An Introduction to Protein Contact Prediction

Nicholas Hamilton
Institute for Molecular Bioscience and Advanced Computational Modelling Centre,
The University of Queensland,
St. Lucia,
Queensland, 4072,
Australia.
n.hamilton@imb.uq.edu.au

Thomas Huber
Department of Biochemistry,
The University of Queensland,
St. Lucia,
Queensland, 4072,
Australia.
huber@maths.uq.edu.au

1 Introduction

The evolutionary analyses of a protein’s history has often seem irrelevant to protein structure prediction. Indeed, a protein’s ability to fold will depend entirely and only on the underlying physics which is dictated by the amino acid sequence and its environment. It will not make any difference whether the protein sequence was designed *de novo*, resulted from random shuffling experiments or has evolved naturally.

However for the purpose of protein structure prediction, the evolutionary context of a protein provides an important additional layer of information that one can employed to increase success. In the most simplistic form, a method may produce predictions from average (or consensus) properties from a family of related proteins instead of using a single member. Noise due to variations in individual proteins is then reduced in the prediction and accuracy increases. Most if not all of prediction methods today take advantage of this fact and use so-called *profiles* from homologous sequences (for examples see chapters xx), some methods even go further and average over similar but not necessarily homologous protein structures (chapter yy).

In this chapter we review some of the ideas of more sophisticated evolutionary information in protein sequences that are important to structure and their application to protein contact prediction. In the next section the general methods for using multiple sequence alignments to make contact predictions are covered. In general it is found that predictors, while performing well above chance levels, will make predictions that are not in fact physically realizable. Hence filtering methods are often used to improve predictions, and these are described in section 3. It is fair to say that the contact prediction literature has suffered from predictors being tested on different data sets using varying measures of predictive quality, making comparison difficult. For this reason blind tests of methodologies on a common data set such as the CASP experiments are invaluable. Section 4 begins by giving the more common measures of predictive quality and then presents results of the state of the art predictors participating in the most recent round of CASP experiments.

2 Sequence based approaches to contact prediction

A variety of approaches have been taken to contact prediction using multiple sequence alignments (MSA's) and other information. At the most fine grained level a multiple sequence alignment is constructed for a given protein and the information in pairs of columns is used to predict which residues are in contact. The fundamental assumption is that for residues that are in contact the corresponding columns will be in some way more highly related to each other than for those residues that are not.

Several measures of the relatedness of MSA columns have been developed. One approach is to look for *correlated mutations* in the MSA [1]. The idea is that if residue substitutions in one column of a MSA are correlated to those in another, then one reason for this may be that the residues are physically close. For instance, substitutions might occur in pairs so as to preserve the total charge in a region and so maintain the structural integrity of the protein. To calculate a mutational correlation score for a pair of columns in an MSA a measure of the physiochemical similarity of any pair of residues is required, and for this the McLachlan matrix [2] is often used. For each column of the MSA of n sequences, an n by n matrix is constructed with entries (i,j) being the McLachlan interchange score for the i^{th} and j^{th} residues of the column. The correlation mutation score between two columns is then calculated as the standard Pearson correlation between the entries of their matrices. Hence the correlation is found between the physical similarity scores for residues substitutions at pairs of sites on the protein. A weighting scheme is also used in the correlation calculation to favour those row pairs of the MSA that are least similar. Column pairs are then ranked according to their correlated mutation score, those with high scores deemed as being more likely to be in contact. Formally, the correlation score between columns i and j is given as

$$r_{ij} = \frac{1}{N^2} \sum_{k,l} \frac{w_{kl}(s_{ijk} - \bar{s}_i)(s_{jkl} - \bar{s}_j)}{\sigma_i \sigma_j}$$

Where N is the length of the sequence alignment, w_{kl} is a weighting function that measures the similarity between rows k and l of the alignment, s_{ijk} is the McLachlan interchange score for the j^{th} and k^{th} residues in column i, and $\bar{s}_i$ and σ_i are the mean and standard deviation of the interchange scores in column i.

Rather than using a general “physical similarity” score, research has also gone into exactly what physical factors effect compensatory mutations, and to look for *biophysical complementarity principles.* For a given physical quantity, such as side chain charge, the correlation between the values of this quantity for pairs of residues in two columns may be calculated [3, 4]. In this way it has been found that for residues in contact compensatory mutations are highly correlated with side chain charge. In other words for residues in contact, the sum of their side chain charges tend to be preserved if the pair mutate. In contrast it has been found that there is little correlation between side chain volume and compensatory mutations.

At the simplest level the likelihood of a given residue pair being in contact may be predicted based on empirical observations of how often such a pair has been observed to be in contact in other proteins of known structure. *Contact likelihood* tables for all residue pairs based in a set of 672 proteins of known structure have been constructed [5], and these have been used to construct a contact likelihood score for columns of an MSA by summing the likelihoods of all residue pairs in the columns. Similarly, Markov models have been built of mutations for residues in contact [6], and these may also be used as a (crude) predictor of when residues are in contact.

