PROTEINS: Structure, Function, and Genetics 18:309–317 (1994)

Correlated Mutations and Residue Contacts in Proteins

Ulrike Göbel, Chris Sander, Reinhard Schneider, and Alfonso Valencia
Protein Design Group, European Molecular Biology Laboratory, D-69012 Heidelberg, Germany.

ABSTRACT **The maintenance of protein function and structure constrains the evolution of amino acid sequences. This fact can be exploited to interpret correlated mutations observed in a sequence family as an indication of probable physical contact in three dimensions. Here we present a simple and general method to analyze correlations in mutational behavior between different positions in a multiple sequence alignment. We then use these correlations to predict contact maps for each of 11 protein families and compare the result with the contacts determined by crystallography. For the most strongly correlated residue pairs predicted to be in contact, the prediction accuracy ranges from 37 to 68% and the improvement ratio relative to a random prediction from 1.4 to 5.1. Predicted contact maps can be used as input for the calculation of protein tertiary structure, either from sequence information alone or in combination with experimental information.**

Key words: protein structure prediction, predicted contact maps, correlated mutations

INTRODUCTION

Evolutionary Constraints on Protein Sequences

The problem of predicting a protein fold from sequence information alone is a difficult one. The increasing number of protein families for which many homologous sequences are known affords a new opportunity to exploit evolutionary information. At the level of protein molecules, selective pressure results from the need to maintain protein function, e.g., efficient catalysis or specific protein–DNA interaction, which in turn requires maintenance of the specific three-dimensional structure consistent with that function. Accordingly, conservation and mutation patterns observed in multiple sequence alignments are evidence of functional or structural constraints plus mutational drift. Functional constraints typically involve surface residues, mutational drift most easily occurs in loop regions not involved directly in functional interactions, and structural constraints usually are strongest in the protein interior, or "core." The historical process is thought to involve a series of (random) point mutations, differential replication of genetic information depending on genotype and phenotype, and elimination of cells containing dysfunctional or noncompetitive sequences. When functionally negative point mutations are compensated for by other mutations, the cells may survive and with it the protein sequence information. We need to learn how to extract information about the various types of evolutionary constraints from multiple sequence data and then exploit this information for the prediction of three-dimensional structure, via distance constraints.

Functional or Structural Constraints?

Given many sequences in a protein family, the first tasks in decoding the observed mutational patterns are to identify constrained sequence changes on a background of neutral mutational drift, i.e., to separate signal from noise, and to separate the effects of structural constraints from those of functional constraints that have no structural implications. The task is difficult for several reasons. First, evolutionary constraints are not directly observable, as they specify which attempted point mutations would lead to the elimination of a particular genotype from a population. Second, the distinction between functional and structural constraints is not clear-cut: e.g., mutations in some residues may affect interactions crucial for substrate interaction as well as for correct folding. Third, functional constraints are stronger than structural constraints in that only a few residue combinations are admissible in active site positions while considerable sequence variation is admissible in most positions in the structural framework, away from active sites. Fourth, sequence information is always incomplete, as the database of known sequence represents a highly nonrandom sample of all natural sequences and all natural sequences are probably a small subset of all possible sequence that are consistent with a known structure. So some protein families contain a widely dispersed set of sequences while others contain just one narrow group of highly similar sequences. Fifth, because of the high cooperativity of protein folding and the plasticity of protein struc-

Received August 17, 1993, revision accepted November 12, 1993.
Address reprint requests to Chris Sander, Protein Design Group, European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69012 Heidelberg, Germany.

ture, the compensatory response to a point mutation may be distributed over a cluster of residues rather than occur at a single paired residue. Sixth, complicated chains of interactions may lead to compensatory mutations at spatially distant sites, rather than only in contacting residues. Finally, experimental studies of these effects and their theoretical simulation are still at the developmental stage.[1–3]

Focus on Structural Constraints

In the face of these difficulties, we have to make simplifying assumptions if we want to extract information about 3D structure directly from a family of sequences. The key assumption is that residues in physical contact show correlated mutational behavior: if one residue mutates, its contact partners also tend to mutate. In addition, we assume that this effect is detectable at the level of pair correlations. Higher order clusters are then built up from mutually correlated pairs.

The method is based on defining an exchange matrix or other similarity measure at each sequence position in a multiple alignment and then calculating a correlation coefficient between the exchange matrices at any two positions. The map of position–position correlations is interpreted as a predicted residue–residue contact map and is then compared with the contact map from experimentally observed structures. The extent of agreement between the two gives an indication of the validity of the underlying assumptions and of the usefulness of the predicted contact maps for prediction of protein three-dimensional structures.

METHODS

Mutational Behavior at One Position

The observed mutations at a sequence position in the protein family are taken to describe the allowed variation of residue types at that structural position (Fig. 1). To quantify this, consider all exchange pairs observed at a sequence position i, i.e., all exchanges that would transform the residue observed in one protein labeled k, into the one observed in another protein, labeled l. For each such pair, enter the similarity $s(i,k,l)$ of the two residue types in an N by N matrix, where N is the number of aligned sequences. The similarities between the 20 basic residue types are taken from a 20 by 20 table of similarity constants derived from statistical or physical considerations.[4] The best choice of similarity table is a matter for future research. The matrix $s(i,k,l)$ is taken to represent the mutational behavior at position i.

