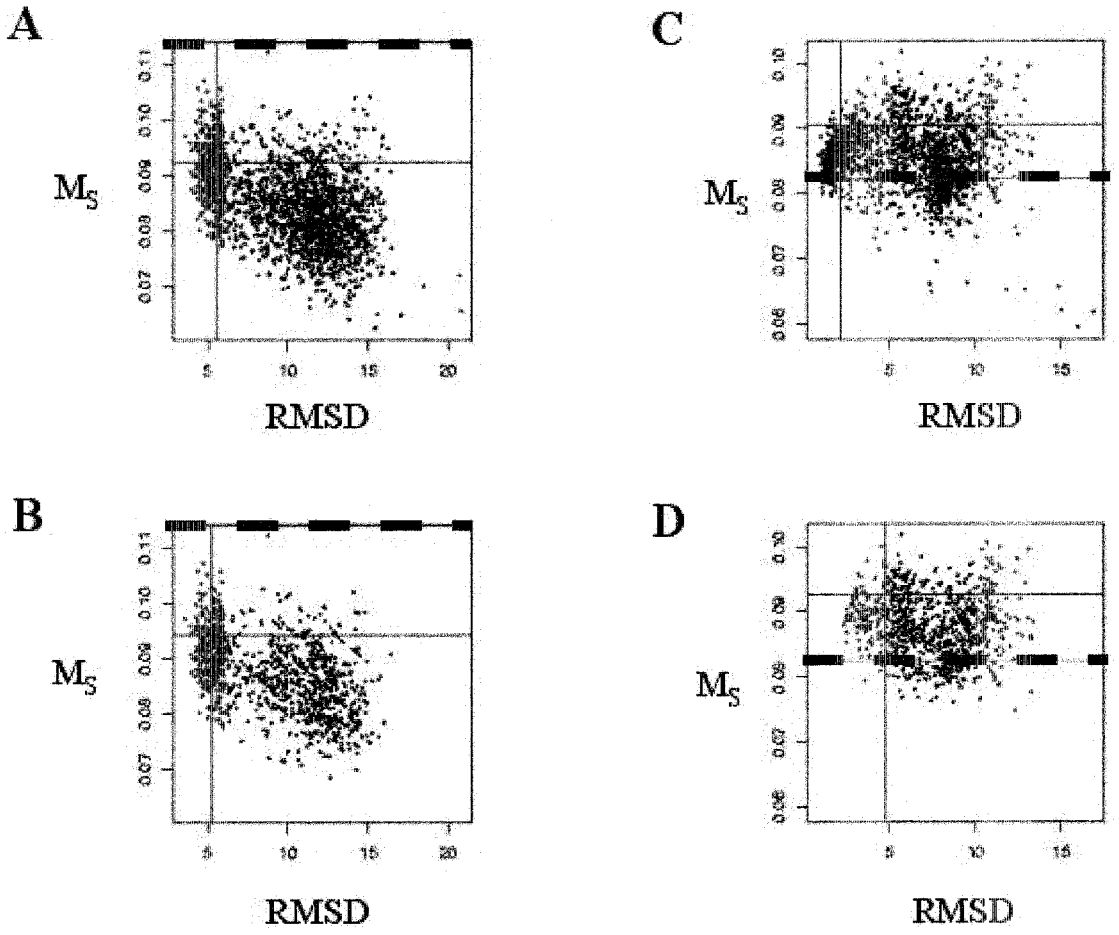

Fig. 3. Discrimination of close-to-native decoys based on clustering of conserved residues (c_{15} -C). For each decoy, M_S (eq. 1) is plotted versus RMSD to native structure. The score of the native structure is marked by a thick dashed line. The thin lines mark the cutoffs for the *low-RMSD decoys* (i.e., the 15% lowest RMSD structures; vertical line) and the *high scores* (i.e., the 15% highest scored structures; horizontal line). The enrichment E (eq. 2) is defined as the proportion of *low-RMSD decoys* with *high scores* (i.e., decoys located in the upper left quadrant) out of all *high scores* (i.e., the upper half), relative to the overall proportion of *low-RMSD decoys* (i.e., 15%). (**A**) Example for good discrimination: ACBP (1aca) with enrichment E = 2.6 (pval = 0). (**B**) As in (A), but only including compact structures (50% of the decoys with lowest R_{gyr}): E = 2.15 (pval = $6*10^{-8}$). (**C**) Example for no discrimination: ribosomal protein S15 (1a32) with enrichment E = 0.11 (pval = 1). (**D**) As in (C), but for compact structures: E = 0.83 (pval = 0.76).

may not enrich for low-RMSD structures, as exemplified in the ribosomal protein S15 (1a32[42]) [Fig. 3(C and D)].