On a more abstract level, *information theory* has been applied to contact prediction from MSA's [7-9]. Here the *mutual information* between two columns of a MSA is used as estimate of contact likelihood and is defined as

$$\sum_{(a_i,a_j)} P^2_{a_i,a_j} \log(P_{a_i,a_j} /(P_{a_i} P_{a_j}))$$

Where a_i and a_j are the amino acids found in columns i and j, P_{ai} is the observed relative frequency of amino acid a_i in column i, similarly for a_j in column j, and $P_{ai,aj}$ is the observed relative frequency of the pair (a_i,a_j). This is based on the definition of entropy from information theory, and gives an estimate of the degree to which identity of one amino acid in one column allows us to predict an amino acid in another. The column pairs with the highest degree of mutual information are then those predicted to be in contact.

Many of these measures have also been useful in predicting the function of groups of residues [7, 10]. While each approach has been moderately successful in that they will usually predict contacts at a level significantly above random chance, the accuracy of the predictions is generally quite low and certainly not high enough to enable accurate reconstruction of the three dimensional structure of a protein from its sequence. One essential problem is that phylogenetic relationships are disregarded and these can lead to strong correlations between physically distant residues. Several attempts have been made to separate out phylogenetic covariation [4, 9, 11-13], but predictive accuracy remains in need of improvement.

Interest has hence turned to *combining prediction methods* and other information to improve predictions. Other sources of information include predicted secondary structure [14], residue sequence separation, sequence length, residue conservation, hydrophobicity of residues, predicted disulphide bridges and conservation weight. While each of these is not a good predictor of protein contacts in itself, the idea is that by combining relatively poor predictors a better predictor may be made[1]. Windows around the residue pairs being predicted on have also been shown to improve accuracy. Here information is also provided such as correlation scores, residue frequencies in columns of the MSA and predicted secondary structure on the residues near to the pair being scored. This improves prediction accuracy since if two residues are in contact then it is also likely that those residues around them will also be in contact and correlate in some way.

Once a set of measures has been chosen, the problem then becomes how to combine the information into a single prediction score for a given protein and residue pair. Virtually every type of learning algorithm has been applied to the problem such as neural networks [16, 17] [18-20], self organizing

[1] In the theory of boosting (15. Shapire RE, *The boosting approach to machine learning: An overview*. MSRI Workshop on Nonlinear Estimation and Classification. 2002: Springer.) it has been proved that weak predictors may be combined to provide a predictor that is significantly more accurate than any of the weak predictors, though the boosting method appears not to have yet been exploited in contact prediction.

maps [21], support vector machines [22-24], and hidden Markov models (HMMs) [25] [26, 27]. Typically an algorithm will learn the problem on a training set of data, and then be tested on an independent test set.

Each of these learning methods has its own advantages and disadvantages, and it might be said that training a learning algorithm is something of a black art. Some claim that neural networks have a tendency to over-train, that is to fit the training data to closely and hence lose their ability to generalize on unseen data sets, though this can be to some extent avoided by the use of a validation set to halt the learning during training, while support vector machines suffer less from this problem. Also, support vector machines are sometimes said to be more tolerant of noisy data [28]. Balanced training is often favoured, that is to use equal numbers of each category of data in training [18, 20], while others have obtained better results training with the proportions that the categories naturally occur in [19]. Encoding of inputs is certainly important. In one study of predicting disulphide connectivity it was found that by taking the log of the sequence separation (with other inputs) of the residues the predictive accuracy of the SVM improved by 4% over simply using the linear sequence separation [29], and choice of window size varied the predictive accuracy by up to 10% (larger window sizes were also found to increase accuracy in [21]). In general, training a good predictor involves much testing of encoding schemes and experimenting with data as well a little good luck.

3 Contact filtering

In many cases a contact predictor will incorrectly predict residue pairs to be in contact. For this reason contact filtering is often applied to a given set of predicted contacts, the aim being to remove those predicted contacts that are in some way physically unrealizable.

The simplest and perhaps most effective method of filtering is *contact occupancy*. For a given residue in a protein sequence there is a limit to the number of residues with which it may be in contact. This number will vary according to the type of residue, the secondary structure around the residue, whether the residue is exposed on the surface of the protein and so on. By examining proteins of known structure, tables of the average number of contacts a given residue of a particular class has may be created. Linear regression model [30], support vector machine [31] and neural network [32] approaches to predicting contact occupancy have also been developed. A list of predicted contacts may then be filtered by removing those pairs for which one or both of the residues is in predicted to be in contact with more than the average for its class. Significant improvements of up to 4% in predictive accuracy have been found using this methodology [20, 33]. Similarly, for any pair of residues that are in contact there is a limit to the number of residues that may be in contact with both of the residues, and this may also be used as a filter [27].

