Bnd jnl

Request # 18388167 **OCT 25, 2005**

Ariel To: 64.40.17.85/140.163.217.217

Memorial Sloan-Kettering Cancer Center
Medical Library Nathan Cummings Center (METRO #146)
1275 York Avenue
New York, NY 10021

#542

DOCLINE: Journal Copy Epayment

Title:	Protein engineering.
Title Abbrev:	Protein Eng
Citation:	1996 Nov;9(11):941-8
Article:	The prediction of protein contacts from multiple s
Author:	Thomas DJ; Casari G; Sander C
NLM Unique ID:	8801484 Verify: PubMed
PubMed UI:	8961347
ISSN:	0269-2139 (Print) 1460-213X (Electronic)
Publisher:	Oxford University Press, Oxford
Copyright:	Copyright Compliance Law
Authorization:	A Artale
Need By:	OCT 26, 2005
Maximum Cost:	**$25.00**
Patron Name:	Gangi-Dino, Rita
Referral Reason:	Not owned (title)
Library Groups:	BQSIMB,EFTS,RESOURCE,METRO
Phone:	1.212.639-7441
Fax:	1.646.422-2316
Email:	ill@mskcc.org
Comments:	**Please, we prefer as PDF or Ariel. Thank you! EFTS, METRO, BQSI.**
Routing Reason:	Routed to NYUHHQ in Serial Routing - cell 1
Received:	Oct 25, 2005 (11:51 AM EST)
Lender:	Cold Spring Harbor Laboratory/ Cold Spring Harbor/ NY USA (NYUHHQ)

Bill to: NYUMSK
Memorial Sloan-Kettering Cancer Center
Medical Library Nathan Cummings Center
1275 York Avenue
New York, NY 10021

The prediction of protein contacts from multiple sequence alignments

David J.Thomas, Georg Casari and Chris Sander

European Molecular Biology Laboratory, Postfach 10.2209, 69012 Heidelberg, Germany

We have studied the question of how much extra predictive power the correlated mutational behaviour of pairs of amino acid residues separated along a sequence has concerning the likelihood of those residues being in contact in the folded protein. The mutational behaviour is deduced from multiple sequence alignments. Our findings are that there is, indeed, some valuable information available from this source and that it is sufficient to make a significant improvement in our ability to predict contacts, when compared with earlier methods that do not take into account the correlations between the mutations. This improvement is approximately twice as large as can be obtained by the more economical method of simply averaging pair preferences over the same sequence alignment. Even when using a method based on pair preferences, a further significant improvement can be made by penalizing more variable regions (on the reasonable assumption that invariant residues are relatively more likely to be in contact), though we have found no way of improving the pair preference method to the extent that it matches the method based on correlated behaviour. Our new method is thought to be the best data-based method of contact prediction developed so far, achieving, on average, an improvement over a random (i.e. information-free) prediction of a factor of five when the number of contacts predicted is chosen to match the number that actually occur.

Key words: correlated mutations/multiple sequence alignments/protein contact prediction

Introduction

A successful prediction of contacts between amino acid residues in a protein would be a significant step forward in pragmatic attempts to tackle the protein folding problem. Indeed, the combination of correctly predicted secondary structure and even a few contacts would be sufficient to deduce the major features of a protein's fold. There is thus some interest in methods of contact prediction. Although no current method to predict contacts directly can claim to be successful, it has been found that correlated mutations have some predictive power (Göbel *et al.*, 1994; Shindyalov *et al.*, 1994; Taylor and Hatrick, 1994) and that different pairs of residue types show differing propensities to be close in a folded protein (Miyazawa and Jernigan, 1985; Hendlich *et al.*, 1990; Sippl, 1990). The study described in this paper was an attempt to find out whether these two different types of information can be extracted simultaneously from a multiple sequence alignment and combined to provide more reliable predictions. The method is entirely probabilistic, being based on statistics derived from a database of known protein structures and sequences aligned to them, with the intention of trying to see whether specific mutational patterns convey extra information that cannot be gleaned by other means. In addition, predictions are biased in favour of residues located relatively near to each other in the sequence, which provides a substantial additional increase in performance.

The problem (known as 'threading') of trying to see if a putative fold is likely to be consistent with a given sequence is, of course, much simpler, and claims of adequate solution have been made (Hendlich *et al.*, 1990; Jones *et al.*, 1992; Lüthy *et al.*, 1992; Ouzounis *et al.*, 1993). Our experience with direct contact prediction suggests strongly that our estimated contact probability maps will improve these algorithms, too.

Materials and methods

The raw data for our method are multiple sequence alignments. We used a pre-computed set of HSSP files (Sander and Schneider, 1991) selected to minimize redundancy using the 'PDB_select' algorithm (Hobohm *et al.*, 1992). These contain sequences from the SWISS-PROT database (Bairoch and Boeckmann, 1994), aligned using the MAXHOM program (Sander and Schneider, 1991) against lead sequences whose 3-D structures are available from the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977). Structures derived by X-ray or electron crystallography and by NMR spectroscopy were admitted, but structures based on modelling studies and those whose provenance could not be determined were excluded. Sequences of <30 residues were also excluded. It is, of course, an assumption in our method that the sequences in an HSSP file do, indeed, share the same folding pattern and, hence, related interresidue contact maps, but this assumption is generally thought to be reliable.

Interresidue contact maps were calculated using two different algorithms: first, a routine kindly supplied by L.Holm (unpublished work) that uses an all-atom representation and, second, a much simpler routine using just the Cβ–Cβ distances.