Comparing Mutational Behavior at Two Positions

Comparison of the two mutation matrices $s(i,k,l)$ and $s(j,k,l)$ can be made in various ways. Here, we

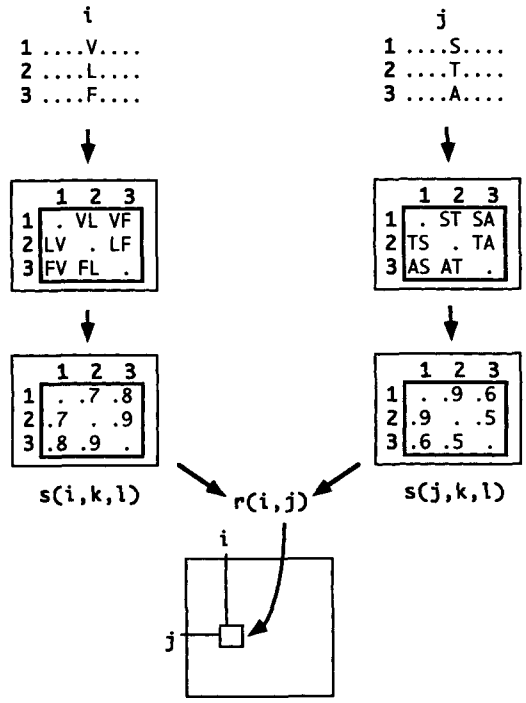

Fig. 1. Schema going from putative correlated mutations to predicted contacts. A protein family is presented as a multiple sequence alignment (series of horizontal lines) and assumed to correspond to a common three-dimensional structure. Mutational behavior at each single position is summarized in a mutation matrix. The mutation matrix contains the amino acid similarity of any pair of residues at that position (indices k,l run over proteins in the family; i,j run over common positions in the sequences). Correlated mutational behavior of different positions carries information about functional and structural constraints acting at these positions. This information can be extracted by calculating correlation coefficients between the mutation matrices for each pair (i,j) of positions. When the correlation is above a chosen threshold, a residue–residue contact is predicted. The accuracy of prediction is evaluated by comparing the pattern of predicted contacts with that derived from experimentally known three-dimensional structures (Fig. 6).

use the correlation coefficient, i.e., the inner product between two unit vectors of length N^2. In order to down-weight information coming from very sequence-similar pairs of proteins, a weighted correlation coefficient is defined: each term in the sum can be weighted with the mutual distance $w(k,l)$ $(0 < w < 1)$ between proteins k and l in sequence space, defined as the fraction of residue type mismatches over the entire aligned sequence length, defined as in Sander and Schneider,[5] but normalized to sum to 1.0. In this way we obtain a measure of correlated mutational behavior between sequence position i and j as

$$r_{ij} = \frac{1}{N^2} \sum_{kl} \frac{w_{kl} (s_{ikl} - \langle s_i \rangle) (s_{jkl} - \langle s_j \rangle)}{\sigma_i \sigma_i}$$

where σ_i is the standard deviation of s_{ikl} about the

mean $\langle s_i \rangle$ and the indices k,l run from 1 to N, the number of sequences in the family.

The significance of the observed correlations can be assessed by randomizing the sequence order, independently at each position, calculating the mean and standard deviation for, say, 200 randomizations, and measuring the correlation in units of standard deviations. The use of weights and normalization gave better results in some cases (data not shown). The best choice of weights is currently an open question. For simplicity, we report here the raw data, without normalization, using the unweighted coefficient throughout.

Prediction of Pair Contacts and Contact Clusters

In the working hypothesis, the extent of correlation between two residues is related to the interresidue contact strength. In the simplest form, two positions are predicted to be in direct contact if their mutational correlation $r(i,j)$ exceeds a certain cutoff r_c. Completely conserved residues are excluded from the analysis, as are positions with more than 10% gaps in the alignment. The predicted contacts can be compared with the actual contacts derived from the experimentally known structure.

Mutational response to a point mutation is in general distributed over more than one (neighboring) residue. To reflect this effect, we define clusters of correlated residues. A cluster of rank n is defined as follows: a residue is part of a cluster if it is correlated with at least n other residues in the cluster. For example, if residues i,j as well as j,m and m,i are correlated as pairs, then the triplet i,j,m form a cluster of rank 2. A particular residue may belong to more than one cluster, i.e., clusters may overlap. The rank n is a measure of cooperative correlations in a cluster. An interesting hypothesis to be checked is that residues in mutationally correlated clusters also cluster in 3-D space.

Accuracy, Coverage, Completeness

The utility of a predicted contact map depends on several aspects. How many of the *predicted* contacts correspond to actually observed contacts (accuracy)? How many of the *observed* contacts are correctly predicted (completeness)? And, how uniformly are the predicted contacts distributed over the main contact regions or how many segment–segment interfaces are covered by at least one predicted contact (coverage)?

