PROTEINS: Structure, Function, and Bioinformatics 56:679–684 (2004)

Protein Contact Prediction Using Patterns of Correlation

Nicholas Hamilton,[1]* Kevin Burrage,[1] Mark A. Ragan,[2] and Thomas Huber[1]
[1]*Advanced Computational Modelling Centre, Department of Mathematics, The University of Queensland, St. Lucia, Queensland, Australia*
[2]*Institute for Molecular Bioscience, The University of Queensland, St. Lucia, Queensland, Australia*

ABSTRACT **We describe a new method for using neural networks to predict residue contact pairs in a protein. The main inputs to the neural network are a set of 25 measures of correlated mutation between all pairs of residues in two "windows" of size 5 centered on the residues of interest. While the individual pair-wise correlations are a relatively weak predictor of contact, by training the network on windows of correlation the accuracy of prediction is significantly improved. The neural network is trained on a set of 100 proteins and then tested on a disjoint set of 1033 proteins of known structure. An average predictive accuracy of 21.7% is obtained taking the best *L*/2 predictions for each protein, where *L* is the sequence length. Taking the best *L*/10 predictions gives an average accuracy of 30.7%. The predictor is also tested on a set of 59 proteins from the CASP5 experiment. The accuracy is found to be relatively consistent across different sequence lengths, but to vary widely according to the secondary structure. Predictive accuracy is also found to improve by using multiple sequence alignments containing many sequences to calculate the correlations. Proteins 2004;56:679–684.**

Key words: protein structure prediction; predicted contact map; correlated mutation; neural network; CASP5

INTRODUCTION

Patterns of a Contact

A fundamental problem in molecular biology is the prediction of the three-dimensional structure of a protein from its sequence of amino acids. However, full molecular modeling to find the structure is at present intractable, and so intermediate steps such as predicting which residues pairs are in contact have been developed.

A variety of approaches to automated contact prediction have been taken. In RNA structure prediction, correlated mutation analysis was introduced with much success.[15] The same concept was then transferred to protein structures by Göbel et al. to predict contacts by finding correlated interchanges in multiple sequence alignments.[13] Likelihood matrix methods have also been applied to the problem.[8] There the idea was to use a large sample of proteins of known structure, and use these to to estimate the likelihood of contact of pairs of residues of given type. Pairs of contacts are then predicted by taking a multiple sequence alignment and summing the likelihoods of all residue pairs in the corresponding columns. Both the correlated mutation and likelihood approaches performed best when residues were local on the sequence, but tended to perform poorly on longer sequences where the residues were non-local on the sequence. Another approach to the problem has been to train neural networks with various encodings of multiple sequence alignments with other inputs such as predicted secondary structure.[12,11,10] These tend to perform better over a wide range of sequence lengths. Hidden Markov Models (HMM) combined with association mining rule techniques have also been successfully applied to the problem.[18] Filtering techniques where physically impossible configurations are removed from lists of predicted contacts have also been developed.[9,16]

One approach to contact prediction that does not appear to have been greatly exploited before is to looking for *patterns of contact*. If two (non-adjacent) residues are in contact then we would expect that the residues adjacent to those residues are also in contact with high probability. For instance, in an antiparallel β-sheet it might be that the fifth residue is in contact with the fifteenth, the sixth with the fourteenth, the seventh with the thirteenth, and so on. More complicated patterns of contact might form in the case of an α-helix in contact with a strand.

In the paper by Göbel et al.,[13] contacts were predicted by finding correlated interchanges of pairs of amino acids in multiple sequence alignments. However, while the prediction accuracy can be quite high for some proteins, generally predictions based on single pairs are poor. Here, we use a neural network approach to find patterns of correlation in combination with other inputs such as predicted secondary structure. The main inputs to the neural network are a set of 25 correlations mutations between two "windows" of size 5 centered on the residues of interest. Visualizing this set of inputs as a 5×5 matrix, each row corresponds to a residue in the first window, each column to a residue in the second window, and the entry in the

Kevin Burrage is a Federation Fellow of the Australian Research Council.

Mark A. Ragan is supported by the ARC Centre in Bioinformatics.