Bootstrapping may also be applied to assess the stability the prediction of contact pairs. Olmea and Valencia performed bootstrapping experiments by excluding 10% of sequences in the alignments given to their prediction method [33]. By excluding predicted contact pairs that occurred in less than 80% of bootstraps a 20% improvement in accuracy was obtained.

More detailed consideration of the (predicted) secondary structure can also be applied. For instance, within an helix, the i^{th} residue should only be in contact with residues i+4 and i-4; a residue can not be in contact with residues on opposite sides of a helix; and within a single strand of a β-sheet only adjacent residues should be in contact [27]. A more direct method to checking the physical realisability of contact maps is to align fragments of a predicted contact map to template contact maps, where the templates are fragments of contact maps of proteins of known structure. The predicted contact map fragment then becomes the contact map of the most closely aligned template [27]. However, care needs to be taken with the definition of contact and the application of these rules and maps since, say, a 4Å cutoff for contact will give a very different pattern of contact than that for an 8Å cutoff.

5 Evaluating contact predictors and the CASP6 experiment

There are many definitions of residue contact used in the literature. Some use the C-α distance, i.e. the distance between the α carbon atoms of the residue pair [34], while others prefer the C-β distance [20, 35] or even the minimal distance between the heavy atoms of the side chain or backbone of the two residues [36]. The most common minimum separation used to define a contact pair is 8Å. It is also usual to exclude residue pairs that are separated along the amino acid sequence by less than some fixed number of residues, since short range contacts are less interesting and easier to predict than long range ones.

For a given target protein, the *prediction accuracy* A_N on N predicted contacts is defined to be $A_N = N_c/N$ where N_c is the number of the predicted contacts that are indeed contacts for a given minimum sequence separation. Typically N is taken be one of L, L/2, L/5 or L/10 where L is the length of the sequence. For most proteins, the actual number of contacts (using the 8Å definition) is in the range L and 2L. It has become relatively standard to report results on the best L/2 predictions with a maximum distance of 8Å between C-β atoms (C-α for glycine), with a minimum sequence separation of 6. This standardization is in large part thanks to the CASP ([37]) and EVA ([38]) protein structure prediction blind tests, and has been invaluable in enabling comparison between predictors.

The *prediction coverage* is defined to be N_c/T_c, where T_c is the total number of contacts pairs for the protein. The *random accuracy* is given by the fraction of all residue pairs that are in contact (for a given sequence separation), and gives a measure of the probability of picking a pair to be in contact by chance alone. The *improvement over random* is then prediction accuracy divided by the random accuracy, and gives a measure of how much better than chance the predictor performs. This can be a useful measure since the number of contacts can vary widely between proteins, and prediction accuracy may be artificially high due to an unusually large number of contacts in a given protein.

Another measure that is sometimes used is the *weighted harmonic average distance* [39]

$$X_d = \sum_{i=1}^{15} \frac{P_{ip} - P_{ia}}{15 d_i}$$

Where the sum runs over 15 distance bins in the range 0 to 60Å, d_i is the upper bound of each bin, normalized to 60, P_{ip} is the percentage of predicted pairs whose distance is included is included in bin i, and P_{ia} is the same percentage for all pairs. The harmonic average is designed to reflect the difference between the real and predicted distances of residue pairs: when the average distance between predicted residue pairs is less than the average distance between all pairs in the structure then $X_d>0$, though interpreting the meaning of a particular value of X_d can be difficult.

Prediction accuracy and coverage are the most commonly reported measures in the literature. However, the choice of sequence separation can greatly affect the prediction accuracy since residues that are close on the sequence are more likely to be in contact. Choosing a minimum sequence separation of 12 instead of 6 may well reduce the accuracy by 50% or more depending on the characteristics of the predictor. Similarly, the accuracy is usually strongly dependant on the number of predictions made. A predictor that has an accuracy of 0.3 on its best L/5 predictions may well drop to 0.15 on its best L predictions. Also, a contact predictor that does relatively well on one data set may predict poorly on another, perhaps due to there being many proteins in the first data set for which several homologous structures are known. For these reasons it can be difficult to evaluate the relative performance of contact predictors in the literature.

To overcome these problems standardized blind tests of protein structure and protein contact predictors have been introduced. The Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiments are run biannually and involve releasing over several months the sequences for a set of proteins for which the structure has been solved, but are not yet publicly available [32, 37]. Groups from around the world then submit their predictions, and these are independently evaluated by a team of experts and the results published. The experiment also includes an automated section where sequences are submitted to prediction servers and predictions returned without human intervention. Similarly, the EVA project provides a continuous, fully automatic analysis of structure prediction servers [38]. Both EVA and CASP include sections for comparative 3D modeling, fold recognition, contact and secondary structure prediction.