The essence of the method is to predict the likelihood of a pair of residues being in contact, based only on their types and any evidence of concerted mutational behaviour. To this end, ordered tetruplets of amino acids are defined according to the following scheme:

$$\begin{aligned} a_1 &= \text{amino acid at position } i \text{ in sequence } k \\ a_2 &= \text{amino acid at position } j \text{ in sequence } k \\ a_3 &= \text{amino acid at position } i \text{ in sequence } l \\ a_4 &= \text{amino acid at position } j \text{ in sequence } l. \end{aligned} \quad (1)$$

This is illustrated in Figure 1.

Working through a rather large 'learning' database of sequences aligned to known structures (i.e. HSSP files), tetruplets in which a_1 and a_2 (both in the lead sequence of the alignment) are in contact are counted and the results stored in variables, $C(a_1,a_2,a_3,a_4)$, referred to as the contact cumulants. Similarly, the same tetruplets whose lead sequence pair of amino acids are judged not to be in contact are also counted

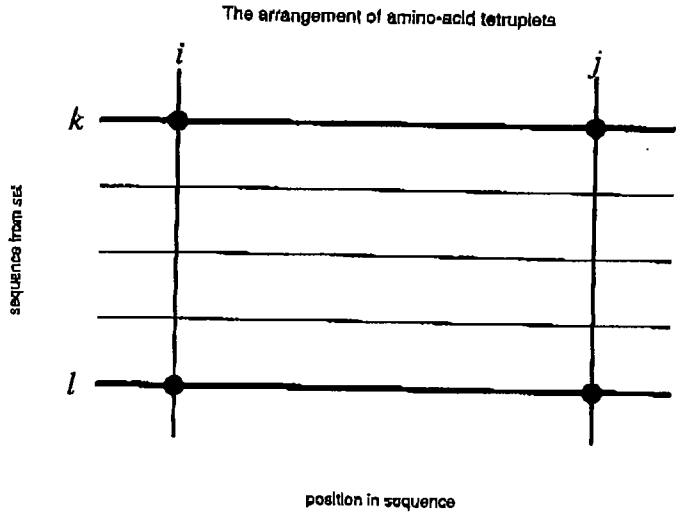

Fig. 1. Diagram of the sequence arrangement of the amino acid tetruplets being studied. The two bolder horizontal lines represent the amino acid sequences in the conventional linear fashion. The two lighter vertical lines denote equivalent positions within those sequences.

and the results stored in non-contact cumulants, $F(a_1,a_2,a_3,a_4)$. Within each HSSP file, the sums are obtained by looping over all unique residue position pairs separated by more than s residues (i.e. $\forall i, j : i < j - s$) and over all unique sequence pairs (i.e. $\forall k, l : k < l$). Apart from cutting out residue pairs closer than s places in the sequence, our learning method ignores positions entirely, so that the cumulants are 4-fold degenerate, with the following symmetry arrangement:

$$\begin{aligned} & C(a_1,a_2,a_3,a_4) && \text{original} \\ = & C(a_2,a_1,a_4,a_3) && \text{swapped positions} \\ = & C(a_3,a_4,a_1,a_2) && \text{swapped sequences} \\ = & C(a_4,a_3,a_2,a_1) && \text{both swapped.} \end{aligned} \quad (2)$$

The learning program takes this symmetry into account explicitly, simply by incrementing all four cumulants simultaneously as a group. The non-contact cumulants, F, naturally follow the same symmetry rules. There are $20^4 = 160\,000$ of each type of cumulant, although only 40 300 unique values because of the symmetry. This number is sufficiently large that the limited size of the available database does not justify any further extension, say by encoding the local environment or the sequence separation explicitly. Our present learning database is sufficiently large that all of our cumulants have at least a few entries and most have many, so we have not had major problems with poor statistics.

For any given pair of residue positions in a multiple alignment, there is a possibility of any tetruplet of amino acids occurring repeatedly. This immediately raises the question of bias, because some tetruplets could then be over- or underweighted. For this reason, a 'seen-once' mechanism was implemented and made available in both the cumulating (i.e. learning) and in the predicting parts of the program. The seen-once mechanism ensures that for any given contact position in a given map, the cumulant for any given tetruplet of amino acids (and the symmetry-related partners) is incremented only once, even if that tetruplet can be extracted many times from the corresponding stripes in the multiple sequence alignment. It also has a welcome side-effect of speeding up the computation, because it is quicker to establish that a tetruplet has been seen before than it is to update the four symmetry-related cumulants. Tests showed that the seen-once rule did, indeed, improve predictive performance significantly.

The size of a multiple sequence alignment depends both on the length of the lead sequence and on the number of sequences aligned to it; hence it varies widely. If uncorrected, this would mean that files with long lead sequences or with a large number of alignments would contribute disproportionately to the results. To avoid this, the contributions to the cumulants are divided by a number, N_{data}, which is approximately equal to the square root of the number of tetruplets that can be extracted from the alignment table. This choice of N_{data} gives each learning protein an approximately equal weight in the stored cumulants. Tests have shown that the results are rather insensitive to the precise value of N_{data}, so it is actually calculated only approximately from the area of the alignment table with first order corrections when the seen-once rule is not used or from the average length of the alignment when it is. The actual formula used in the former case is:

$$\begin{aligned} N_{\text{data}} = {} & \text{total number of residues in multiple alignment table} \\ & - \text{length of lead sequence} - (s+1) \times \\ & (\text{number of alignments} - 1). \end{aligned} \quad (3)$$

A regression test showed that this formula approximates the square root of the number of tetruplets that can be extracted from the alignment table to high accuracy (the regression coefficients were: slope = 0.484; intercept = 20.8 and r = 0.996). Replacing N_{data} with a constant (meaning that every observed tetruplet contributes equally, so large files dominate) degrades performance slightly but significantly, whilst squaring it (which has no theoretical justification and means that large files are disproportionately downweighted) degrades performance markedly.