*Correspondence to Nicholas Hamilton, Advanced Computational Modelling Centre, Department of Mathematics, The University of Queensland, St. Lucia, Queensland, 4072, Australia. E-mail: nick@maths.uq.edu.au

Received 16 November 2003; Accepted 19 February 2004

Published online 14 May 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20160

matrix is the correlation between those residues as calculated in by Göbel et al. In the case of β-sheets, where two residues are in contact in adjacent strands, we might hope to see a pattern of high correlations on one of diagonals of the matrix, and lower correlations elsewhere. Which diagonal the high correlations would occur on would depend on whether the strands were parallel or antiparallel. The aim then is to train a neural network to use this information to predict when the middle two residues of the windows are in contact.

MATERIALS AND METHODS

Definition of a Contact

In the following we consider two residues to be in contact if the distance between their C_α atoms is less than 8 Å. There are a variety of measures of residues contact used in the literature. Some use the C_α distance,[6] while others prefer the C_β distance,[10,11] or even the minimal distance between the heavy atoms of the side-chain or backbone of the two residues.[12] Separations of less than 4.5 Å are also sometimes used to define contact. It is also common in the literature to exclude residue pairs that are separated along the amino acid sequence by less than some fixed number of residues; values of 0, 3, 6, and 10 have been used previously.[13,12,11,8] We chose only those pairs that are separated by at least four residues. Tests on over one thousand proteins showed that the C_α to C_α average distance between residues on the same helix, but separated by exactly four residues, was 8.665 Å with a standard deviation of 0.305 Å (data not shown). And so the majority of contacts that occur just because the residue pair are on the same helix are excluded by using the separation chosen here.

The input for our method to predict contact pairs is the amino acid sequence for a protein.

Generation of a Multiple Sequence Alignment and Predicted Secondary Structure

The Psipred[3,7] version 2.3 software by D. Jones is used to generate a prediction for the *secondary structure* as well as giving a pair-wise *multiple sequence alignment* for the proteins sequence. For each pair of residues in the protein sequence (subject to certain restrictions given below) we generate a pattern of inputs for a neural network.

Neural Network Inputs for Protein Contact Prediction

Pairwise correlations (25 inputs)

The multiple sequence alignment is used to calculate the (mutational) correlation between two columns of the multiple sequence alignment. The correlations are calculated as in Göbel et al.,[13] with the minor modification that the Blosum62 matrix rather than that of McLachlan[1] is used to score the residue interchanges. Our tests showed that the McLachlan and blosum matrices performed at approximately the same level as predictors of contact pairs of residues. For a given protein the correlations are normalized by subtracting the average pair correlation and dividing by a constant to bring the value into approximately the range $[-1,1]$. Columns that are either conserved or have more than 10% gap entries are excluded from the training set, while up to 40% gap entries are allowed for the test set of proteins.

Windows of length 5 of consecutive non-excluded columns are found. For each pair of non-overlapping windows the 25 correlations between columns of the first window with columns of the second are used as inputs to the neural network. The aim is to predict whether the middle residue of the first window is in contact with the middle residue of the second.

Residue classes (ten inputs)

Residues may be classified as non-polar, polar, acidic, or basic.[2] For a pair of residues there are ten possible pair cases, both non-polar, both polar, and so on. Thus we have ten binary inputs, exactly one of which is set to one to encode the residue type of the pair we are attempting to predict on.

Predicted secondary structure (18 inputs), (2 windows, length 3)

Given the input sequence, Psipred gives a predicted secondary structure for each residue as either helix, sheet, or neither. For a given residue, its predicted secondary structure type is encoded as three binary inputs, one of which is set to one. For a given residue pair that we are attempting to predict with, the predicted secondary structure is input for the two residues as well as the four residues that are adjacent to them. This gives a total of $2 \times 3 \times 3 = 18$ inputs.

Affinity score

A given residue pair is assigned an affinity score based on the type of each of the amino acids. From a training set of 50 proteins, the fraction f_{xy} of each residue pair $\{x,y\}$ type in contact was calculated (see supplementary material). For instance, of all the alanine-cysteine pairs in the 50 proteins the fraction in contact was 0.0326. For a given residue pair $\{x,y\}$ the affinity score is given by $30(f_{xy} - f_{ave})$ where f_{ave} is the average over all residue pair types. The normalization of subtracting the average and multiplying by 30 is to bring the value into approximately the range $[-1,1]$.