In the 6th round of CASP in 2004 there were 87 target proteins released, and 16 groups competed in the contact prediction category, 5 of which were registered as automated servers [37]. Unfortunately, the only published evaluation performed was for the 11 hard new fold (NF) targets for which additional and structural information was not available [37]. These targets are not representative of the wide range of possible protein folds, and with such a small sample set it is difficult to evaluate the effectiveness of each contact predictor accurately. Fortunately the raw prediction scores for each contact predictor are available from the CASP6 web site[2], and so we can present results for the full set of targets.

In Table 1, average accuracy and coverage results are shown for the contact predictors that submitted to CASP6. The data shown is for the best L/2 predictions with a minimum sequence separation of 6 residues. The tables are separated according to the target type. Not all predictors submitted for all targets, and so the averages presented are over those proteins for which a prediction was submitted. The number of targets predicted by each group is also shown. For most purposes, accuracy is more important than coverage, since the aim is to get a number of high quality contact predictions.

Several groups attained accuracy of 20% or better on most of the classes of protein. Here we emphasize those that do not involve 3 dimensional modeling, or in which 3 dimensional modeling incorporates a contact predictor. For more information on the predictors, see also the CASP6 methods abstracts available from http://predictioncenter.org/casp6/abstracts/abstract.html

RR100 Baker [37, 40]: The Baker predictor is particularly interesting in that while it takes a whole structure 3D modeling approach, a contact predictor is integrated into the structure building for fold recognition targets. The approach to contact prediction is to train several neural network on the predictions made by a set of 24 protein (3D) structure predictors that participated in recent LIVEBENCH experiments [41]. For a given residue pair, the principle input to a neural network is the ratio of the number of servers that predict the residues to be in contact (contact meaning closer than 11Å), along with other inputs such as secondary structure prediction and amino acid property profiles. Ten neural networks were trained and validated on different subsets of the same training set. For prediction, the average score of the 10 trained networks is taken, and the highest scoring residue pairs taken as predicted contacts. The consensus contact predictor is then used as an indictor of distant contacts that should be present in the *de novo* predicted models.

www.jens-meiler.de/contact.html

[2] CASP6 website http://predictioncenter.org/casp6/Casp6.html

		All targets							Comparative modelling targets						
group		#submitted	accuracy		coverage		Xd		#submitted	accuracy		coverage		Xd	
			av	stddev	av	stddev	av	stddev		av	stddev	av	stddev	av	stddev
RR011	MacCallum	83	15.5	8.4	4.4	2.5	7.4	3.3	41	15.0	8.0	4.2	2.3	8.1	3.3
RR012	GPCPred	82	19.3	11.5	6.9	4.6	9.3	3.8	40	21.0	12.1	6.2	3.2	10.5	3.4
RR018	Baldi	31	33.6	15.3	10.7	9.2	14.1	4.7	14	32.0	16.0	11.7	12.6	13.7	5.1
RR019	Baldi-server	85	36.8	17.3	10.0	9.0	15.4	5.0	42	36.2	15.5	9.9	8.8	15.9	4.6
RR088	Bystroff	34	13.4	11.7	6.4	8.2	5.6	4.8	11	17.0	13.9	9.7	12.3	7.0	3.8
RR089	KIAS	85	15.3	13.3	3.2	2.5	7.3	4.2	42	15.0	10.1	3.3	2.3	7.8	3.8
RR100	Baker	83	40.1	22.2	19.9	14.4	15.8	8.0	41	52.1	20.8	24.0	15.2	20.0	7.2
RR166	SAMT04-hand	43	20.8	11.8	6.3	3.4	9.5	4.3	30	23.3	12.5	7.0	3.5	10.7	4.0
RR185	Huber-Torda	75	38.3	30.3	17.0	18.9	13.8	10.2	41	57.5	25.2	24.5	19.1	20.3	8.3
RR301	rostPROFcon	85	28.3	13.5	13.5	9.4	12.7	4.8	42	31.3	14.4	13.3	9.9	14.2	5.0
RR327	Hamilton-Huber-'	65	24.0	16.2	5.6	3.8	10.7	5.9	34	26.3	14.4	6.6	3.7	12.4	5.1
RR348	Distill	81	9.1	7.8	5.9	9.4	5.0	3.8	41	8.8	7.9	6.5	11.9	5.3	3.3
RR361	karypis	74	14.7	11.9	4.6	4.5	10.4	3.8	34	16.6	14.5	4.7	4.4	11.6	3.6
RR491	cornet	72	4.7	6.3	1.0	1.4	7.1	4.1	36	5.2	7.2	1.0	1.3	7.2	4.6
RR545	cracow.pl	19	8.8	6.4	2.4	1.8	3.4	3.7	8	9.3	6.0	2.2	1.8	5.1	2.8