To determine whether M_S can be used reliably for decoy discrimination, we carried out an analysis of the enrichments in a large set of proteins.

Subsets of Evolutionary Conserved Positions Can Be Used to Select Good Decoys

We calculated the enrichment values for our set of proteins in the same manner as for ACBP, described in the previous paragraph, and determined for how many of the proteins the enrichment was significant (Table II). Evaluation of different combinations of parameters shows that the best enrichment is obtained for subsets of size 15 and excluding conserved cysteine positions (e.g., c_{15} -C): 79% of the proteins are significantly enriched. Both hydrophobic and polar conserved residues contribute to the enrichment, as indicated by reduced performance of the nonhydrophobic and nonpolar subsets. When only compact structures are considered (i.e., decoys with the 50% lowest R_{gyr} values), the performance drops by about 10%.

Enrichment of Good Decoys Based on Scoring of Evolutionary Conserved Surface *or* Core *Positions*

We also assessed the degree of enrichment for the evolutionary conserved positions on the protein surface (by excluding the 30% most buried residues), and in the protein core (by excluding the 50% most exposed residues). For subsets of size 15, and excluding conserved cysteine positions (e.g., c_{15} -C), the percentage of proteins with significant enhancement was reduced from 79%, based on the whole protein, to 57%, based on the protein surface, and 58%, based on the protein core. Analysis of the contributions of hydrophobic and polar conserved residues to the enrichment showed that hydrophobic residues influence the enrichment based on the whole and the protein core, whereas the polar residues are important for good enrichment based on the whole and the protein surface.

In summary, we conclude that evolutionary conserved residues are significantly more clustered in low-RMSD decoys, and that this feature can be used to select low-RMSD decoys from a set of given decoys. For this, the

TABLE II. Enrichment of Good Decoys by M_S for Different Residue Subsets[†]

Enrichment of low-RMSD decoys	Whole protein	Whole protein, low compact decoys[a]	Protein surface	Protein core
All amino acids	67%	60%	62%	63%
Best condition for single exclusion	−C: 79%	−C: 67%	−L: 66%	−P: 66%
Worst condition for single exclusion	−I: 66%	−I: 57%	−G: 52%	−L: 56%
Non-hydrophobic	68%	57%	68%	48%
Non-polar	68%	56%	49%	66%

[†]The percentage of proteins with statistically significant enrichment is given for subsets of 15 residues (c_{15}).
See Table I for legend.
[a]50% of the decoys with lowest R_{gyr}.

whole protein should be included, because both conserved residues on the protein surface and in the protein core contribute to the significance of the enrichment. For best results, the subset of conserved residues should not include conserved cysteines.

DISCUSSION

This study investigates the clustering of conserved residues in protein structures. Two major questions are addressed: (1) How are conserved residues distributed in protein structures? (2) Can these distributions be used for de novo structure prediction?

Evolutionary conserved residues are clustered within protein structures. In this analysis of a representative set of 79 proteins, the conserved residues are significantly more clustered than random, same-size subsets in the same protein for 92% of the proteins. For 65% of the proteins, the conserved residues are also significantly more clustered in the native structure than in decoy structures, demonstrating that clustering is not due to sequence connectivity or specific residue types, because these are constant in all decoys. For 79% of the proteins, we were able to select good models from a set of decoys, as judged by enrichment.