Length of input sequence and residue separation (two inputs)

The length of the sequence and the sequence separation, each divided by 1000, are input for the pair we are predicting with.

Network Architecture and Training

The predictor neural network is a standard feed-forward network, with 56 inputs as given above, ten hidden units, and a single output. The expected output is 1 for contacts and 0 for non-contacts. Experiments were performed with different numbers of hidden units (data not shown here) and ten units was found to be a good balance between generalization and over-training.

TABLE I. Average Prediction Accuracy by Sequence Length for the Best *L*, *L*/2, *L*/5, and *L*/10 Predictions[†]

	L	*L*/2	*L*/5	*L*/10
All (1033 proteins)	0.174 (7.70)	0.217 (9.69)	0.270 (12.34)	0.307 (14.11)
$0 \leq L < 100$ (262 proteins)	0.163 (3.33)	0.201 (4.12)	0.241 (4.90)	0.283 (5.81)
$100 \leq L < 170$ (296 proteins)	0.175 (5.69)	0.214 (7.07)	0.269 (9.12)	0.291 (9.94)
$170 \leq L < 300$ (268 proteins)	0.189 (9.62)	0.237 (12.20)	0.301 (15.64)	0.347 (18.32)
$L \geq 300$ (207 proteins)	0.169 (13.60)	0.213 (17.23)	0.271 (22.09)	0.308 (25.12)

[†]The bracketed numbers are the average of the ratios of the prediction accuracy to the random prediction accuracy, i.e. the improvement of the prediction over a random predictor.

Training and test proteins were randomly chosen from a representative set of proteins (pdb_select[14] Sept 2001) of the Protein Data Bank. Approximately 1600 proteins were first selected, then those with broken chains (i.e., those sequences in which the C_α to C_α distance of some pair of successive residues was greater than 5 Å) or less than 15 sequences in the generated multiple sequence alignments were removed, leaving 1133 proteins. A single chain was chosen from those PDB files with multiple chains. From the 1133 proteins, 100 were randomly chosen for training, and the remaining 1033 for testing. The training data was then a collection of 739,753 patterns of contact and noncontact from the set of 100 proteins. See the supplementary material for a list of PDB identifiers for the proteins.

The Stuttgart Neural Network Simulator[5] version 4.2 was used to train the neural network using standard back propagation with a momentum term. (It is perhaps worth noting that we found SNNS to be well-documented, easy to use, with a variety of useful features.) A variety of training schemes were tested and back propagation was found to perform best (data not shown). A validation set of 50 proteins was used to determine at what point to stop the training. On the final architecture, approximately 30 random weight initializations and trainings were run, and then the best performing network on the validation set was selected.

Of the 739,753 patterns used for training, 17,996 were contacts. In the papers of Fariselli et al.[12,11] *balanced* training is favoured, that is taking an equal number of contacts and noncontacts to train a neural network. However, our experiments with balanced training performed typically 2% worse than just training with all the patterns from a protein in the ratio of contacts to noncontacts as they naturally occur.

Testing the Trained Network

Once a "best" network was found, its performance was tested on a set of 1033 proteins from the Protein Data Bank, selected as described in the previous section. The test set is large enough to give statistically meaningful results that should generalize to other proteins. For closer comparison with results obtained by other groups we obtained PDB files for 59 of the proteins used in the recent CASP5 experiments, and tested the predictor on them. The CASP5 target identifiers are given in the supplementary material.

For a given target protein, we define the *prediction accuracy* A_N on N predicted contacts to be

$$A_N = N_c/N$$

where N_c is the number of the predicted contacts that are indeed contacts. In the following, N will typically be one of L, $L/2$, $L/5$, or $L/10$ where L is the length of the sequence. This follows the convention for evaluating protein contact prediction set out by the EVA project[4] (though they further distinguish several different minimal sequence separations for the pairs being predicted on, and we also do not report on the case $N = 2L$ since this is typically larger than the actual number of contacts).

