Improving contact predictions by the combination of correlated mutations and other sources of sequence information
Osvaldo Olmea and Alfonso Valencia

We have previously developed a method for predicting interresidue contacts using information about correlated mutations in multiple sequence alignments. The predictions generated with this method were clearly better than random but not enough for their use in de novo protein folding experiments. We assess the possibility of improving contact predictions combining information from the following variables: correlated mutations, sequence conservation, sequence separation along the chain, alignment stability, family size, residue-specific contact occupancy and formation of contact networks. The application of a protocol for combining these independent variables leads to contact predictions that are on average two times better than those obtained initially with correlated mutations. Correlated mutations can be effectively combined with other types of information derived from multiple sequence alignments. Among the different variables tried, sequence conservation and contact density are particularly relevant for the combination with correlated mutations.

Address: Protein Design Group, CNB-CSIC, Campus U Autonoma, Cantoblanco, Madrid 28049, Spain.

Correspondence: Alfonso Valencia
E-mail: valencia@cnb.uam.es

Electronic identifier: 1359-0278-002-S0025

© Current Biology Ltd ISSN 1359-0278

Introduction
Contact prediction by correlated mutations
Correlated mutations correspond to position networks in protein structures characterized by their coordinated pattern of change in multiple sequence alignments. These networks probably correspond to paths of sequence variation introduced during evolution to compensate for unfavourable interactions created by natural genetic drift or adaptation to new functions.

The intrinsic cooperativity of proteins makes a direct experimental approach to the study of correlated mutation networks difficult. Different experimental approaches indicate a wide range of possibilities for the introduction of compensatory mutations in protein structures. These approaches include: double cycle mutagenesis applied to pairs of positions in many different structural environments [1]; the study of the possible evolutionary pathways between similar proteins [2]; and screening of large libraries for functional complementation [3]. Different methods have been published for detecting correlated behaviour between positions of multiple sequence alignments [4–8]. It is generally assumed that compensation may be easier, and therefore more frequent, between spatially close residues; consequently, correlated mutations have been used to predict residue–residue contact and quantified in terms of physical distance [6,8,9]. The conclusions of different authors about the possibility of detecting correlated positions in multiple sequence alignments have been very different, which is not surprising considering the very different nature of the methods used to detect them (see [10] for a review).

In our formulation [6], correlated mutations are detected as pairs of positions with similar patterns of variation. Highly correlated positions are predicted to be residue–residue contacts and the accuracy of the predictions is assessed systematically. The ratio of correctly predicted contacts for a typical case, for example γ-cryostallin (4gcr, 174 positions in the family alignment), is 21 correct predictions out of 174 predicted contacts (12%), when only long-range sequence contacts are considered. These predictions are still far from perfect, representing a moderate improvement of around threefold over random predictions.

Practical applications of contact predictions by correlated mutations
The most direct use of contact predictions should be de novo protein folding using correlated positions as long-range sequence constraints in distance geometry related approaches. Unfortunately, the quality of the predictions is still far from making this approach straightforward. We have taken two other more indirect routes toward protein structure prediction with information from correlated mutations. First, we have investigated the combination of protein–protein docking methods with the prediction of neighbouring residues by correlated mutations. In test cases, the information about correlated mutations is able to select the correct docking solutions among many alternatives when enough sequences are available for both docking proteins (Pazos et al., unpublished data). Second, we have combined information from correlated mutations with threading methods. In this case, we have used contact prediction as a filter for alignments (implicit protein models) generated by standard threading methods.
In a significant number of cases, this hybrid approach leads to better detection of remote homologous proteins (Pazos et al., unpublished data).

In any of these cases or in direct de novo protein folding, improving the methods used for contact prediction will be of great value. We present here a first systematic approach toward the combination of correlated mutations with other sources of sequence information.

Combining correlated mutations with other information
Since the quality of predictions derived from correlated mutations is statistically significant but low, we seek to incorporate other types of information about interresidue contacts. In particular, statistics about residue contact preferences [12] and residue contact density [13–15] have been used previously with some success for contact prediction. Our aim is to include these and other sources of information under a single prediction protocol.

In this first approach, we carry out a systematic combination of correlated mutations with: (1) sequence conservation; (2) sequence separation along the chain; (3) alignment stability; (4) family size; (5) residue-specific contact occupancy; and (6) formation of contact networks. The protocol obtained for the combination of these variables leads to a more than twofold improvement in contact prediction.