As the method is still in its infancy, we report here only the overall accuracy, a_{pred}, and the improvement in accuracy relative to a random prediction, $R_{improve}$. The accuracy is defined simply as $a_{pred} = C_{correct}/C_{pred}$, where $C_{correct}$ is the number of correctly predicted contacts and C_{pred} is the total number of predicted contacts. A random prediction corresponds to placing predicted contacts randomly anywhere in the contact map. Its accuracy is $a_{random} = C_{observed}/C_{max}$, where $C_{observed}$ is the number of contacts observed in the three-dimensional structure and C_{max} the total number possible contacts, i.e., the size of the contact map. The improvement ratio is defined as $R_{improve} = a_{pred}/a_{random}$.

The accuracy a_{pred} can be interpreted as the estimated probability of correct prediction, given a predicted contact. A perfect prediction has accuracy $a_{pred} = 1.0$. The accuracy of a random prediction depends on the average density of contacts, i.e., on the size of the protein: for example, for 56 residues of trypsin inhibitor (6pti) we have $a_{random} = 0.39$, while for the 223 residues of trypsin (PTP) we have $a_{random} = 0.13$. For this reason, the improvement ratio is a more appropriate measure than the bare accuracy. The larger the protein, the more difficult it is to achieve a perfect prediction, as the set of experimentally observed contacts is an increasingly smaller fraction of the set of all possible contacts. As proteins with homologous three-dimensional structure do not necessarily share 100% of all contacts, a realistic goal of near-perfect prediction is about 90%.

RESULTS

Having defined a simple measure of correlated mutational behavior for pairs of positions, one needs to assess whether such correlations carry any information about spatial proximity. If two positions are correlated in their mutational behavior, does this imply that they are in physical contact in the three-dimensional structure? What is the accuracy of contact prediction? Comparison of predicted and observed contacts in several protein families provides the first answers.

Correlated Pairs and Predicted Contacts

More correlated pairs tend to have smaller inter-side chain distances

A simple measure of physical contact of two residues is the distance of the first side chain atoms, i.e., the C^β–C^β distance. The scatter plot of C^β–C^β distances against the mutational pair correlation reveals the extent to which correlated residues are close in space (Fig. 2). For the small protein pancreatic trypsin inhibitor (6pti), the six most highly correlated pairs have C^β–C^β distances of about 5, 10, 6, 5, 20, and 16 Å, respectively. Below a correlation value of 0.5, more and more of the correlated pairs have larger distances, up to about 30 Å. Overall, there is a weak but unambiguous tendency for more correlated pairs to have smaller inter-side chain distances. This tendency is present in all families analyzed (e.g., ribonuclease 1rbb, Fig. 3a). Choosing an appropriate cutoff in r, one can attempt to isolate the correlated pairs that carry some information about physical contact from the less informative background (Fig. 3). The trade-off is in a decrease in the

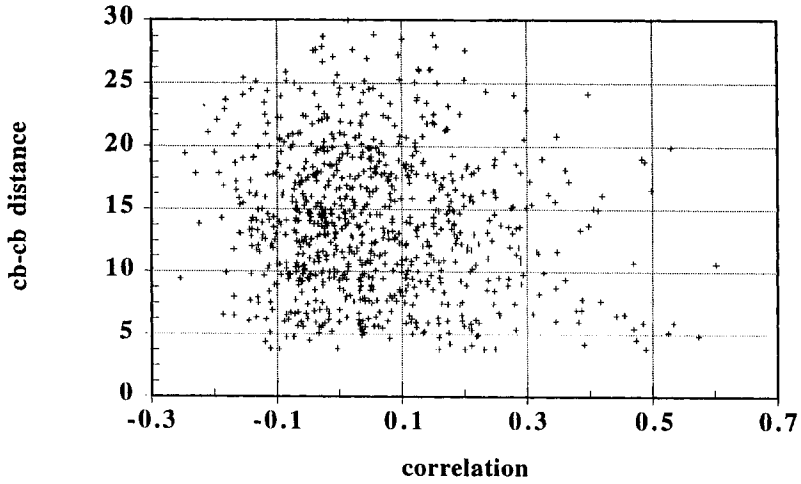

Fig. 2. Scatter plot of C^{β}–C^{β} distances [Å] of pairs of residues, derived from the crystal structure of 6pti, and mutational pair correlation r, derived from the multiple sequence alignment. Each point represents a single pair (i,j) of positions in the protein. In spite of considerable background, there is a discernible trend: the C^{β}–C^{β} distance of a residue pair tends to smaller values the higher its mutational pair correlation. Completely conserved residues are excluded from the analysis. Also, positions are excluded at which more than 10% of proteins in the alignment have a gap.