For a given protein, the random prediction accuracy, A_R, is defined to be the fraction of the patterns generated that are contacts. The *improvement* of a set of predictions is then given by

$$\text{Improvement} = A_N/A_R.$$

The *coverage* (fraction of observed contacts predicted) is given as

$$\text{Coverage} = N_c/N_{obs},$$

where N_{obs} is the observed number of contacts.

RESULTS AND DISCUSSION

The results of the predictive accuracy of the neural network on the 1033 test proteins are presented in Tables I, II, III, and IV. In Table V results obtained on 59 proteins from the CASP5 assessment are presented.

In Table I the averages over the 1033 proteins of predictive accuracy on the best L, $L/2$, $L/5$, and $L/10$ predictions are given for each protein, where L is the sequence length. Over all 1033 proteins, for the best L predictions we obtain an average accuracy of 0.174 with a standard deviation of 0.096. For $L/2$, we obtain 0.217 ± 0.13; for $L/5$, we obtain 0.270 ± 0.18; and $L/10$, we obtain 0.307 ± 0.22. The relatively high standard deviation reflects the fact that the distribution is not normal and has a long trailing tail. Table Ib of the supplementary material gives the standard deviation for all the data in Table I. The average accuracy on the best $L/2$ predictions is often reported in the literature, and here an average accuracy of 0.217 is obtained over all 1033 sequences.

In Table I, the sequences are also separated according to their sequence length to give the average accuracy for sequences in certain length ranges. For direct comparison, the length bins are chosen to be in accordance with the results reported in the papers of Fariselli et al..[12,11] It can

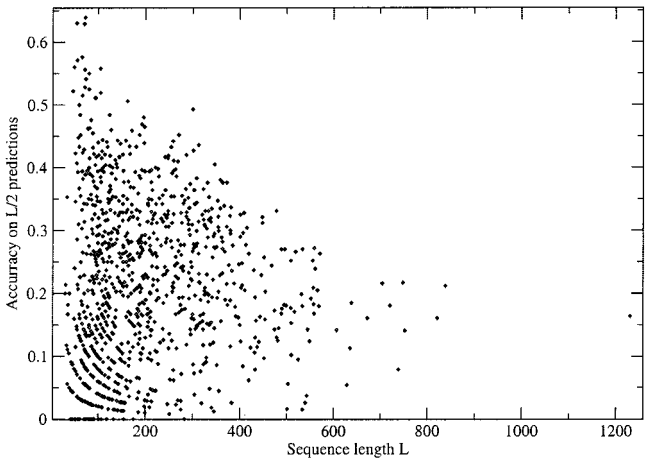

Fig. 1. Accuracy on best $L/2$ predictions versus sequence length L for the 1033 test sequences.

TABLE II. Average Prediction Accuracy by Sequence Length for Those of the 1033 Test Proteins for Which There Were at Least 100 Sequences in the Multiple Sequence Alignment

	L	$L/2$	$L/5$	$L/10$
All (434 proteins)	0.189 (9.56)	0.235	0.293	0.333
$0 \le L < 100$ (73 proteins)	0.191 (3.54)	0.237	0.280	0.321
$100 \le L < 170$ (85 proteins)	0.198 (6.39)	0.248	0.311	0.345
$170 \le L < 300$ (133 proteins)	0.200 (10.51)	0.247	0.312	0.356
$L \ge 300$ (143 proteins)	0.172 (13.63)	0.216	0.272	0.311

be seen that the predictive accuracy is relatively consistent across the different protein lengths (see also Figure 1.) We also give the average improvement over random prediction, that is the average ratio of the prediction accuracy to the overall fraction of residue pairs that are in contact for the protein. Overall, for the 1033 sequences, the average fraction of residue pairs in contact is 0.0305.

Of the 1033 proteins, 434 had more than 100 sequences in the alignment used to calculate the correlations. We would expect that a larger number of sequences in the alignment would give rise to more significant correlations. Hence, if the patterns of correlation approach is valid, we should see better predictive power for those proteins for which there are more sequences in the alignment. Table II shows that this is in fact the case. Comparing the average accuracy on the 1033 proteins with the 434 proteins that had more than 100 sequences in the alignment the improvement ranges from 1.5% to 2.6%.