		Fold recognition targets							New fold targets						
group		#submitted	accuracy		coverage		Xd		#submitted	accuracy		coverage		Xd	
			av	stddev	av	stddev	av	stddev		av	stddev	av	stddev	av	stddev
RR011	MacCallum	31	16.4	9.1	4.9	2.9	6.6	3.4	11	14.4	7.3	3.9	1.8	6.7	2.4
RR012	GPCPred	31	16.6	9.9	7.8	5.9	7.6	3.9	11	20.6	11.5	7.0	4.7	9.6	3.3
RR018	Baldi	14	31.6	12.9	9.6	4.7	13.9	4.4	3	50.2	12.3	10.8	4.5	16.6	1.8
RR019	Baldi-server	32	35.8	19.6	11.2	10.3	14.4	5.3	11	41.7	16.2	6.6	3.7	16.5	5.2
RR088	Bystroff	18	12.8	10.8	4.8	4.4	5.5	5.3	5	8.0	5.7	4.7	4.3	2.8	2.9
RR089	KIAS	32	16.3	17.8	3.4	3.1	6.8	5.0	11	13.7	7.0	2.2	1.2	6.9	2.8
RR100	Baker	31	32.0	16.9	18.1	12.7	12.7	6.8	11	18.3	10.7	9.7	8.0	8.7	4.3
RR166	SAMT04-hand	9	12.2	6.7	4.0	2.6	5.5	3.6	4	21.6	2.2	6.1	1.4	9.9	1.9
RR185	Huber-Torda	29	15.3	17.8	8.7	15.3	5.9	6.2	5	14.5	7.3	4.3	1.4	6.9	3.4
RR301	rostPROFcon	32	25.8	13.1	14.6	9.8	11.1	4.7	11	24.2	6.5	10.8	4.3	11.4	2.1
RR327	Hamilton-Huber-'	22	21.7	18.1	5.1	4.2	8.7	6.5	9	20.5	16.1	3.5	2.1	9.6	5.6
RR348	Distill	29	9.7	8.5	5.8	6.5	5.0	4.6	11	8.5	4.1	4.0	2.2	4.1	3.0
RR361	karypis	30	13.6	9.3	4.7	4.7	9.5	3.6	10	11.7	7.4	4.0	3.8	9.3	3.7
RR491	cornet	26	4.1	5.6	1.1	1.5	7.3	3.8	10	4.9	3.8	0.9	0.8	6.1	2.0
RR545	cracow.pl	9	8.4	7.0	2.6	2.0	1.7	3.9	2	8.2	3.9	2.3	1.0	3.6	2.5

Table 1 Performance results from all contact predictors submitting to the CASP6 experiment for L/2 predicted contacts and a minimum separation of six residues along the sequence.

RR185 Huber-Torda [42]: The Huber-Torda predictor is not a dedicated contact predictor but builds 3D models by threading, which combines structure and sequence based terms for scoring alignments and models. Protein contacts are extracted from the models in a post processing step. It is interesting to observe the performance of a threading method that is based on a fundamentally different philosophy than protein contact predictors, since it shows limitations of the methods and may suggest new ways to improve overall performance.

http://www.zbh.uni-hamburg.de/wurst

RR019 and RR018 Baldi-server and Baldi [43, 44] : Similarly to the Baker group, the Baldi group predictors are whole structure 3D modelers that incorporate contact predictions. The energy function used in constructing the 3D coordinates incorporates a contact map energy term that “encourages” the models to follow the predicted contact structure. The contact predictor is a 2D recursive neural network in which outputs feed back as inputs [45]. The recursive approach allows local information to be combined with more distant contextual information to provide better prediction. The inputs include the residue type of the pair being predicted, the residue frequencies in a MSA for the corresponding columns, the frequencies of residue pairs in the columns of the MSA, the correlated mutation score for the column pair, secondary structure classification and solvent accessibility. To make a contact map from the residue contact probabilities given by the neural network two approaches are taken. One method is to use a fixed threshold that maximize precision and recall on a test set, the other is a variable, band dependant, threshold determined by estimating the number of contacts in a band from the sum of all the predicted contact probabilities in that band.

http://www.igb.uci.edu/servers/psss.html

RR301 rost_PROFCon [18]: The rost_PROFCon server takes a neural network approach to contact prediction. For each residue pair, information in two windows of length 9 centered around each residue is encoded. For each residue position in the windows there are 29 inputs to the network including frequency counts of the residues types in the corresponding MSA column, predicted secondary structure and the reliability of that prediction, predicted solvent accessibility and conservation weight. Inputs are also included to give a biophysical classification of the central residues, as well as whether or not the residues are in a low complexity region. Unusually for a contact predictor, inputs describing a window of length 5 half way between the pair of residues being considered are also included. In this window the same 29 input encoding scheme for each position is used as for the windows of length 9. A binary encoding scheme is used to describe the separation of the residues of interest. Finally, there are inputs describing global information such as the length (via a coarse grained binary encoding), the composition of amino acids and secondary structure for the protein.

http://www.predictprotein.org/submit_profcon.html

RR327 Hamilton-Huber-Torda [19]: The Hamilton-Huber-Torda server (recently named PoCM “possum” for Patterns of Correlated Mutations) is also a neural network predictor. The approach is to train the network on patterns of correlation. For a given residue pair, there are two windows of length 5 centered on the residues. The correlated mutation score for all 25 pairs of residues between the windows are then calculated, the idea being that if the central residues are in contact, then adjacent residues are also likely to be in contact and so correlated. Inputs are also included for predicted secondary structure, biophysical classification of residues, a residue affinity score based on observed contacts in proteins of known structure, sequence length and residue separation.

http://foo.maths.uq.edu.au/~nick/Protein/contact.html

RR166 SAMT04-hand [26]: Is a whole structure 3D modeler based on homology and templates.