The contribution to clustering of residues in the protein core (defined by excluding the 50% most exposed residues) and surface (defined by excluding the 30% most buried residues) shows that the most significant clustering was obtained when all residues were considered simultaneously. Considering the contributions of hydrophobic and polar residues, the dominant role of hydrophobic residues is clearly seen in the whole protein or the protein core, whereas conserved polar residues contribute primarily on not only the protein surface but also the whole protein when decoys are compared. The clustering of conserved residues may be more evident for the whole protein than for the core and surface alone for several reasons. First, almost all proteins have hydrophobic cores, and many have functional sites. Thus, in addition to the conserved hydrophobic core residues important for protein stability, the protein surface contains also some significantly clustered conserved residues that contribute to the signal in the current analysis. For decoy discrimination, the surface signal might be particularly important, because it provides a discriminating signal even when most of the decoys contain a hydrophobic protein core. Second, the clustering of the conserved exposed residues alone is often not detectable by simple inspection by eye or by clustering of exposed conserved residues. A conserved patch containing both buried and surface positions is only defined if both surface and core positions are included in the analysis. The example of ACBP demonstrates how core positions can contribute to the continuity of the conserved, functional patch involved in ligand binding, by connecting relatively dispersed conserved polar surface positions [Fig. 1(A, D, and E)]. The contribution of core residues both to protein stability and function might be a more general feature in proteins with functional sites located in buried clefts, such as in lysozyme. Indeed, the description of a functional patch as a number of relatively isolated, conserved exposed polar residues, connected into a contiguous patch by other, buried and nonpolar residues, such as in ACBP, was used by Aloy et al.[13] to predict functional sites in proteins. In their study, a functional site is modeled as a set of residues within a sphere whose definition is based on invariant polar residues. This scheme includes both conserved polar residues, which are predominantly exposed, and adjacent, nonpolar residues that are not necessarily conserved. Madabushi et al.[19] have used the evolutionary trace method[11] to identify functionally important sites. They too observe that single and dispersed residues with a high conservation rank tend to be connected into contiguous patches by adding additional residues with lower rank.

In their attempt to define functional patches on proteins, Madabushi et al.[19] compared the degree of clustering of conserved residues to random, same-size subsets on a set of 46 proteins. Based on the number of clusters and cluster size created by those residues, significant clustering was observed for all but one protein in this set, with at least one of the parameter combinations they tested. For future applications on as yet uncharacterized proteins, they suggest choosing for each protein separately the combination of parameters that results in a set of residues with the most significant clustering on its native structure. The largest cluster that is identified with this specific parameter combination can then be used to characterize the functional site of the protein. Here, in contrast, we are primarily interested in an overall measure for evaluating alternative conformations rather than identifying functional sites per se. In developing such a measure, we have, first, carried out a global evaluation of the optimal subset size and amino acid composition, and, second, utilized a

robust measure of residue clustering, $M_S = \langle 1/r \rangle$. M_S upweights the contributions from pairs of residues that are near in space, is robust to outliers, and combines features of the two alternative measures used previously, namely, cluster size and number of clusters. Moreover, it facilitates the ranking of alternative conformations, because the score for conformations is not restricted to integer values. It would be interesting to combine the two approaches by using the $\langle 1/r \rangle$ measure with the evolutionary trace definition of conserved residues, which takes into account residue conservation patterns within subfamilies.

We have shown that conserved residues are more clustered in the native structure than in the majority of the decoys. Olmea et al.[20] have threaded the sequence of a protein on an alternative protein structure of the same length and evaluated the ability of a clustering measure to distinguish the correct fold from this model. Conserved residues, correlated residues, apolar residues, and finally, conserved residues with a high tendency to occur in binding sites were all able to select the correct model for over 92% of the proteins in their test set. The reason for the significantly better discrimination is that the alternate models do not have any reasonable hydrophobic core: 95% of the pairs can only be distinguished based on the clustering of the hydrophobic residues, without considering their conservation. In our study, we concentrate on the contributions to structure and function beyond the creation of a hydrophobic core by attempting to distinguish between incorrect and correct models that all are optimized for hydrophobic core packing.

The ability to distinguish the native protein structure from alternative decoy conformations based on the degree of clustering of conserved residues suggested that this measure can be applied to recognize close-to-native conformations in de novo structure prediction. We demonstrate that, indeed, native-like conformations tend to show more clustering of the evolutionary conserved residues. Selection based on clustering of conserved residues should be orthogonal to selection methods based on traditional energy functions and should contribute to increasing the accuracy and reliability of de novo structure prediction.

ACKNOWLEDGMENTS

We wish to thank members of the Baker laboratory for stimulating discussions and support.