Information contained in correlated mutations
The results obtained with our basic approach for the calculation of correlated mutations are represented in Figure 1. The observed decrease in the proportion of correctly predicted contacts with protein size is not surprising, since the difficulty of prediction increases with protein size. Indeed, the random chance of predicting contacts also decreases with protein size (the number of contacts increases linearly with protein size while the contact map grows quadratically).

In Table 1a, the results are summarized for four categories of protein size, with an average of 13% correctly predicted contacts for medium-sized proteins. The average numbers of correctly predicted contacts are 1.91, 3.96, 4.15 and 2.71 times better than random for the four different protein size ranges, respectively. Predictions are better for proteins between 100 and 300 residues. Smaller proteins in this set are in many cases nontypical globular proteins, e.g. disulphide-rich proteins, and among the proteins with >300 residues there are several cases of multidomain proteins. These may be the reasons for the poor performance of the method in these two size ranges.

A second observation (Fig. 1) is that the results obtained for different protein families are strongly scattered (see standard deviation [SD] in Table 1a). The main reason for this is the different nature of the underlying multiple sequence alignments.

The ‘sequence space’ occupied for a protein family can be defined in terms of the extension of the space covered (distance between sequences), density of sequences (number of sequences at different distances) and degree of sequence clustering (aggregation in subfamilies). In our data set, different protein families vary in size, from 15 to 224 sequences, with very different configurations of their sequence space, from very homogeneous families to many different subfamilies.
This intrinsic characteristic of the data represents a real problem that is difficult to compensate externally. We explore two variables: the size of the sequence space and the stability of the predictions toward changes in the number of aligned sequences.

**Range of sequence similarity covered by the multiple sequence alignments**

In Table 1b, we show how the proportion of correct predictions depends on the range of similarity covered. Alignments including sequences down to 30% similarity (as taken from the HSSP database) give better predictions than alignments covering only up to 50% sequence similarity and slightly worse than alignments going further down to >25% similarity. Therefore, small improvements in the predictions can be obtained by including more distantly related sequences, but the inclusion of very distant sequences, and consequently more uncertain alignments, brings only minor improvements.

The extension of the space has been demonstrated to be an important factor in secondary structure prediction [17], where including remote homologous sequences does improve predictions. In contrast, correlated mutations seem to have a behaviour more similar to accessibility predictions, where including distant sequences is beneficial but going down to remote homologies does not bring further improvements [14,18]. The different behaviour of accessibility and secondary structure predictions has been correlated with their different degree of conservation in distantly related proteins, the secondary structure being more conserved than accessibility [15,18]. Along these lines, it can be argued that contacts are only partially conserved among distant members of the same protein family and introducing very distant family members does not bring further improvement.

**Stability of the contact prediction toward changes in the sequence alignments**

One of the simplest ways of assessing the stability of the predictions toward changes in the alignments is a simple bootstrapping experiment. Figure 2 represents the number of different bootstrapping assays in which each one of all the possible pairs of positions are predicted as correlated. Results are split into two distributions: contacting and noncontacting pairs of residues. The most obvious difference between the two distributions is the accumulation of noncontacting pairs predicted as correlated only in a small number of the bootstrapping experiments. Thus, by disregarding those correlated mutations that appear in a
smaller number of repetitions, many wrong contact predictions can be avoided. Predictions obtained excluding those pairs of residues that appear as correlated in <80% of the bootstrapping experiments are on average 20% better, with more incidence on small and medium-sized proteins (Table 1c).

**Sequence distance**
Contact maps always present contacts distributed in clouds corresponding to the interaction of different secondary structure elements. We investigate the possibility of improving the predictions by considering neighboring residues along the sequence. Two different procedures were tried: averaging correlation values over sequence windows, and taking as prediction value the difference between the correlation value of a residue pair and the average of the window around it. The two procedures were tried out with different window sizes (3, 5, 7 and 9 residues), but we did not find improvements with any of these procedures (data not shown). The results for the averaging over a sequence window of five positions are presented in Table 1d.

**Combining sequence distance and correlation values**
We also attempted to search for a simple linear combination of correlation values and sequence distance. We did not find any linear combination of the two variables able to improve contact prediction (not shown, see Methods). As contact density decreases slightly after the first region of contact between neighboring secondary structure elements, it seems to be too weak for its application to the prediction of specific contacts. In other words, the information about their sequence distances is not a strong enough indicator to choose between pairs of correlated positions.