The question then arises whether the improvement might have been obtained by chance. By taking 10,000 random subsets of size 434 of the 1033 proteins we can estimate the p-values for getting the predictive accuracies obtained in the first row of Table II. Of the 10,000 random subsets, none had accuracy greater than 0.189 for the best L predictions, none had accuracy greater 0.235 for the best $L/2$ predictions, three had accuracy greater than 0.293 for the best $L/5$ predictions, and four had accuracy greater than 0.333 for the best $L/10$ predictions. Hence the p-values are estimated to be at most 10^{-4}, 10^{-4}, 3×10^{-4}, and 4×10^{-4}, respectively, for these figures.

TABLE III. Average Prediction Accuracy on the 1033 Test Proteins by SCOP Secondary Structure Classification

	L	$L/2$	$L/5$	$L/10$
α (271 proteins)	0.097 (5.75)	0.126	0.166	0.201
β (248 proteins)	0.186 (6.04)	0.216	0.251	0.266
α/β (215 proteins)	0.213 (12.57)	0.272	0.352	0.397
α + β (199 proteins)	0.233 (8.35)	0.292	0.365	0.412

TABLE IV. Average Prediction Accuracy by SCOP Secondary-Structure Classification for those Proteins for Which There Were at Least 100 Sequences in the Multiple Sequence Alignment

	L	$L/2$	$L/5$	$L/10$
α (99 proteins)	0.120 (7.48)	0.156	0.207	0.236
β (89 proteins)	0.192 (7.34)	0.210	0.250	0.266
α/β (145 proteins)	0.211 (13.06)	0.268	0.340	0.385
α + β (64 proteins)	0.247 (9.09)	0.308	0.386	0.457

The secondary-structure type of the test proteins was obtained from the SCOP database. Of the 1033 proteins, 933 had a classification of either all alpha (α), all beta (β), alpha and beta proteins (α + β) or alpha and beta proteins (α/β). Table III presents the average prediction accuracy sorted by secondary-structure class. It can be seen that the predictive accuracy is highest (up to around 40%) for the mixed α and β cases, with the predictive accuracy on the all α case being less than half that of the mixed case.

Previous attempts at contact prediction have also found that prediction seemed to be particularly difficult for proteins whose secondary structure was of α-type.[12,11] In Fariselli et al.,[11] it was suggested that the poor predictions of their methods on this subset of proteins might be a result of the underrepresentation of α type proteins in their training set. However, in our case 20 of our 100 training proteins were of α type, and so here it appears not to be a problem of underrepresentation. We also attempted to train a network just on proteins of α-type to predict on α-type protein (data not shown), but no improvement in prediction accuracy was obtained. Given the details of our method, we might expect α-type prediction to be less accurate because the patterns of contact are less local on the sequence and require larger window sizes.

Of the 1033 sequences for which a SCOP classification was available, 662 sequences were not of α type. On these the average predictive accuracies were 0.209 (best L predictions) 0.257 ($L/2$), 0.318 ($L/5$), and 0.352 (L/10).

Table IV shows the effect of including only those proteins with more than 100 sequences in the multiple sequence alignment, on the predictive accuracy on each of the secondary structure types. The largest increases in accuracy occur when taking $L/10$ predictions, or on the all α type secondary structure. The increase in accuracy for the α type is as we might expect. The larger number of sequences leads to stronger correlations, and so the weaker patterns of contact generally obtained from helices are more easily recognised by the neural network.

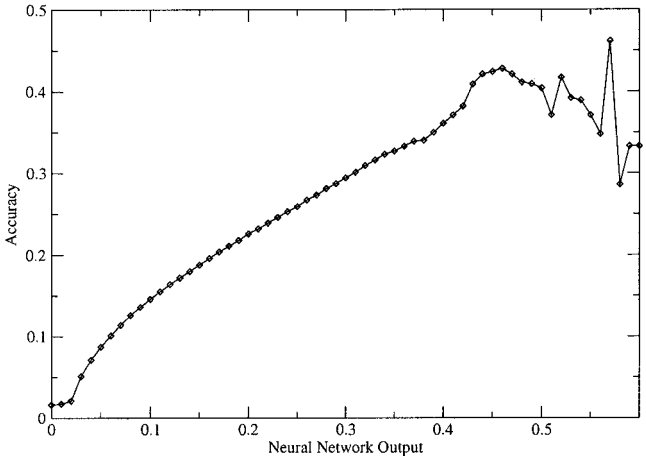

Fig. 2. Predictive accuracy of neural network output for 1033 test sequences. For a given neural network output, all the patterns from the 1033 proteins that have this network output or greater are found, and then the fraction of these that are contacts is plotted.