There is another systematic bias that must be removed: the density of contacts in a contact map depends on the length of the sequence. This is corrected by multiplying or dividing by the square root of the ratio of the number of actual contacts to the number of possible contact positions. The number of possible contacts is calculated as $N_P = i(i-1)/2$, where i = length of sequence $- s$. Thus, the actual weights applied to the contact and non-contact cumulants are:

$$w_C = \frac{1}{N_{\text{data}}} \div \sqrt{\frac{N_C}{N_P}} \quad (4)$$

and

$$w_F = \frac{1}{N_{\text{data}}} \times \sqrt{\frac{N_C}{N_P}} \quad (5)$$

The reason for applying the square root to each set of cumulants rather than simply multiplying or dividing one of them by the relative ratio N_C/N_P is to minimize the perturbation of the overall weight based on N_{data}. This weighting, being different for contacts and non-contacts and for different proteins in the learning set, means that the predictions can be made only in terms of relative trends rather than absolute probabilities. As far as we can tell, this loss of absolute scaling and rigour appears to be unimportant.

We would have preferred to formulate our approach in a strictly Bayesian form, in which (potentially) a series of conditional probabilities ending in either a prior or an un-

conditional probability would be multiplied to give the best observational estimate of the probability of a given contact actually occurring. In circumstances such as ours, where rather large quantities of information may be involved, it is conventional to re-express the probabilities as their logarithms (often called log-odds) and to add them instead. This has certain numerical and computational advantages, though no effect on the results, but it does mean that we discuss our probabilities in terms of their logarithms. In fact, our approach is Bayesian only in spirit, because we multiply estimates of unconditional probabilities rather than the conditional ones strictly required.

In the absence of weighting terms, we would express the logarithm of the probability of a contact as:

$$\log \left(\frac{C(a_1, a_2, a_3, a_4)}{C(a_1, a_2, a_3, a_4) + F(a_1, a_2, a_3, a_4)} \right) \quad (6)$$

being the logarithm of the ratio of the number of observed contacts for a given tetruplet of amino acids to the total number of occurrences of that tetruplet. However, the use of variable weighting makes the C and F terms formally incompatible, so instead we use:

$$\rho(a_1, a_2, a_3, a_4) = \log \left(\frac{C(a_1, a_2, a_3, a_4) + \beta}{C(a_1, a_2, a_3, a_4) + \beta} \right) \quad (7)$$

The constant β was introduced to downweight the effects of sparse data, but it turns out that the overall results are exceedingly insensitive to its value over many orders of magnitude, for reasons that are discussed later. This latter formula does not represent a probability in the direct way that equation (6) does, but it shows the same trend, which seems to be sufficient in practice. The performances of the two formulae were, of course, compared (with $w_C = w_F = 1$) with our standard test of the success rate of predicting contacts and, remarkably, the results differed insignificantly.

As is usual in this type of work, the cumulants were generated from a 'learning set' of proteins and the method evaluated with a 'test set' whose members were chosen to be sufficiently independent of those of the learning set. Some preliminary testing was performed with sets prepared from a relatively short file of non-homologous proteins kindly supplied by Burkhard Rost, but all of the results reported here are based on subsets of the 'pdb_select.aug_1994' file prepared by Hobohm *et al.* (1992). Specifically, the learning set was the list of protein chains with <25% sequence homology at the head of this file. The test set was the incremental set of protein chains with <30% homology to each other and to the proteins in the learning set. It was not considered appropriate to use the 'jackknife' approach, where a larger learning set is used repetitively, excluding a single test protein on each run, because of the excessive computational cost. (A single learning run can take a whole day. Even on the fastest available scalar processor and even if the largest alignments are removed, leaving 343 proteins, several hours are still needed. With 51 proteins in the test set, this would mean ~2 months for a single jackknife test, which is hardly practical.)

For test purposes, predictions were made by the following simple method. For each pair of residue positions in the protein to be predicted, the aligned sequences in the appropriate HSSP file were read and the corresponding term ρ (cf. Equation 7) averaged over all sequence pairs (normally using the seen-once rule), to give $\bar{\rho}$ at each position in the contact map. The symmetry of the map was, of course, taken into account to reduce computational cost. The weighting scheme used implies that the average of ρ over the whole contact map should be near to zero, and this was found to be true. Then the contact map was sorted in order of decreasing $\bar{\rho}$ and the top N_C entries were taken as predictions of contacts. (N_C is the number of contacts in the contact map, so we predict exactly as many contacts as are actually judged to occur. As is explained later, our claimed results do not depend strongly on this choice.) The success rate was then scored as the number of correct predictions divided by N_C. It should be noted that this testing method is exceptionally stringent and could never yield a success rate of 100%, as will be explained in greater detail later.

In order to evaluate the results, they were compared with predictions made via simpler routes, i.e. from the lead sequence alone and from the multiple alignment, both using (single-sequence) pair preferences calculated by contraction of the tetruplets on the second sequence:

$$C(a_1,a_2) = \sum_{a_3} \sum_{a_4} C(a_1, a_2, a_3, a_4) \quad (8)$$

The non-contact pair cumulants $F(a_1,a_2)$ are calculated in the same way. The predictions using the pair cumulants $C(a_1,a_2)$ and $F(a_1,a_2)$ were made using the same methods as those for the tetruplet cumulants $C(a_1,a_2,a_3,a_4)$ and $F(a_1,a_2,a_3,a_4)$.