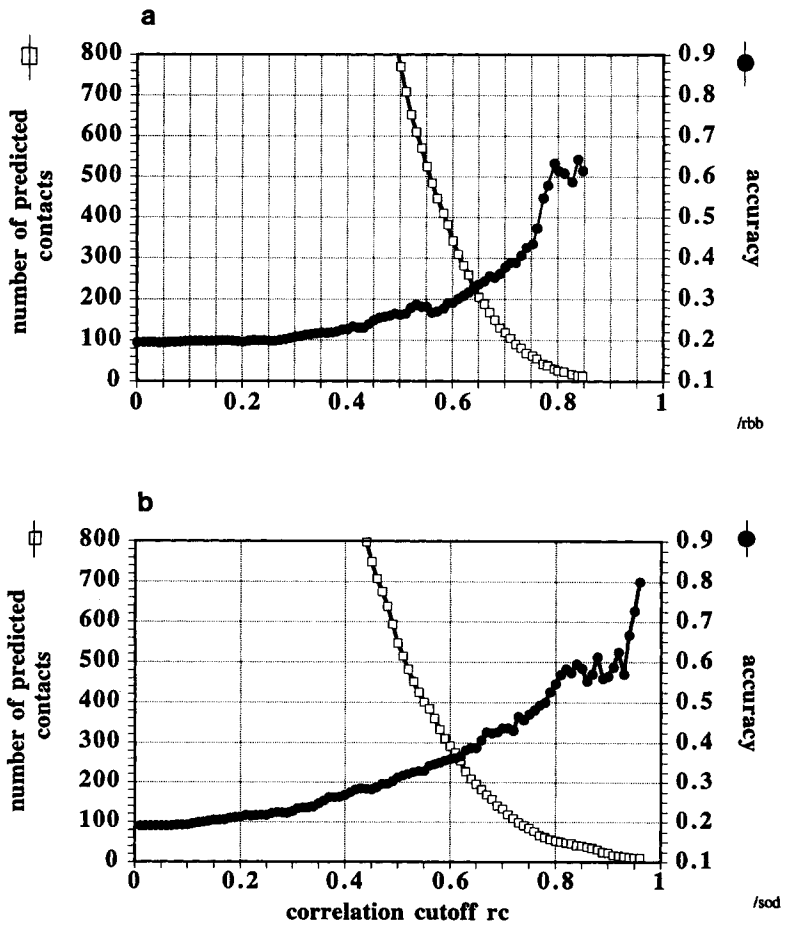

Fig. 3. The number of predicted contacts (bullets) and the accuracy (squares) are plotted as a function of the pair correlation cutoff r_c for one protein. Each point corresponds to a prediction of contacts in the particular protein, using only pairs with $r > r_c$. A higher cutoff r_c results in a higher accuracy of contact prediction. The price paid is a smaller number of predicted contacts. (**a**) Pancreatic ribonuclease (1rbb). An accuracy of 0.4 is achieved for about 90 predicted contacts ($r > 0.7$). (**b**) Superoxide dismutase (2sod). In this protein, an accuracy of 0.4 is achieved for about 190 predicted contacts ($r > 0.65$).