**Sequence conservation**
Conserved residues are often used in the analysis of multiple sequence alignments for detecting active sites and functionally important features. Other conserved residues are also expected to form part of conserved structural cores (see [19] for examples of both behaviours in the Ras-p21 family). In general, sequence conservation is larger in the protein core than in the external regions [20,21].

Sequence conservation can be used directly for contact prediction under the same protocol implemented for correlated mutations. In this case, pairs of residues are sorted by their average conservation value (defined as in [22], see Methods) and the sorted list is used for predicting contacts in the same way that we used the list of correlated positions.

The quality of the contact prediction by conservation or correlation independently are similar on average but they differ when different protein size ranges are compared (Table 1e). Conservation is a better predictor for small proteins (<99 residues in Table 1) than for medium-sized ones. The main reason could be the proximity of the structural core to the active site in small proteins. A similar observation on the predictive power of conserved residues has been made by Taylor and Hattrick [8]. It is interesting to note that the pairs of residues selected for their conservation and those selected by their correlation value are strictly different. The most obvious difference is the presence of the completely conserved residues, representing the dominant class in the prediction by conservation but excluded by definition of the calculation of correlated mutations. It is also true that other not completely conserved pairs of residues are essentially different between correlation and conservation. This is different from Taylor...
and Hatrick’s study [8], in which they found no significant difference between the pairs selected by their definition of correlation and sequence conservation. This is not surprising, since they used a different definition for correlated mutations.

Interestingly, predictions with conserved residues not only do not improve but even get worse when averaged over sequence windows (see Table 1e for a 5 residue window; window sizes of 3, 7 and 9 residues were also tried). In this regard, the behaviour of conservation is very similar to what happens with the correlation values (see above).

**Combining variability and correlation values**

In our approach, correlation and conservation are very different, and we attempt to combine them under a single prediction method. Sequence conservation is the most important factor in the analysis of multiple sequence alignments, and its combination with correlation information is a complex and promising subject that is certainly not concluded here, where only a first strategy is presented.

Completely conserved residues represent a technical difficulty since no correlation value is calculated for them. Under the procedure for combining a sorted list of variables, we chose to set completely conserved residues to a minimal value of correlation, 0. Despite this considerable downweighting of their importance in correlation, the optimization procedure overcomes this low value and in the final prediction many conserved residues score among the best predictions by conservation and correlation.

Figure 3 shows the values of the free parameter that are optimal for the linear combination of correlation and variability. The wide dispersion of values in the figure is in part a consequence of the different meaning of conservation in the underlying alignments, where the number of conserved residues goes from 0 to 144 in different families. Choosing a unique value for the linear combination of conservation and correlation is clearly not the best solution for many proteins and has to be seen as finding a middle ground for prediction purposes. For the average value of 0.3 (free parameter in the linear combination), the gain in prediction capacity is modest and more significant for small proteins (Table 1f). The fact that the combination of the two variables works better than either of them independently is an additional argument in favour of their relative independence, while the small size of the improvement obtained is indicative of their strong relation.

Our rationale for setting conserved residues to the minimal level of correlation is that in the absence of sequence variability it is impossible to find evidence of selection of compensating variants in some of the sequences. This idea can be challenged with the opposite view: conserved positions are the perfect place for finding cases of compensation and it is just that the sequences containing these examples have not yet been found.

In support of our view we tried the opposite experiment, giving a maximal value of correlation to pairs containing completely conserved residues. This procedure brings the completely conserved residues to the top of the conservation and correlation lists. The linear combination of both variables is in this case completely biased toward conservation, giving almost no importance to correlation. The final predictions obtained with this parameter are worse than for correlation or conservation alone (not shown). Therefore,
from a practical point of view, it is also better to consider completely conserved residues as noncorrelated.

It is interesting to note that the final list of pairs obtained by the combination of correlation and conservation does not necessarily contain all pairs of conserved residues. The contribution of conserved pairs is weighted by a factor of 0.7 and other, less conserved but more correlated, pairs may have higher overall values.

**Contact occupancy**

Any amino acid can make only a limited number of contacts (‘contact occupancy’), which are determined by its size, physical characteristics and its environment in the folded protein structure. The typical values of contact occupancy can be obtained from current protein structures by averaging for different residue types over different structural environments. Galaktionov and Marshall [23] have already used this type of information for contact prediction. In our case we have compiled contact occupancies for each residue class, secondary structure and degree of exposition (see Methods) for our database of 71 proteins. There are indeed clear differences between different residue types and environments (available as Supplementary material published with this paper on the internet).