The method as described so far ignores completely the separation of residues along the sequence. For the tetruplets this is almost inevitable, because we do not have sufficient data to determine reliably the implied number of extra parameters and our pair preferences are normally derived from the tetruplets by contraction. It is possible to compensate for these lacking data, however, on the simplifying assumption that the probability of a contact represented by a given tetruplet at a certain separation, $P(a_1,a_2,a_3,a_4,t)$, can be estimated as the product of the overall probability for the tetruplet, $P(a_1,a_2,a_3,a_4)$, multiplied by the probability of a contact at a given sequence separation $P(t)$. In practice, the equation used is:

$$\log P = \overline{\rho(a_1,a_2,a_3,a_4)} + \gamma \log P(t) \quad (9)$$

The dramatic effect that this equation has on the results is discussed in the next section, as is the curious behaviour as γ is varied.

We thought that we would have sufficient data to derive pair preferences at various distances, say $\log P(a_1,a_2,t)$, and these were calculated and tested against our normal approximation $\log \rho(a_1,a_2) + \gamma \log P(t)$. Remarkably, the approximation performed significantly better. We suspect that the poorer performance of the formally better formula was because of poor statistics, but attempts to damp down the less well determined terms with a constant analogous to β degraded the performance even more, for reasons that we have not investigated fully.

Results

Our results are displayed for clarity in graphical form and each graph is designed to address a specific question about the performance of our algorithm. Perhaps the most important question concerns the way in which the overall performance

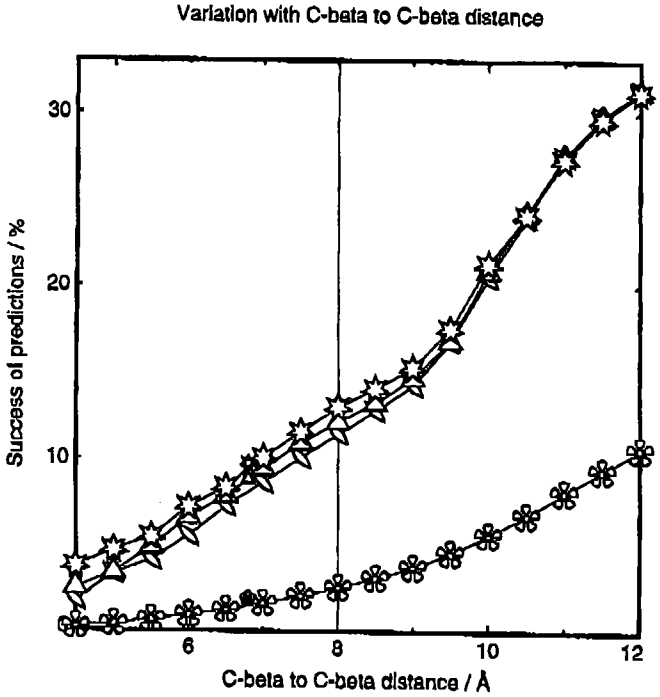

Fig. 2. The effect of varying the Cβ–Cβ cut-off distance (in half-Angstrom steps). The four lines are (A) the expected success rate of a completely random prediction (marked with flowers); (B) the success rate using pair preferences on only the lead sequence of the alignment (marked with lenticles); (C) the success rate using pair preferences averaged over the alignment (marked with triangles); and (D) the success rate using the new tetruplet method (marked with stars). In all cases, the number of contacts predicted is the same as the number that actually occur, and the success is counted as the fraction of predicted contacts that are actually correct. The vertical line marks our 'standard run', which has an 8 Å cut-off and achieves a success rate of 13%, five times more than the random prediction. In all cases, the same contact definition was used both while learning and while predicting. It will be noticed that for distances up to ~9 Å, predictions using the new tetruplet method consistently perform ~2% better than pair preferences on the lead sequence alone, whilst pair preferences on the whole alignment show an intermediate level of performance. In this region, the performance is almost linearly dependent on the cut-off distance, with a slope of ~2.9%/Å and zero-crossing at ~3.8 Å. It is striking that the difference in performance of the three methods disappears and the curves have a clearly different shape above 9 Å. This is believed to reflect the fact that such large distances no longer represent contacts in a physically meaningful way. The smaller symbols show the equivalent results when L.Holm's (unpublished work) all-atom contact definition is used. It will be seen that the results correspond approximately to those which would be obtained when using a Cβ–Cβ cut-off distance of ~6.8 Å, a value almost exactly centred on the linear region of the graph. This gives further support to our belief that the tetruplets and pair preferences encode genuine physical information.

depends on the definition of contact used. Figure 2 shows that if a Cβ–Cβ cut-off is used, the success rate rises monotonically with the allowed Cβ–Cβ distance. The same figure includes the results obtained when L.Holm's (unpublished work) all-atom contact definition is used. It will be seen that this definition gives results corresponding approximately to a Cβ–Cβ cut-off of 6.8 Å. In all cases, the same contact definition was used both while learning and while predicting. The ratio over random can be increased enormously by making the Cβ–Cβ cut-off distance smaller, but this apparent improvement is illusory because the absolute success rate falls. It is particularly noteworthy that the curve tips up sharply at 9 Å and the difference between the three predictive methods very nearly vanishes. This must be a reflection of the fact that with a