Given an output from our neural network for a given input pattern, it would be useful to be able to assign a probability that the residue pairs are in contact. In Figure 2, the average predictive accuracy for a given neural network output is shown. For a given neural network output, we find all of the patterns from the 1033 proteins that have this network output or greater, and then plot the fraction of these that are contacts. It is interesting to note that the peak predictive accuracy of 0.428 in Figure 2 is lower than the average accuracy obtained in some of our data sets. This suggests that rather than setting neural network output cutoff points to select which pairs of residues we predict are in contact, it is better to take the "best L/10" predicted contacts.

Comparison With Previous Contact Predictors

As far as the authors are aware, the CORNET predictor,[11] which is an extension of the work reported Fariselli and Casadio[12] and Olmea and Valencia,[9] claims to have the best contact prediction results to date. Their method is a neural network approach that involves encoding frequencies of residues in columns of a multiple sequence alignment, as well as having inputs based on predicted secondary structure, length of input sequence, and residue separation (which having read their work we also chose to use.) Overall this resulted in a rather large network having some 1071 inputs, eight hidden units and a single output.

On a test set of 29 proteins, making $L/2$ predictions on each, CORNET obtained an average accuracy of 0.14. As with our results, all α proteins were found to be difficult to predict, and if the 7 α's were removed from the dataset an accuracy of 0.16 was obtained.

It should be noted though that in Fariselli et al.[11] a residue pair is defined to be in contact if the C_β atoms are less than 8 Å apart, while here C_α separation is used. (Though our tests show that the neural network predicts C_α or C_β distance with close to the same accuracy.) And in Fariselli et al.[11] only residues that are separated by at least six residues are considered, while here residues are separated by at least four residues. By way of comparison, using our network to predict to C_β distances and with the restriction that residues that are separated by at least six residues, the neural network here gives average predictive accuracies of 0.165 (best L predictions) 0.205 ($L/2$), 0.255 ($L/5$), and 0.288 (L/10), on the 1033 test proteins.

In Table V results of testing the predictor on a set of 59 proteins from the the CASP5 assessment that ran in 2002 are given. Note that we did not participate at the time in CASP5, but the results of the our predictor are presented in an attempt to give a comparison with other predictors on a standard data set. Also note that there were 78 target proteins in CASP5, but only 59 were available to us. For a summary of the results of other groups on CASP5 proteins see Table VI in Aloy et al.[17] The results in Table V of this paper are presented to allow direct comparison. In particular C_β to C_β residue separations are predicted as in the CASP5 experiments. The first column gives the number of predictions made per protein. These predictions are then divided according to the residue separations, and the average accuracy and coverage over the 59 proteins is then given for each class. It can be seen that the average predictive accuracies are substantially higher than those obtained for the 1033 test proteins. This appears to be a consequence of a relatively high number of contacts in the 59 CASP5 proteins. On average, just over 4% of pairs separated by four or more residues were in contact in the CASP5 set, compared to an average of 3% in the 1033 test proteins. In fact, the average improvement in prediction for the best L predictions was 6.55 for the CASP5 set, compared to 7.7 for the 1033 proteins.

Of the six groups participating in the CASP5 experiment for contact prediction, the patterns of contact approach has generally better accuracy and coverage than all but two of them: the GeneSilico (517) and Bujnicki-Janusz (020) groups, which achieved substantially higher accuracy for approximately the same degrees of coverage (see Table VI in Aloy et al.[17]). But since both of these approaches involved generation of multiple models of the three-dimensional structure of each protein together with human intervention, this is not unexpected.

CONCLUSION

Fariselli et al.[11] state that their goal is to obtain an accuracy of 50%, because then the folding of a protein of less than 300 residues length could be reconstructed with good accuracy (within 0.4-nm RMSD). While we are still short of this aim, it can be seen that under restrictions such as having a large number of sequences in the multiple-sequence alignment and sequences being of particular secondary-structure type, the patterns of correlation approach is a step closer to this goal.