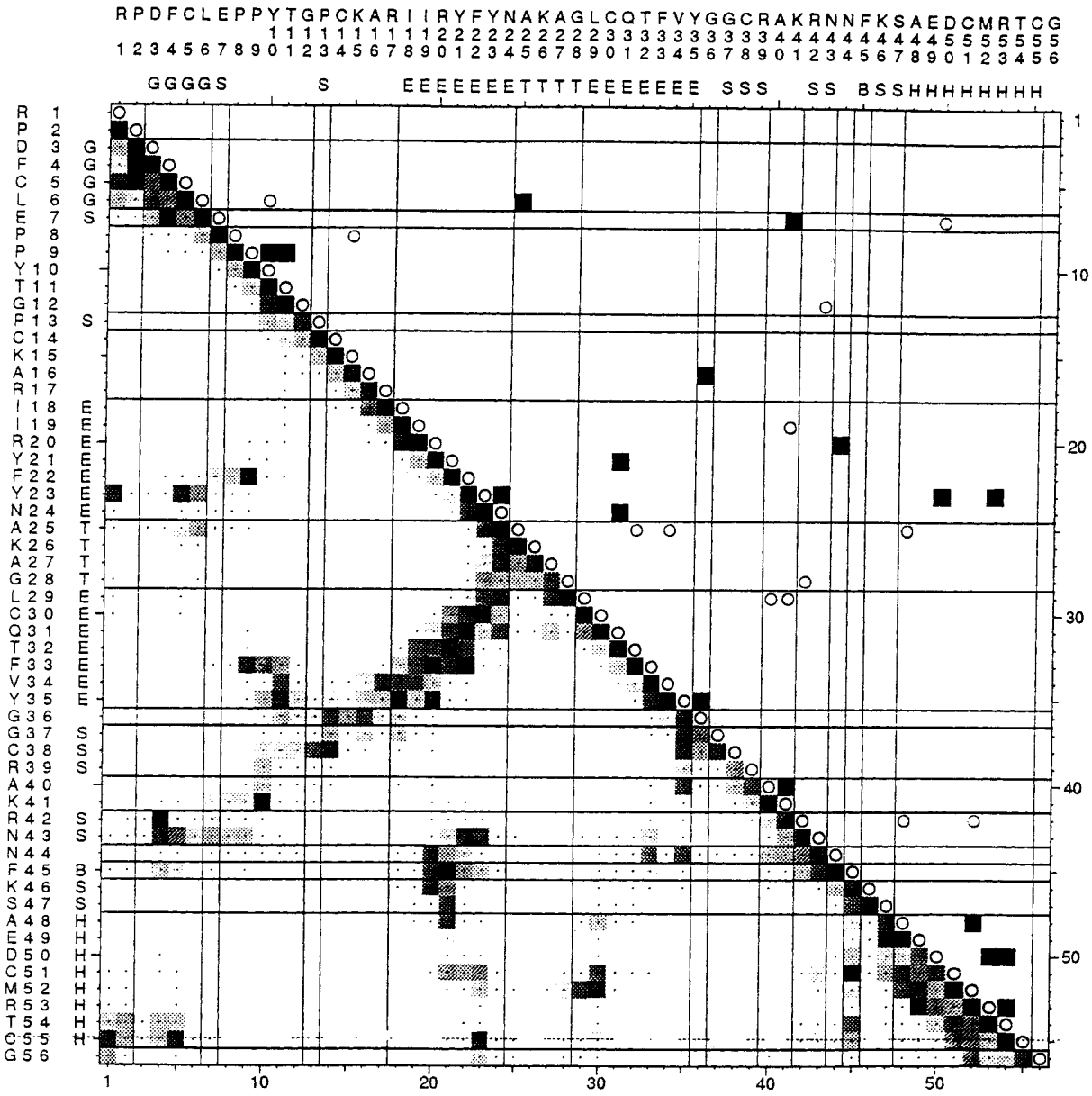

Fig. 4. Comparison of predicted and observed contacts for trypsin inhibitor (6pti). A contact is considered predicted correctly if it matches an observed contact (filled squares). Lower left: residue–residue contact map derived from atomic contacts in the crystal structure; the stronger a contact, the more intense the marker at *i,j;* top right: contacts predicted for correlated pairs with $r > 0.425$; correctly predicted, filled squares; incorrectly predicted, open circles. The number of predicted contacts is 30; prediction accuracy is $a = 0.57$. Details: contact strength is calculated as in Sander et al.,[16] using a 6–9 van der Waals potential. The weakest contacts shown (dots) correspond to a C^{β}–C^{β} distance of about 10 Å. The results are similar if these weak contacts are omitted. Secondary structure symbols are extracted from the crystal structures using the program DSSP,[17] with H = α-helix, E = β-strand, G = 3_{10}helix, T = hydrogen bonded turn, S = bend.

total number of predicted contacts. For example, for ribonuclease, a cutoff of $r_c = 0.4$ leads to 81 predicted contacts, of which 33 are correctly predicted. With this cutoff, the number of predicted contacts is small compared to the number of contacts observed in the crystal structure (1603). If the cutoff is lowered, more contacts are predicted, but with lower accuracy.

Predicted contacts in a small protein are fairly accurate

Exploiting the overall tendency, residue pairs with correlation values $r > r_c$ are predicted to be in contact. Comparison of the predicted contact map of 6pti with the observed contact map (Fig. 4) illustrates how well the prediction works in this case.

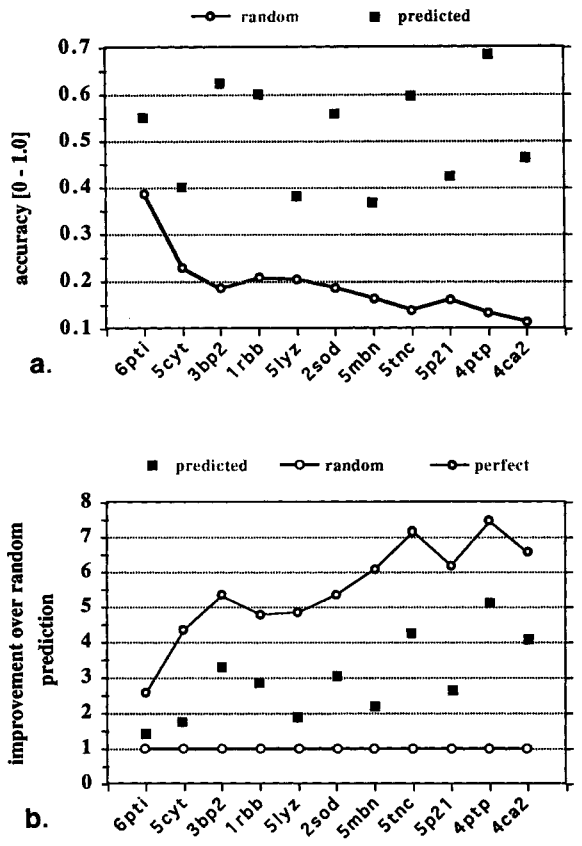

Fig. 5. (a) Accuracy of contact prediction from correlated mutations for each of 11 different protein families and (b) improvement relative to random prediction. In each case, the number of predicted contacts is $L/5$ (by appropriate choice of r_c), where L is the length of the guide sequence of known structure used to align the sequences in the family. The accuracy of random prediction varies with protein size, contact density, and observed sequence variation in a particular protein family. The improvement ratio indicates how much better the prediction is relative to that from a random method, e.g., the accuracy of 0.55 for protein 6pti is 1.4 times better than random. A perfect prediction, given for reference, corresponds to 100% accuracy for the predicted contacts. Proteins are identified by their names in the Protein Data Bank[18]: 6pti, pancreatic trypsin inhibitor; 5cyt, cytochrome *c*; 3bp2, phospholipase A2; 1rbb, pancreatic ribonuclease; 5lyz, lysozyme; 2sod, superoxide dismutase; 5mbn, myoglobin; 5tnc, troponin C; 5p21, transforming protein H-ras (p21); 4ptp, trypsin; 4ca2, carbonic anhydrase II.