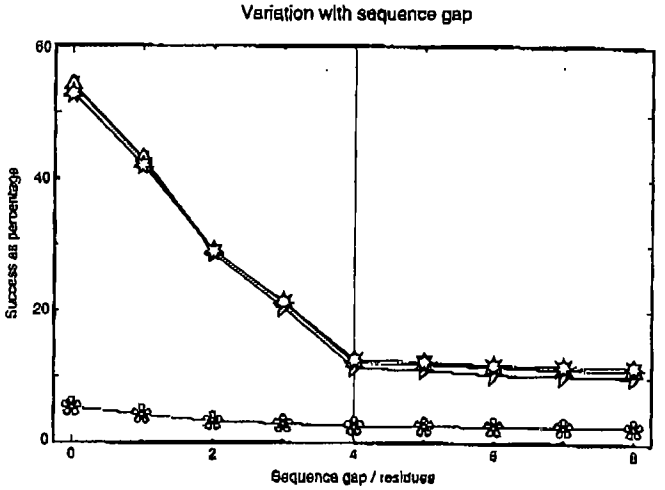

Fig. 3. Effect of varying s. This graph was calculated under our 'standard conditions', including in particular a Cβ–Cβ cut-off of 8 Å and shows a dramatic and completely artefactual apparent improvement in performance if the minimum sequence gap between residues, s, is reduced below 4. The leftmost points in the graph, with $s = 0$, correspond to immediate neighbours rather than 'self-contacts', since s measures the gap between residues and not the difference in their indices. The position of the sharp kink in the curves is, of course, a direct function of the cut-off distance, and the illusory increase in performance is due simply to the fact that any pair predicted to be in contact that has such a small sequence separation is bound (within a certain tolerance) to satisfy the criterion for being in contact. Such an artefact is, incidentally, much less pronounced when using L.Holm's (unpublished work) definition of contacts, which actually requires side-chain atoms to touch.

contact definition so generous, the two residues concerned could in fact be widely separated and any mutational behaviour would obviously be devoid of information regarding propensities to be in contact. Below ~9 Å it can be seen that the tetruplet prediction performs consistently ~2% better than pair preferences on the lead sequence alone and, except for the smallest cut-offs, pair preferences on the whole alignment achieve a value close to the average of the other two methods.

Figure 3 shows the variation in performance as the minimum sequence separation is altered. The graph is annotated in terms of the gap between residues, s, which is always one less than the conventionally measured sequence separation. Again, the same setting was used both while learning and while predicting. It will be seen that the performance appears to rise as shorter separations are allowed, but tends to level off for $s > 4$, which is the region of greater interest when predicting the fold of a protein. The apparent rise in performance for the smaller values of s is partly illusory and arises from the fact that contacts are much more likely between close sequence neighbours, though, in fact, the success rate is higher than the unconditional probability of a contact at these sequence separations (except, of course, for sequence neighbours), so there must still be useful compositional information. Our 'standard run', marked with the vertical line, uses $s = 4$.

For test purposes we usually predict the same number of contacts as are judged to occur in the lead sequence protein. It might at first be thought that this is an unfair inclusion of knowledge. It is also interesting to know whether the most strongly predicted contacts are in fact more likely. Figure 4 addresses these questions. The graph shows the expected form, with a gently falling performance if too many contacts are predicted and a fairly marked improvement if fewer are. This

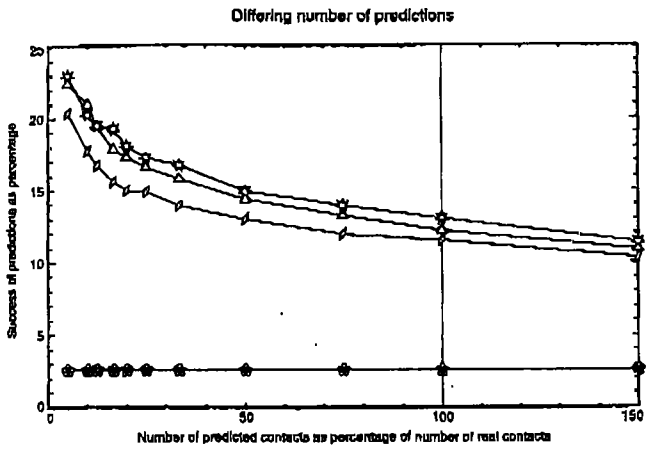

Fig. 4. The effect of predicting fewer contacts. This graph shows the more strongly predicted contacts are more likely. The very strongly predicted contacts have a surprisingly high probability of being correct, approaching 25%. The irregularity of the left-hand side is a sign of poorer statistics as the total number of contacts being predicted falls. It will be noticed that the slope at the normal test point (100%) is very low, proving that the exact number of contacts predicted does not have a significant influence on the claimed results. Clearly, if the overall success rate were very high, the slope at this point would increase dramatically.

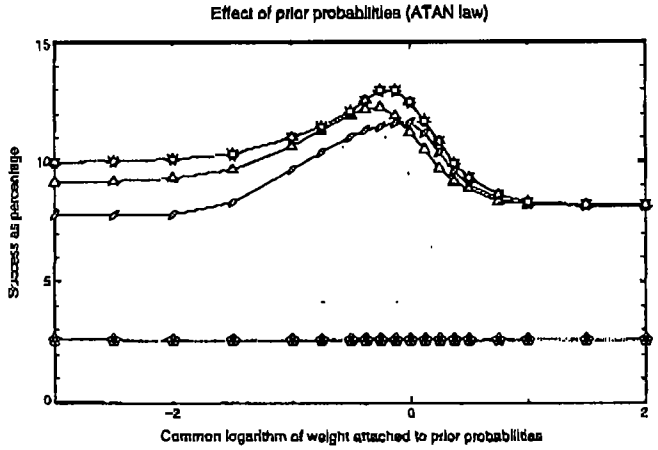

Fig. 5. The effect of varying γ. Intuitively, assuming the veracity of Bayesian statistics, we would expect the performance of our algorithm to peak when $\gamma = 1$, but this graph shows that the true peaks are shifted towards lower values of γ for reasons that we do not understand. Notwithstanding this, it shows the enormous increase in performance that can be achieved by combining two different types of information: (i) the mutational behaviour of pairs of residues and (ii) the reduced propensity of contacts to be formed between residues distant in the sequence.

indicates that the contacts predicted with higher probability are, indeed, more likely to be correct.