In the following results we use contact occupancy as a filter for the predicted contact list. It has to be kept in mind that the assignment of amino acids to their structural classes is done using their real environment in the protein and not secondary structure and accessibility predictions. In a real scenario these assignments, derived from predictions, will be wrong in many cases and will decrease the value of the contact occupancy filtering. To minimize this error, we have selected those environments for which prediction methods work better with three secondary structure states (alpha, beta and coil) and only two accessibility classes (exposed and buried) defined with a 50 Å exposed surface cut-off.

Thus, the most effective way of applying the contact occupancy filtering is to exclude predicted residue pairs if any of the two residues in the pair has made more than its corresponding average number of contacts minus one standard deviation. The application of this rigid boundary leads to an underestimation of the typical number of contacts made for each residue, since they are not allowed to reach the average –SD value, with a reduction of the number of predicted contacts of ~25%. The decrease in prediction coverage is well compensated by a net gain in predictive power. Filtering the contact list predicted by the combination of correlation and conservation (see above) through contact occupancy leads to a clear and well-distributed improvement for all protein size classes (Table 1g).

**Contact networks**

In the list of correlated pairs, networks of concatenated pairs, i.e., pairs like A with B, B with C and C with A, are found very frequently. We investigated whether these networks are enriched in correctly predicted contacts. The set of residues contained in the largest connected clusters (cliques) were calculated as in [24] using as input the list of \( L/2 \) correlated pairs of each protein family (where \( L \) is the protein length). Pairs contained in clusters (Table 1h) contain slightly more correct predictions for small proteins than for other pairs. For big proteins, even if it is possible to find extended networks of contacts, they do not represent an increase in the number of correctly predicted contacts.

Filtering contact predictions by their inclusion in contact networks has one additional drawback: it considerably reduces the number of predicted contacts. This is especially disadvantageous when applied to small proteins in which the number of predicted contacts (half of the protein size) is necessarily small.

**Prediction in an independent test set**

The final procedure for the prediction of contacts involves the following steps: (1) selection of correlated mutations that occur in >80% of the bootstrapping experiments; (2) combination of the independent lists of correlated and conserved pairs; and (3) filtering of the combined list of predicted contacts with a contact occupancy criteria cut-off. Since the procedure involves a certain amount of training (derivation of the occupancy numbers and selection of the linear combination of variables), it is appropriate to use an independent test set. As such, we used the new nonredundant PDB files of August 13 1996 [25], which include a different selection of structures and many new files entered between August 1995 and August 1996. After selecting families with more than 15 alignments the final test list contains 71 new protein structures.

In Table 1i (and Supplementary material), it can be seen that the values obtained for the training and test sets are very similar for different protein size categories. Therefore, the gain in predictive capacity combining correlation, conservation, alignment stability and contact occupancy is stable and does not come from overtraining. The algorithmic combination of these variables represents a general improvement of more than twofold in the proportion of correctly predicted contacts. For the four different protein size categories of Table 1, the corresponding improvements over random values (3.79, 6, 5.47 and 6.26) are also clearly better than the values obtained with correlated mutations alone (1.91, 3.96, 4.15 and 2.71). The significant gain obtained with this first approach is a good sign for further work on the selection of different variables and new methods for combining them.
Methods

Data sets

Methods and parameters have been optimized with the ‘PDB-select list’ [25] of August 1995, which includes 474 proteins. From this list, those proteins with <15 alignments in the HSSP database of multiple sequence alignments [16] were excluded. The final list contains 71 proteins, corresponding to the PDB structures of 1aa1, 1aa2, 1ad1, 1aps, 1bba, 1ca5, 1caj, 1ccr, 1cgt, 1cr1, 1eaf, 1ec0, 1ezm, 1fns, 1fnd, 1ha1, 1ga1, 1gb1, 1hra, 1ijf, 1fpd, 1jx1, 1lcd, 1le4, 1mg1, 1mrt, 1ndk, 1nx1, 1ovf, 1osa, 1pdc, 1p1c, 1poo, 1pnp, 1ppt, 1rm, 1so1, 1sga, 1tik, 1zaa, 2bop, 2cas, 2cbh, 2cmd, 2cpl, 2ech, 2h1, 2h2, 2mrb, 2pf1, 2sn3, 2snv, 2spo, 2tgi, 2tmv, 3adk, 3b5c, 3cd4, 3chy, 3cla, 3egf, 3grs, 3pgk, 4enl, 4fgf, 4gcr, 5nn9, 5p21 and 8rxn.