RR012 GPCPred [21]: Perhaps the most unusual approach to contact prediction in CASP6 is via “striped sheets”. For a given protein, a PSI-BLAST sequence profile is constructed, that is a 21 by L matrix that records the frequencies of amino acids in each of the positions of a MSA, where L is the length of the protein. From this matrix, windows of length w are extracted, w=1,5,9,25. During training, to reduce the number of dimension in the data in the windows, a self organizing map (SOM) [46] was created for each w, with output 3 integers in the range 0 to 5. Any profile window, or indeed central residue, could then be mapped to three integers by the trained SOMs. Genetic programming techniques were used to classify whether a pair of residues were in contact from the SOM outputs for the windows around them.

http://www.sbc.su.se/~maccallr/contactmaps

RR088 Bystroff [27]: The Bystroff predictor uses a threading method to predict contact maps. The target sequence is aligned to a set of template sequences with template contact maps, and target contact maps generated. Predicted contact maps are then scored using a “contact free energy” function, and physicality rules applied such as those outlined in Contact Filtering section.

http://www.bioinfo.rpi.edu/~bystrc/downloads.html

In the CASP6 experiment it can be seen that the contact predictors that performed best were those that took a whole structure 3D modeling approach, though several "pure" contact predictors also performed well. It is interesting to note that it is becoming more common for 3D modeler builders to rely on pure contact predictors to refine and select amongst models. No doubt as the pure predictors improve and the newer ones are incorporated into the 3D predictors, this will lead to both better 3D structure and contact prediction.

For the pure contact predictors there are a number of general trends in accuracy that have been observed in the literature. Since most contact prediction methods rely on multiple sequence alignments, they tend to have a lower accuracy on proteins for which there a few homologues. Many predictors also report a decrease in accuracy for longer sequence proteins [1, 5, 21, 35], though there are exceptions [18, 19]. In some cases the average predictive accuracy may be reduced by up to a factor of 2 for long proteins. This may be due to the fact that for shorter proteins a randomly chosen residue pair is more likely to be in contact than for a longer one. Similarly, residues that are close on a sequence are more likely to be in contact and so are usually easier to predict than distant residues. Also, most predictors will significantly improve their accuracy if allowed to make fewer predictions. For instance, on a test set of 1033 proteins the PoCM predictor gave average accuracies of 0.174, 0.217, 0.27, 0.307 on the best L, L/2, L/5 and L/10 predictions, respectively [19]. This can be useful if only a few higher quality predictions are required.

Predictive accuracies also tend to vary widely between secondary structure classes such as those of the SCOP classification [47]. Proteins classified as "all α" are almost always poorly predicted in comparison to other classes. For example, the rostPROFCon server obtained an average accuracy of 0.24 on all α proteins, but 0.35 and 0.36 on the other classes, on a test set of 522 proteins with minimum sequence separation of 6 residues and the best L/2 predictions taken [18]. This order of decrease in accuracy is typical of contact predictors and may be due to a number of factors. It may be that to maintain the structure of the α-helices the kinds of substitutions possible are restricted, and so there is less information within the areas of the multiple sequence alignments corresponding to helices. Another problem may be the "windows" of residues approach that some of the predictors take. Since, on average, a single turn of a regular α-helix is 3.6 residues long, if two sequence distant residues contained in alpha helices are in contact, the residues adjacent to these residues are unlikely to be in contact. Hence one approach that might improve prediction on residues contained in alpha helices would be to use non-standard windows for these residues. For instance, for a window of size 5 around a given residue, the window would be taken as the 4th and 7th residues before and after the central residue. In this way it would be ensured that the residues in the window were on the same side of the helix.

All of these factors in combination can lead to a wide variation in predictive accuracy. On a data set of 1033 proteins, the PoCM predictor had an average accuracy of 0.174 on the best L predictions, with sequence separation of at least 5 [19]. Taking the subset of 64 proteins of class α+β for which there were at least 100 sequences in the alignment, the average accuracy rises to 0.457 for the best L/10 predictions.

5 Conclusions

Fariselli et al. [20] state that their goal is to obtain an accuracy of 50%, for then the folding of a protein of less than 300 residues length could be reconstructed with good accuracy (within 0.4-nm RMSD). While current contact predictors are still well short of this aim, predictive accuracy has significantly improved in recent years and has provided a valuable source of additional information in protein structure prediction.