Perhaps the most curious and interesting results are displayed in Figure 5, which shows the effect of varying the relative weighting between the terms dependent upon the amino acid composition and that depending on the sequence separation (using γ, cf. Equation 9) and the enormous improvement in performance which can be achieved by adjusting γ suitably. Intuitively, if our probabilities were estimated correctly and the chance of a contact could truly be expressed as a product of one probability based on mutational behaviour multiplied by one based on sequence separation alone, we would expect the rate of success to peak at $\gamma = 1$ (corresponding to 0 in the graph because of its logarithmic scale), which is when Equation 9 most closely, if inaccurately, resembles a Bayesian form. In

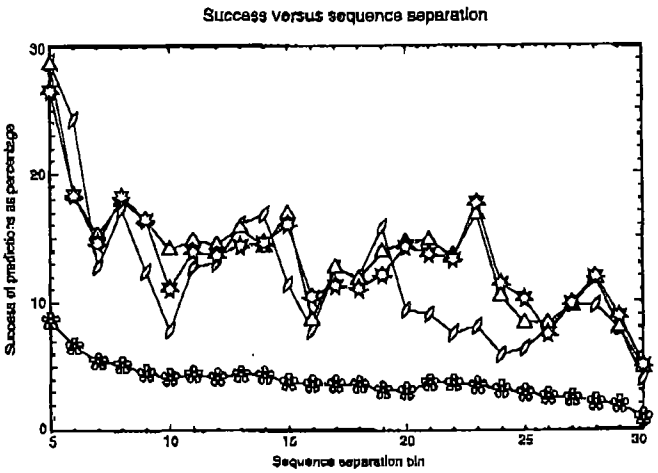

Fig. 6. Success of predictions versus sequence separation. The horizontal axis represents sequence separation in a non-linear fashion and the exact (error-function) law is given in Table I. This graph is also computed under our standard conditions, with $s = 4$ and a Cβ–Cβ cut-off of 8 Å. The statistics in the left-hand half of this graph are poor, for lack of data, but the opposite is true at the right-hand extreme. It can be seen that in sequence bins 20–25 (corresponding to sequence separations 22–31) both the new tetruplet algorithm and the use of pair preferences over a sequence alignment produce very much better results than do pair preferences on the lead sequence alone and it is undoubtedly in this region that the algorithm achieves its maximum usefulness.

fact, the peaks for the various options occur at slightly offset positions, corresponding to lower values of γ. We do not understand the reason for this. This same figure also demonstrates another striking effect, which is that at the right-hand edge, where the probabilities are totally dominated by the sequence separation part, the results are actually better than those achieved by pair preferences on the lead sequence alone, as seen at the left-hand edge.

Figure 6 shows the success rate as a function of sequence separation using just the compositional information (i.e. with $\gamma = 0$). The horizontal axis is non-linear, in such a way that the marked number at the left-hand side corresponds exactly to the actual sequence separation, but then separations are progressively packed ever more densely into successive bins. The rightmost bin includes all separations up to infinity. The non-linear law is based on the error function, which was chosen for its smoothness, resulting in reasonably sized bins. It will be seen that the success rate is not a smooth function of separation. However, it should be remembered that the statistics in this graph are poor, except for the rightmost bins where the converse is true. The general trend of high success at very small separations and very low success at large ones is to be expected because the probability of a contact between residues drops as their sequence separation increases, as can be seen from Figure 7 (which uses an arctangent function for the horizontal axis resulting in a better bin-size distribution than in Figure 6, particularly for very long sequence separations). In contrast with the graph of success rate, the statistics on this graph are excellent and the small hump at sequence bin 19 (separations 23–25) is a real effect.

Certain other tests of performance were made. One was to include the test proteins in the learning set, which resulted in a markedly better success rate for the predictions using tetruplets. This result is not as trivial as it may appear and indicates that the various tetruplets do, indeed, carry very specific information that cannot be encoded within a small

Table I. Sequence bins for graphs with error-function axes

Bin		Separations	Bin		Separations	Bin		Separations
4	←	4	13	←	13	22	←	25
5	←	5	14	←	14	23	←	26–27
6	←	6	15	←	15–16	24	←	28–29
7	←	7	16	←	17	25	←	30–31
8	←	8	17	←	18	26	←	32–34
9	←	9	18	←	19	27	←	35–37
10	←	10	19	←	20–21	28	←	38–42
11	←	11	20	←	22	29	←	43–51
12	←	12	21	←	23–24	30	←	52–∞