6pti was chosen here as an example because of its small size, although the prediction accuracy is below average for this protein. Out of the 30 predicted contacts, 17 contacts are actually observed in the crystal structure, while 13 contacts are predicted incorrectly. The corresponding accuracy (number of correctly predicted/number predicted) of $a = 0.57$ is a factor of 1.4 better than that for a random prediction ($a = 0.39$, calculated as number of observed contacts/number possible contacts), the lowest improvement of the 13 proteins tested. Interestingly, four contacts are correctly predicted in the long antiparallel β-hairpin, one in the loop–strand contact region near N44-R20 and two in the structurally important helix–sheet interface near Y23-C55.

In addition, one could add potential disulfide bonds between the six conserved Cys residues as predicted contacts, with an accuracy ratio of 3/12. Completely conserved residues were ignored in the current results, as they are trivially and perfectly correlated with one another, although they contain important information about evolutionary constraints (see discussion of structural versus functional constraints).

Accuracy for eleven protein families

What is the accuracy of contact prediction for different protein families? How good are the predictions relative to an "ignorant" random prediction? To compare accuracies for different protein families, one needs a common baseline. Complications arise from the fact that the average density of contacts in the two-dimensional contact map decreases with protein size and that the average accuracy decreases with the number of predicted contacts. The problem is dealt with as follows. We choose a cutoff r_c in correlation value for correlated mutations such that the number of predicted contacts equals $L/5$, where L is the length of the protein chain. As the number of observed contacts C_{obs} rises approximately linearly with L, the number of predicted contacts is a roughly constant fraction of the number of observed contacts.

The accuracy thus assessed varies from 0.37 in myoglobin to 0.68 in trypsin (Fig. 5a, Table I). The improvement over a random prediction (accuracy of present method/accuracy of random prediction) varies from a factor of 1.4 in trypsin inhibitor (6pti) to a factor of 5 in trypsin (4ptp) (Fig. 5b). Note that a perfect prediction is much more difficult to reach for larger proteins, as a result of the smaller fraction of observed contacts compared to the number of all possible contacts. We see no clear tendency of accuracy depending on the size or folding type of the protein. However, larger sequence diversity in a family usually results in higher accuracy of prediction. This was demonstrated for ribonuclease (1rbb) and superoxide dismutase (1sod) by omitting some remote homologues from the family. This reduction in diversity resulted in decreased accuracy of predicted contacts (Table I). For example, for the ribonuclease family, accuracy drops by 12 percentage points when the family size is reduced from 54 sequences to 42 sequences. Overall, prediction accuracy is clearly better than random for most proteins and is surprisingly close to a perfect prediction in one case: 30 out of the 44 predicted contacts, or 68%, coincide with contacts observed in the crystal structure of trypsin, 5.0 times better than for a random prediction.

Networks of Correlated Pairs and Predicted 3D Clusters

It is useful to extend the analysis beyond mere pairs of residues and ask if a network of mutational

TABLE I. Prediction of Residue Contacts in 11 Protein Families

Family*	Length of alignment	Number of sequences in alignment	Least related sequence: % identical residues	Correlation cutoff r_c	Number of predicted contacts (1/5 chain length)	Number of correctly predicted contacts	Accuracy of predicted contacts	Improvement over random prediction
6pti	56	39	37	0.43	11	6	0.55	1.42
5cyt	103	110	33	0.67	20	8	0.4	1.74
3bp2	122	103	31	0.73	24	15	0.63	3.34
1rbb	124	54 (42)†	30 (77)	0.8	25	15 (4)	0.6 (0.17)	3.01 (1.19)
5lyz	129	47	33	0.82	26	10	0.38	1.85
2sod	151	45 (12)†	30 (68)	0.87	32	18 (8)	0.56 (0.25)	2.24 (1.21)
5mbn	153	76	30	0.92	30	11	0.37	4.26
5tnc	161	108	30	0.65	32	19	0.59	2.63
5p21	166	107	30	0.71	33	14	0.42	2.88
4ptp	223	164	31	0.52	44	30	0.68	5.10
4ca2	255	23	34	0.9	52	24	0.46	4.05

*Names in the PDB data base. Protein names are given in Figure 5.
†Numbers in parentheses for 1rbb and 2sod indicate results for a subset of aligned sequences.

correlation can be used to predict successfully a network of physical contacts. This is physically reasonable, as the observed plasticity of protein structure implies that compensating mutations do not always occur in pairs. Rather, different parts of the entire environment of a mutated residue may adjust to a mutation. Networks of residues mutually correlated may reflect a series of mutational events confined to a small region of the molecule. So, in favorable cases, a network of mutational correlation is correctly predicted to correspond to a network of physical contacts.