Definition of parameters

Contacts are defined as Cα–Cα distance closer or equal to 8; for Gly, Cα atoms are taken. The quality of the predictions is quantified by the number of correctly predicted contacts divided by the total number of predicted contacts. This ratio is strongly dependent on protein size; since the density of contacts is larger in small proteins, it becomes easier to predict contacts in small proteins than in large ones. To make them comparable, results are always given in protein size categories.

Calculation of correlated mutations

Correlated mutations are calculated as in [6]. In brief, each position in the alignment is coded by a distance matrix. This position-specific matrix contains all the residue–residue distances between all possible pairs of sequences at that position. Distances between amino acids are defined by the scoring matrix of McLachlan [26]. The correlation value between each pair of positions is calculated as the average of the correlation for each corresponding bin of the position-specific matrices. Corresponding bins contain the distance between the same two sequences in the two positions under comparison. Correlated mutational behaviour between sequence positions i and j is defined as:

\[ r_{ij} = \frac{1}{N} \prod_{k} \left( \frac{\sum_{i,j} \left( \sigma_i - \bar{\sigma}_i \right) \left( \sigma_j - \bar{\sigma}_j \right) \exp(-b \cdot d_{ij})}{\sigma_i \sigma_j} \right) \]  

(1)

where \( \sigma_k \) is the standard deviation of \( \sigma_{ki} \) about the mean \( \bar{\sigma}_i \) and the indices \( k \) run from 1 to the number of sequences in the family (N).

Positions with >10% of gaps or completely conserved are not included in the calculation. Recently [27], we have introduced two minor variations to the method, first setting the similarity value of gaps to a dummy value of 0 (gaps are considered in position with <10% of gaps) and second, avoiding a previous step of normalization of the position-specific distance matrix. We demonstrated earlier that the proportion of correctly predicted contacts increases with correlation values [6], being meaningful only for the most correlated residues. It was also shown that the absolute value of correlation for each file is not a good indicator of the quality of the predictions, i.e. the same correlation value in two different protein families may contain very different numbers of predicted contacts depending on the maximum value of correlation in each one of the families. To avoid this problem we predict, for each protein family, a number of contacts proportional to their size, taken from the sorted list of correlation values. For this study, we use a safe cut-off, taking as many predicted contacts as half the length of the alignment, this value being a compromise between quality of the predictions and the number of predicted contacts.

Other variables derived from multiple sequence alignments

Sequence variability, accessibility and secondary structure were taken from the HSSP database [16]. Accessibility and secondary structure correspond to the DSSP definition [28]. In the HSSP definition [22], variability is 0 when positions in the multiple sequence alignment are completely conserved and it increases proportionally to the number of amino acid changes occurring at that position.

Selection of the best linear combination of variables

In order to combine the information from different variables, i.e. correlation and sequence conservation, we implemented a simple algorithm based on the combination of a sorted list of values generated by independent methods. First, the values obtained with any independent predictor (correlation, conservation or sequence distance) are normalized for all pairs of residues between 0 and 1. Second, each pair of residues gets a score that is the linear combination of the values of the normalized values of the two variables under consideration (i.e. correlation and sequence distance). The free parameter for the linear combination of the two variables is tried out exhaustively from 0 to 1 in intervals of 0.1, each one of the attempts producing a list of predicted contacts. The parameter that renders the best list of predicted contacts is chosen.

The quality of the list of predictions is evaluated in the following way. First, the list of residue pairs is sorted by the combined score. Then, the list of sorted pairs is scored by the number of correctly predicted contacts at the top of the list. In practice, the score assigned to a given list is the sum of the ordinal number of the pairs in the list. A perfect list will have all the true contacting pairs at the top of the list and its score will be the sum of the order number of the pairs starting from the first position of the list (value 1) to the position of the last pair representing a true contact (in this case the number of contacts). For example, the perfect list for a protein with four contacts will have a score of 10 (1+2+3+4).

The procedure is repeated for each one of the proteins in the training set and the optimal linear combination for each protein is analyzed in the search for the best common value for the combination of the two variables under analysis.

Supplementary material

Supplementary material (published with this paper on the internet) includes a table of contact occupancies for each residue class, secondary structure and degree of exposition for our database of 71 proteins. There is also a figure showing the average proportion of correct predictions for different protein sizes obtained for the training and test sets.

References