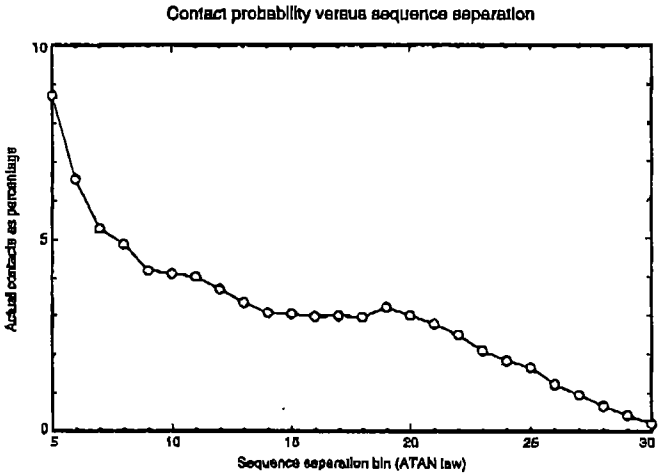

Fig. 7. Unconditional probability of a contact versus sequence separation. This is plotted on the improved (arctangent) non-linear axis given in Table II. The graph shows a result well known to polymer physicists, that the probability of contacts between residues in a high polymer falls with their separation. This graph is obtained from our normal learning database and has excellent statistics. The slight peak at sequence separation bin 19 (separations 23–25) is thus a real effect. This distribution is used to penalize predictions at very large sequence separations, which are relatively less likely to be correct and to promote predictions for closer neighbours.

number of pair preferences. It also indicates that our learning set is not yet large enough, though there is little evidence of poor statistics for most of the tetruplets. The pair preferences, obtained by contraction from the tetruplets, are thought to be so well determined that an increase in the size of the learning set would have little effect.

Another test was performed to try to boost the performance of the algorithm. This was based on the putative supposition that positions that are absolutely conserved are relatively more likely to be in contact, whereas those that are highly variable are not. A minor alteration was made in the prediction algorithm such that the number of pairs or tetruplets actually seen was counted and a small variable multiple of this number was subtracted from log P. It is very unlikely that this method is optimal, but, as can be seen from Figure 8, there seems little reason to pursue the approach. Clearly, this additional test cannot be performed on the lead sequence alone, nor does it have any effect on the expected results for a random prediction. It can be seen that the effect on the tetruplet prediction is always detrimental, but a small improvement can be gained when using pair preferences across the alignment. However, the improvement is small, and insufficient to match the better performance of the prediction using tetruplets. It would seem that the tetruplets already encode the variability information in a superior way.

The tetruplets were compared with estimates of their values made from the underlying pairs to see if any simple rules could be found. The strongest indicator of a contact is the tetruplet CCCC, representing a conserved pair of cysteines, and is consistent with the pair CC occurring twice. However, tetruplets of the type CCCx (where x is any amino acid other than C) generally give a rather strong signal against being in contact, which is in strong contrast to the average of the pairs CC and Cx, which is generally in favour of being in contact. Unfortunately, in a scatter plot of the tetruplets versus the sum of matching pair preferences, it is only the tetruplets of the CCCx type that are obvious as outliers from the main distribution. Unfortunately, it is not possible to display the scatter plot in a usefully annotated way within the confines of a conventionally printed journal. The slope of the scatter plot is approximately equal to one and the correlation coefficient, r, peaks at ~70%, both figures depending weakly on the value of β. Another result was found by deliberate inspection and is that tetruplets such as KDDK are more likely to represent a contact than are the (symmetry-related) pairs KD and DK. Physically, this represents two partners in a salt bridge swapping positions in an otherwise broadly conserved structure (cf. Clarke, 1995).

Discussion

Although it is clear that our new method of contact prediction performs better than more conventional methods, it is clearly not achieving a useful level of accuracy as a direct method of predicting contacts. However, several factors should be borne in mind. The first is that the data (tetruplets) are actually poorer than those used in the best secondary structure predictions (Rost and Sander, 1994), which only achieve accuracies of ~72% in a variable spanning at most two degrees of freedom. The number of degrees of freedom in a raw contact map is more typically in the range of tens to hundreds of thousands. On this ground alone, one would not expect contact map prediction by statistical means to be successful. Further, there is the problem of degeneracy. Suppose that an algorithm such as ours had available to it data of unexceptionable quality, but were faced with a protein containing eight absolutely conserved cysteines. The algorithm does not take into account context such as sequence neighbours, so cannot distinguish the cysteines apart from their sequence separation, which is useful, though weak, information. In this case, there are 105 possible arrangements of disulphide bonds [the number of arrangements if all n cysteines pair is $n!/(2^{n/2}(n/2)!) = (n-1)!! \equiv \Pi_{i=1}^{n/2}(2i-1)$]. If the prediction algorithm was altered to make a

Table II. Sequence bins for graphs with arctangent axes

Bin		Separations	Bin		Separations	Bin		Separations
4	←	4	13	←	13–14	22	←	32–35
5	←	5	14	←	15	23	←	36–41
6	←	6	15	←	16–17	24	←	42–48
7	←	7	16	←	18	25	←	49–58
8	←	8	17	←	19–20	26	←	59–74
9	←	9	18	←	21–22	27	←	75–101
10	←	10	19	←	23–25	28	←	102–162
11	←	11	20	←	26–28	29	←	163–426
12	←	12	21	←	29–31	30	←	427–∞