REFERENCES

1. Shortle D. Mutational studies of protein structures and their stabilities. Q Rev Biophys 1992;25:205–250.
2. Clackson T, Wells JA. A hot spot of binding energy in a hormone–receptor interface. Science 1995;267:383–386.
3. Ma B, Wolfson HJ, Nussinov R. Protein functional epitopes: Hot spots, dynamics and combinatorial libraries. Curr Opin Struct Biol 2001;11:364–369.
4. DeLano WL. Unraveling hot spots in binding interfaces: Progress and challenges. Curr Opin Struct Biol 2002;12:14–20.
5. Godzik A, Sander C. Conservation of residue interactions in a family of Ca-binding proteins. Protein Eng 1989;2:589–596.
6. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein–protein interaction. J Mol Biol 1997;271:511–523.
7. Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18:309–317.
8. Olmea O, Valencia A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Fold Des 1997;2:S25–S32.
9. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. Nat Struct Biol 1995;2:171–178.
10. Livingstone CD, Barton GJ. Protein sequence alignments: A strategy for the hierarchical analysis of residue conservation. Comput Appl Biosci 1993;9:745–756.
11. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
12. Lichtarge O, Sowa ME. Evolutionary predictions of binding surfaces and interactions. Curr Opin Struct Biol 2002;12:21–27.
13. Aloy P, Querol E, Aviles FX, Sternberg MJ. Automated structure-based prediction of functional sites in proteins: Applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol 2001;311:395–408.
14. Armon A, Graur D, Ben-Tal N. ConSurf: An algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 2001;307:447–463.
15. Landgraf R, Xenarios I, Eisenberg D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol 2001;307:1487–1502.
16. Fariselli P, Pazos F, Valencia A, Casadio R. Prediction of protein–protein interaction sites in heterocomplexes with neural networks. Eur J Biochem 2002;269:1356–1361.
17. Hu Z, Ma B, Wolfson H, Nussinov R. Conservation of polar residues as hot spots at protein interfaces. Proteins 2000;39:331–342.
18. Ouzounis C, Perez-Irratxeta C, Sander C, Valencia A. Are binding residues conserved? Pac Symp Biocomput 1998;401–412.
19. Madabushi S, Yao H, Marsh M, Kristensen DM, Philippi A, Sowa ME, Lichtarge O. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. J Mol Biol 2002;316:139–154.
20. Olmea O, Rost B, Valencia A. Effective use of sequence correlation and conservation in fold recognition. J Mol Biol 1999;293:1221–1239.
21. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
22. Sippl MJ, Weitckus S. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. Proteins 1992;13:258–271.
23. Jones DT, Thornton JM. Potential energy functions for threading. Curr Opin Struct Biol 1996;6:210–216.
24. Jones DT. Progress in protein structure prediction. Curr Opin Struct Biol 1997;7:377–387.
25. Venclovas, Zemla A, Fidelis K, Moult J. Comparison of performance in successive CASP experiments. Proteins 2001;45:163–170.
26. Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. Curr Opin Struct Biol 2000;10:139–145.
27. Aszodi A, Taylor WR. Homology modelling by distance geometry. Fold Des 1996;1:325–334.
28. Smith-Brown MJ, Kominos D, Levy RM. Global folding of proteins using a limited number of distance constraints. Protein Eng 1993;6:605–614.
29. Mumenthaler C, Braun W. Predicting the helix packing of globular proteins by self-correcting distance geometry. Protein Sci 1995;4:863–871.
30. Ortiz AR, Kolinski A, Skolnick J. Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. J Mol Biol 1998;277:419–448.
31. Bowers PM, Strauss CE, Baker D. De novo protein structure determination using sparse NMR data. J Biomol NMR 2000;18:311–318.
32. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
33. Eyrich VA, Standley DM, Friesner RA. Prediction of protein tertiary structure to low resolution: Performance for a large and structurally diverse test set. J Mol Biol 1999;288:725–742.
34. Dunbrack RLJ, Wang G. PISCES: A protein sequence culling server. Submitted for publication.

35. Bhat TN, Bourne P, Feng Z, Gilliland G, Jain S, Ravichandran V, Schneider B, Schneider K, Thanki N, Weissig H, Westbrook J, Berman HM. The PDB data uniformity project. Nucleic Acids Res 2001;29:214–218.
36. Simons KT, Strauss C, Baker D. Prospects for ab initio protein structural genomics. J Mol Biol 2001;306:1191–1199.
37. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
38. Shannon CE. A mathematical theory of communication. Bell System Tech J 1948;27:379–423, 623–656.
39. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: Detection of distantly related proteins. Proc Natl Acad Sci U S A 1987;84:4355–4358.
40. Valdar WS. Scoring residue conservation. Proteins 2002;48:227–241.
41. Kragelund BB, Andersen KV, Madsen JC, Knudsen J, Poulsen FM. Three-dimensional structure of the complex between acyl-coenzyme A binding protein and palmitoyl-coenzyme A. J Mol Biol 1993;230:1260–1277.
42. Clemons WM Jr, Davies C, White SW, Ramakrishnan V. Conformational variability of the N-terminal helix in the structure of ribosomal protein S15. Structure 1998;6:429–438.
43. Sayle RA, Milner-White EJ. RASMOL: Biomolecular graphics for all. Trends Biochem Sci 1995;20:374.