Interestingly, contact predictions are not yet widely used in combination with 3D structure prediction and only a few approaches use them routinely. However, 3D modeling approaches which do use evolutionary analyses to predict contacts in protein structures seem to also be the better performing ones. One reason why contact prediction is generally omitted in fold recognition is simply algorithmic difficulties. Dynamic programming, as it is used in sequence(s)-sequence(s) alignment and sequence(s)-structure(s) alignment approaches, is not easily to reconcile with these residue pair distance constraints. The problem does not exist with 3D prediction methods that use heuristic optimization methods instead. Well performing programs of this kind include Skolnick's TASSER protein folding approach [48] and Baker's fragment assembly approach Rosetta [40]. Even when it is not possible to integrate contact predictions into a structure predictor it may still be useful to use the predictions as a way of selecting the "best" structure from a number of generated models. Our own experiments have shown that if a set of 3D models is ranked according to how many predicted contacts it is in agreement with, then the (known) real structure is ranked most highly against other predicted structures in almost all cases.

As we have seen, a number of different methodologies for protein contact prediction have been developed in recent years. The question is then how can contact prediction be improved? One approach would be to attempt to construct a predictor that combines the best and most novel aspects of each. Most predictors have a similar core of inputs to a training algorithm such as predicted secondary structure, but each has some unique feature such as using the predicted solvent accessibility, a stringent contact filtering algorithm, or a totally novel encoding as in the stripped sheets approach of McCallum. Also, within 3D structure prediction, meta-servers that make predictions based on the predictions of other servers have proved highly successful in the CASP experiments, often out performing all other methods. As more contact predictors come online it will be interesting to see if meta-contact predictors will enjoy similar success.

Acknowledgements

The authors gratefully acknowledge financial support from the University of Queensland, the ARC Australian Centre for Bioinformatics and the Institute for Molecular Bioscience. The first author would also like to acknowledge the support of Prof. Kevin Burrage's Australian Federation Fellowship.

References

1. Gobel U, et al., *Correlated mutations and residue contacts in proteins.* Proteins, 1994. **18**: p. 309-317.

2. McLachlan AD, *Tests for comparing related amino acid sequences.* J Mol Biol, 1971. **61**: p. 409-424.

3. Neher E, *How frequent are correlated changes in families of protein sequences?* Proc Natl Acad Sci U S A, 1994. **91**(1): p. 980-102.

4. Vicatos S, Reddy BVB, and Kaznessis Y, *Prediction of distant residue contacts with the use of evolutionary information.* Proteins: Structure, Function, and Bioinformatics, 2005. **58**: p. 935-949.

5. Singer MS, Vriend G, and Bywater RP, *Prediction of protein residue contacts with a PDB-derived likelihood matrix.* Protein Eng, 2002. **15 (9)**: p. 721-725.

6. Lin K, et al., *Testing homology with CAO: A contact-based Markov model of protein evolution.* Comp. Biol. Chem., 2003. **27**: p. 93-102.

7. Clarke ND, *Covariation of residues in the homeodomain sequence family.* Protein Sci., 1995. **7**(11): p. 2269-78.

8. Korber BTM, et al., *Covariation of Mutations in the V3 Loop of Human Immunodeficiency Virus Type 1 Envelope Protein: An Information Theoretic Analysis.* Proceedings of the National Academy of Sciences, 1993. **90**: p. 7176-7180.

9. Martin LC, et al., *Using information theory to search for co-evolving residues in proteins.* Bioinformatics, 2005. **21**(22): p. 4116-4124.

10. Oliveira L, Paiva ACM, and Vriend G, *Correlated Mutation Analyses on Very Large Sequence Families.* ChemBioChem, 2002. **3**(10): p. 1010-1017.

11. Akmaev VR, Kelley ST, and Stormo GD, *Phylogenetically enhanced statistical tools for RNA structure prediction.* Bioinformatics, 2000. **16**(6): p. 501-512.

12. Tillier ERM and L. TWH, *Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments.* Bioinformatics, 2003. **19**(6): p. 750-755.

13. Wollenberg KR and A. WR, *Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap.* Proc. Natl. Acad. Sci. U S A, 2000. **97**: p. 3288-3291.

14. McGuffin LJ, Bryson K, and Jones DT, *The PSIPRED protein structure prediction server.* Bioinformatics, 2000. **16**: p. 404-405.

15. Shapire RE, *The boosting approach to machine learning: An overview*. MSRI Workshop on Nonlinear Estimation and Classification. 2002: Springer.

16. Haykin S, *Neural Networks*. 2nd ed. 1999: Prentice Hall.

17. Zell A et al, *Stuttgart Neural Network Simulator User Manual Version 4.2*. 1998: University of Stuttgart.

18. Punta M and Rost B, *PROFcon: novel prediction of long range contacts.* Bioinformatics, 2005. **21**(13): p. 2960-2968.