One approach to improving predictive accuracy would be to find a better measure of correlated mutations. In Singer, Vriend, and Bywater,[8] a new method using likelihood scores is described to find correlated mutations. This method appears to give better results than the correlated

TABLE V. Average Predictive Accuracy and Coverage on 59 Proteins From the CASP5 Experiments[†]

	All pairs		Middle range		Long range	
Predictions/protein	Acc	Cov	Acc	Cov	Acc	Cov
L	0.210	0.082	0.186	0.152	0.210	0.081
L/2	0.255	0.050	0.229	0.094	0.254	0.051
L/5	0.301	0.023	0.273	0.044	0.300	0.023
L/10	0.321	0.012	0.261	0.025	0.313	0.011

[†]The first column gives the number of predictions made per protein. The data is divided according to the residue separations: greater than or equal to five (All pairs); between five and eight (Middle range); greater than or equal to nine (Long range). The average accuracy (Acc) and coverage (Cov) over the 59 proteins is then given. Note that here we predict C_β to C_β residue separations.

mutation method of Göbel et al. It will be interesting to see if it can be used as a replacement for the correlated mutations in our method to improve overall prediction.

In Olmea and Valencia,[9] the method of *contact occupancy filtering* is described. It uses the fact that an amino acid can have only a limited number of contacts, to filter out physically impossible configurations. This leads to an approximately 25% reduction in the number of predictions, but can improve predictive accuracy by several percentage points.[10] Currently we are investigating the method with the aim of providing improved accuracy on a reduced set of predicted contacts.

Contact Prediction Server

A contact prediction server implementing the patterns of contact approach is available at *http://foo.maths.uq.edu.au/~nick/Protein/contact.html*. All data sets used in this work are also available there.

ACKNOWLEDGMENTS

The authors gratefully acknowledge financial support from the University of Queensland. Kevin Burrage acknowledges the support of the Australian Research Council as a Federation Fellow.

REFERENCES

1. McLachlan AD. Tests for comparing related amino acid sequences. J Mol Biol 1997;61:409–424.
2. Krane DE, Rayner ML. Fundamental Concepts of Bioinformatics Benjamin Cummings; 2003.
3. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
4. Koh Ingrid YY et al. Eva: evaluation of protein structure prediction servers. Nucleic Acids Res 2003;31:3311–3315.
5. Zell A et al. Stuttgart neural network simulator user manual version 4.2. University of Stuttgart, 1998.
6. Mirny L, Domany E. Protein fold recognition and dynamics in the space of contact maps. Proteins 1996;26:319–410.
7. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics 2000;16:404–405.
8. Singer MS, Vriend G, Bywater RP. Prediction of protein residue contacts with a PDB-derived likelihood matrix. Protein Eng 2002;15:721–725.
9. 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.
10. Fariselli P, Olmea O, Valencia A, Casadio R. Prediction of contact maps with neural networks and correlated mutations. Protein Eng 2001;14:835–843.
11. Fariselli P, Olmea O, Valencia A, Casadio R. Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. Proteins 2001;Suppl 5:157–162.
12. Fariselli P, Casadio R. Neural network based prediction of residue contacts in protein. Protein Eng 1999;12:15–21.
13. Göbel U, Sander C, Scheider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18:309–317.
14. Hobohm U, Scharf M, Schneider R, Sander C. Selection of a representative set of structures from the Brookhaven protein data bank. Protein Sci 1992;1:409–417.
15. Winker S, Overbeek R, Woese CR, Olsen GJ, Pfluger N. Structure detection through automated covariance search. Comput Appl Biosci 1990;6:365–371.
16. Shao Y, Bystroff C. Predicting interresidue contacts using templates and pathways. Proteins 2003;53:497–502.
17. Aloy P, Stark A, Hadley C, Russell RB. Prediction without templates: new folds, secondary structure, and contacts in CASP5. Proteins 2003;53:436–456.
18. Zaki MJ, Shan J, Bystroff C. Mining residue contacts in proteins using local structure predictions. IEEE Transactions on Systems, Man and Cybernetics 2003;33:789–801.