In an example from ribonuclease (Fig. 6), several mutually correlated residues are in a contact region between a three-stranded β-sheet and an α-helix. The cluster includes direct contacts between adjacent residues in the helix (the pair M29/R33), between adjacent strands (the pair H48/T82), and between the sheet and helix (the pair Y25/T82 and Y25/H48). Remarkably, all three pairs of the triplet Y25/T82/H48 are mutually correlated and are in pairwise physical contact. In addition, most of the residues in this cluster (Fig. 6) contact F46, a completely conserved buried residue (strong contacts with residues 23, 25, 29, 33, 48, weak contacts with 35 and 36). Completely conserved residues such as F46 are not included in the analysis, but can have a role in correlated mutational behavior. So in this case the prediction of a contact cluster is reasonably successful.

Related Approaches

The study of complementary changes of volume and/or size in the interior of proteins has been a classical topic in protein analysis,[6–8] Lesk and Chothia stated "Although complementary mutations do occur . . . they are not the rule." More recently, studying conservation of protein volume,[9] the same authors conclude that "it is not necessary for mutations to be locally compensating to produce the small observed variation in core volume."[10]

The idea of exploiting evolutionary data on correlated mutations using characteristics other than volume appears to have surfaced independently at different times. Only recently, however, have there been sufficient data and database tools for more quantitative and systematic work so that an assessment of accuracy becomes possible. For example, Altschuh et al. analyzed patterns of residue replacements in evolutionary trees.[11,12] For tobacoviruses they reported that "the positions in the sequence that show an identical pattern of variation in seven related viruses are mainly close together in the three-dimensional structure," but also "the spatial proximity of some residues with identical conservation patterns may be fortuitous." They went on to analyze sequences in the serine protease, cysteine protease, and hemoglobin families and concluded that "coordinated changes have been found in all three protein families mostly within structurally constrained regions" and that their "method works with a varying degree of success depending on the function of the proteins, the range of sequence similarities and the number of sequences considered." Shindyalov et al. reconstructed likely historical mutation pathways from trees of about 60 sequence families.[13] Applying a very stringent cutoff of significance, they showed that a small number of residue contacts can be predicted with reasonable accuracy and concluded that "significant correlation is observed between correlated mutation of some residue pairs and their spatial proximity in the three-dimensional structure."

Our method differs from earlier work in three as-

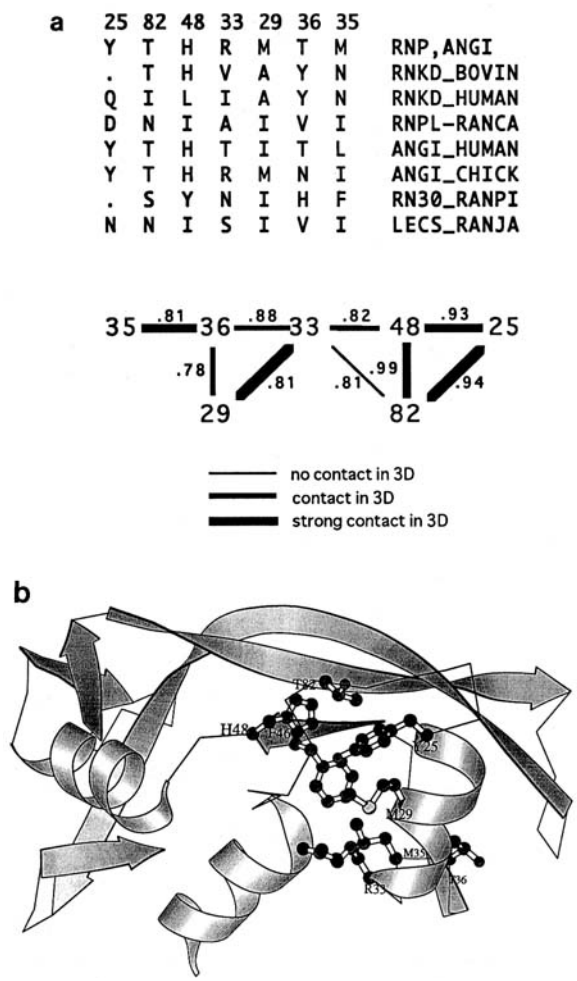

Fig. 6. Example of a cluster of residues connected by pairwise correlated mutations in pancreatic ribonuclease (1rbb[19]). (**a**) Residues in the cluster are internally connected by mutational correlations. Residues are identified by the residue number and are connected by a line labelled with the correlation value r if correlated with $r > 0.8$. Thin line, the contact is not observed experimentally (33–82); medium line, experimentally observed contact; thick line, experimentally observed strong contact. Representative sequences (8 out of 54) give a flavor of the difficulty of seeing the correlations by eye in a multiple sequence alignment. Protein names are RNP, pancreatic ribonuclease; ANGI, angiogenin; RNKD, nonsecretory ribonuclease K2; RN30, P-30 protein "onconase"; LECS, sialic acid-binding lectin from *Rana pipiens* (RANPI), *Rana japonica* (RANJA), *Rana catesbeiana* (RANCA). (**b**) The cluster of correlated residues from (a) mapped onto the backbone ribbon of the three-dimensional crystal structure: residues Y25, M29, R33, M35, T36, H48, T82, and the completely conserved residue F46. These residues form a network of contacts in a helix-sheet interface. Residues M29/R33 are in contact on two subsequent turns of a helix ($i/i + 4$). Residues H48/T82 make a side chain–side chain contact (within the hydrogen bonding O to N distance of 3.0 Å) typical of $i + 2/j$ pairs in twisted beta sheets, where i/j (F46/T82) are backbone–backbone hydrogen bonding partners. The helix–sheet contact involves the residue pairs Y25/T82 (C–C distance 3.3 Å) and Y25/H48 (side chain O–N and C–C distances of 3.5 Å in spite of a C^{α}–C^{α} separation of 12 Å). Residue F46, at the center of the cluster, is completely conserved and can be involved indirectly in correlated mutational behavior. For this part of the cluster, mutational correlation is a good predictor of physical contact.