19. Hamilton N, et al., *Protein contact prediction using patterns of correlation.* Proteins: Structure, Function, and Bioinformatics, 2004. **56**: p. 679-684.

20. Fariselli P, et al., *Prediction of contact maps with neural networks and correlated mutations.* Protein Eng, 2001. **14**: p. 835-843.

21. MacCallum RM, *Stripped sheets and protein contact prediction.* Bioinformatics, 2994. **20**(1): p. i224-i231.

22. Cortes C and Vapnik V, *Support vector network.* Machine and learning, 1995. **20**: p. 273-297.

23. Boser B, Guyon I, and V. V. *A training algorithm for optimal margin classifiers*. in *Proceedings of the fifth annual workshop on computational learning theory*. 1992.

24. Chang C-C and Lin C-J, *LIBSVM : a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm*. 2001.

25. Koski T, *Hidden Markov Models for Bioinformatics*. 2002: Springer.

26. Karplus K, et al., *Combining local-structure, fold-recognition, and new-fold methods for protein structure prediction.* Proteins: Structure, Function, and Genetics, 2003. **53**(S6): p. 491-496.

27. Shao Y and Bystroff C, *Predicting Interresidue contacts using templates and pathways.* Proteins, 2003. **53**: p. 497-502.

28. Conrad C, et al., *Automatic Identification of Subcellular Phenotypes on Human Cell Arrays.* Genome Research, 2004. **14**: p. 1130-1136.

29. Tsai C-H, et al., *Improving disulphide connectivity prediction with sequential distance between oxidized cysteines.* Bioinformatics (Advanced Access), 2005.

30. Hu J, et al., eds. *Mining protein contact maps*. In 2nd BIOKDD Workshop on Data Mining in Bioinformatics. 2002.

31. Yuan Z, *Better prediction of protein contact number using a support vector regression analysis if amino acid sequence.* BMC Bioinformatics, 2005. **6**: p. 248-257.

32. Aloy P, et al., *Predictions without templates: new folds, secondary structure, and contacts in CASP5.* Proteins Suppl., 2003. **6**: p. 436-456.

33. Olmea O and Valencia A, *Improving contact predictions by the combination of correlated mutations and other sources of sequence information.* Fold Design, 1997. **2**: p. S25-S32.

34. Mirny L and Domany E, *Protein Fold Recognition and Dynamics in The Space of Contact Maps.* Proteins, 1996. **26**: p. 319-410.

35. Fariselli P, et al., *Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations.* Proteins, 2001. **Suppl 5**: p. 157-162.

36. Fariselli P and Casadio R, *Neural network based prediction of residue contacts in protein.* Protein Eng, 1999. **12**: p. 15-21.

37. Graña 0 and et al *CASP6 assessment of contact prediction.* Proteins: Structure, Function, and Bioinformatics, 2005.

38. Koh IYY et al. , *EVA: evaluation of protein structure prediction servers.* Nucleic Acids Research, 2003. **31**: p. 3311-3315.

39. Pazos F, Helmer-Citterich M, and Ausiello G, *Correlated mutations contain information about protein-protein interaction.* J Mol Biol, 1997. **271**: p. 511-523.

40. Kim DE, Chivian D, and Baker D, *Protein structure prediction and analysis using the Robetta server.* Nucleic Acids Research, 2004. **32**: p. W526-W531.

41. Rychlewski L and Fischer D, *LiveBench-8: The large-scale, continuous assessment of automated protein structure prediction.* Protein Science, 2005. **14**: p. 240-245.

42. Torda AE, Procter JB, and H. T, *Wurst: A protein threading server with a structural scoring function, sequence profiles and optimised substitution matrices.* Nucleic Acids Research, 2004. **32**(W532-W535).

43. Baldi P and Pollastri G, *The Principled Design of Large-Scale Recursive Neural Network Architectures-DAG-RNNs and the Protein Structure Prediction Problem.* Journal of Machine Learning Research, 2003. **4**: p. 575-603.

44. Baldi P and Pollastri G, *Machine Learning Structural and Functional Proteomics.* IEEE Intelligent Systems (Intelligent Systems in Biology II), 2002(March/April).

45. Pollastri G and Baldi P, *Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners.* Bioinformatics, 2002. **18**(Suppl. 1): p. S62-S70.

46. Kohonen T and Makisari K, *The self-organizing feature maps.* Phys. Scripta, 1989. **39**: p. 168-172.

47. Andreeva A, et al., *SCOP database in 2004: refinements integrate structure and sequence family data.* Nucleic Acids Research, 2004. **32**(Database issue): p. D226-9.

48. Zhang Y, Arakaki AK, and Skolnick J, *TASSER: An automated method for the prediction of protein tertiary structures.* Proteins: Structure, Function, and Bioinformatics, 2005. **Suppl. 7**: p. 91-98.