pects. (1) Generality: the use of correlation coefficients between sets of numbers at any two positions is general and can be applied to different measures of mutational behavior. (2) Similarity versus identity: we use a 20 by 20 table of similarities between the 20 residue types, rather than just residue identity, to quantify the severity of a mutational transition (the particular choice of similarity table has not yet been optimized). (3) Independence: the key conceptual element of the current approach is the description of the mutational behavior at each position by a similarity matrix; and comparison of two positions by comparing their similarity matrices, independent of the construction of evolutionary trees.

Toward Structure Prediction

The tendency for correlated mutations to indicate contacts between two or more residues is clear, in spite of considerable background and false positives. A stronger signal, as a result of better data or improved methods of data analysis, is required before predicted contacts can successfully be used for the prediction of tertiary structure. Other information on contacts can be brought in, especially physical information about the potential of mean force between different residue types. Filters can be applied that express the cooperativity of interactions in regions of space and along stretches of sequence. Finally, distance geometry calculations can be used in attempts to predict protein folds (e.g., 14,15). In the analysis of correlated mutations, one of the key problems to be resolved is the separation of the structural and functional information contained in multiple sequence alignments. As multiple sequence data on more and more protein families become available, we anticipate that the analysis of correlated mutations will be performed routinely and yield information useful in the prediction of three-dimensional structure.

ACKNOWLEDGMENTS

We acknowledge stimulating discussions with Erwin Neher, who independently has developed a similar approach to the analysis of correlated mutations ("Fluctuation analysis on tables of aligned sequences," personal communication). We thank the Human Frontiers Science and EC Bridge Programs for financial support.

REFERENCES

1. Gregoret, L.M., Sauer, R.T. Additivity of mutant effects assessed by binomial mutagenesis. Proc. Natl. Acad. Sci. U.S.A. 90:4246–4250, 1993.
2. Lee, C., Levitt., M. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. Nature (London) 352:448–451, 1991.
3. Wells, J.A. Additivity of mutational effects in proteins. Biochemistry 29:8509–8517, 1990.

4. McLachlan, A.D. Tests for comparing related amino acid sequences. J. Mol. Biol. 61:409–424, 1971.
5. Sander, C., Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–68, 1991.
6. Kendrew, J.C., Watson, H.C. Stabilizing interactions in globular proteins In: "Principles of Biomolecular Organization." Wolstenholme, G.E.W., O'Connor, M., eds. London: J & A Churchill, 1966.
7. Lim, V.I., Ptitsyn, O.B. On the constancy of the hydrophobic nucleus volume in molecules of myoglobins and hemoglobins. Mol. Biol. (USSR) 4:372–382, 1970.
8. Lesk, A.M., Chothia, C. How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. J. Mol. Biol. 136:225–270, 1980.
9. Ptitsyn, O.B., Volkenstein, M.V. Protein structures and the neutral theory of evolution. J. Biomol. Struct. Dyn. 4:137–156, 1986.
10. Gerstein, M., Sonnhammer, E., Chothia, C. Volume changes in protein evolution, 1993, submitted.
11. Altschuh, D., Lesk, A.M., Bloomer, A.C., Klug, A. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. J. Mol. Biol. 193:693–707, 1987.
12. Altschuh, D., Verner, T., Berti, P., Moras, D., Nagai, K. Correlated amino acid changes in homologous protein families. Prot. Engin. 2:193–199, 1988.
13. Shindyalov, I.N., Kolchanov, N.A., Sander, C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Engin. 1994, in press.
14. Galaktionov, S.G., Rodionov, M.A. Calculation of the tertiary structure of proteins on the basis of analysis of the matrices of contacts between amino acid residues. Biophysics 25:395–403, 1980; translation of Biofizika 25:385–392, 1980.
15. Saitoh, S., Nakai, T., Nishikawa, K. A geometrical constraint approach for reproducing the native backbone conformation of a protein. Proteins 15:191–204, 1993.
16. Sander, C., Scharf, M., Schneider, R. Design of protein structures. In: "Protein Engineering, a Practical Approach." Rees, A.R., Sternberg, M.J.E., Wetzel, R., eds. Oxford: Oxford University Press, 1992:89–115.
17. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1984.
18. Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., Weng, J. In: "Crystallographic Databases—Information Content, Software Systems, Scientific Applications." Allen, F.H. et al., eds. Bonn: Data Commission of the International Union of Crystallography, 1987:107–132.
19. Williams, R.L., Greene, S.M., McPherson, A. The crystal structure of ribonuclease B at 2.5 angstroms resolution. J. Biol. Chem. 262:16020–16030, 1987.