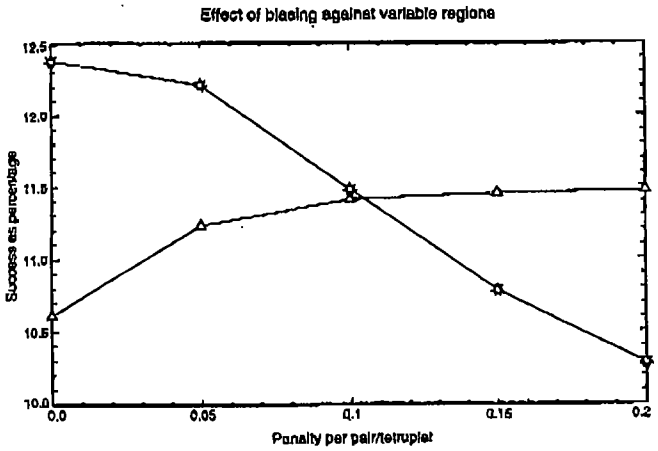

Fig. 8. Effect of penalizing variable positions. It will be seen that when using pair preferences on a multiple alignment, a small improvement in performance can be gained by penalizing variable regions. This is done simply by subtracting a small number (plotted on the horizontal axis) from the estimate of the logarithm of the probability of a contact for each unique amino acid tetruplet seen at a given position in the contact map. When using the new tetruplet method, however, the performance is degraded by such an alteration. The improvement available using the pair preferences is insufficient to allow their performance to reach that of the tetruplet method, but they do have the advantage of being computationally cheaper.

specific decision about which cysteines were in contact, our expectation of its success rate would be <1% in this case. In fact, our algorithm does not cross-check predicted contacts against each other, so might be expected to perform even less well. Similar though admittedly weaker arguments apply to all other tetruplets, depending on the relative abundances of the amino acids and the degeneracy of each tetruplet. Seen in this light it is amazing that our algorithm performs as well as it does! A discursive account of the extent to which putative protein models can be constrained on the basis of cysteine conservation patterns alone can be found in Kreisberg *et al.* (1995).

It was noted earlier that the results depend very little on the setting of β. Specifically, they vary insignificantly as β is raised from a negligible amount (say 10^{-15}) to a 'turnover' value in the order of 10^{-2} when there is a very small but hardly significant improvement in performance. Further increases result in a gradual deterioration of performance. The very slight peak in performance is thought to occur when the setting of β damps down terms with poor statistics sufficiently, but does not yet degrade the terms with good statistics: this is, of course, exactly the purpose for which β was introduced. It is, however, surprising how small this peak is. If the compositional terms $\rho(a_1,a_2,a_3,a_4)$ are sorted, the exact ordering is unstable with respect to changes in β, although there are naturally no major positional shifts for the terms determined with good statistics. When a prediction is made, in most cases a relatively large number of these terms will be averaged and the variation in these sums will be relatively smaller. In addition, we always choose a certain number of winners by sorting the results and, thus, it is only relative shifts and not absolute ones that are important. For these reasons, it is not particularly surprising that the method is so insensitive to the value of β.

Our algorithm estimates a map of contact probabilities with standard deviations. However, our testing method is a very stringent one, because it is based on taking only the most strongly predicted set of contacts, regardless of whether they may be self-consistent or not. It is easy to see that even if the probability map was in some sense perfect, the testing method would still not produce a success rate of 100%. We believe that the contact probability map may offer an improved performance when used to test putative folds (i.e. in what is commonly known as a 'threading' approach) for two reasons: (i) the tetruplets do contain a little more valuable information and (ii) taking into account the standard deviations is usually helpful. Another way of expressing this is simply to say that using the contact probabilities cooperatively rather than singly would be expected to be more successful.

We intend to pursue this 'cooperative' approach, but not by means of threading. One very obvious point is that contact probabilities away from the diagonal would have a very strong predictive value if the larger values clustered in diagonal bands; these would be predictive of paired β-strands, and the relative orientation would be discernible, according to which diagonal is displayed. It would be interesting, too, to bias this approach in favour of regions that are strongly predicted to have the β conformation by one of the better secondary structure prediction programs. It is not yet clear to what extent this might be making use of the same information twice, though it is clear that any point off the diagonal of a contact map refers to two different points in the sequence, whereas any point in a secondary structure prediction refers to just one. Clearly, we could also search for the characteristic pattern of contacts near to the leading diagonal that would be indicative of α-helical structure. In this case, we would have greater reservations about using secondary structure predictions because the contact prediction algorithm would not have any greater information available to it than the secondary structure prediction one.

It was always assumed in our work that the statistics of the cumulants would be normal, as assured by the central limit theorem in the presence of abundant uncorrelated data. Clarke

(1995) has, however, recently published a thoughtful account of a contrasting study based on a much smaller data set in which this assumption failed for reasons attributed to evolutionary and sampling bias. We have no reason to suppose that our own work suffers from such problems.

There is an additional comment which should be made concerning the computational effort involved in executing our algorithm. The number of tetruplets that can be extracted from a multiple alignment scales as the square of the sequence length multiplied by the square of the number of aligned sequences. Applying the seen-once rule saves some work, but nevertheless very long sequences with very wide alignments take a totally disproportionate amount of computational time and do not contribute proportionately to the overall results. However, even if the worst cases are removed from the learning list, the algorithm is still a fairly expensive one: learning takes several hours on the fastest available scalar processor and testing takes several tens of minutes.

Conclusion

Our new contact prediction algorithm, based on a statistical analysis of mutational patterns of pairs of residues and the general distribution of contacts at differing sequence separations, has achieved a small but significant improvement in performance compared with the use of pair preferences alone. Unfortunately, the overall level of performance still leaves much to be desired. The use of tetruplets of amino acids from a sequence alignment gives ~1% extra performance over pair preferences averaged over the same alignment, but at considerable extra computational cost. Pair preferences averaged over the alignment also achieve ~1% extra over pair preferences on the lead sequence alone, but at a much more modest cost. It would seem that the use of pair preferences, either on the lead sequence or on any available alignment, can generally be recommended, but tetruplets should only be used in special cases where the small but expensive increase in performance can be justified.

Our results do indicate, however, that correlated mutational behaviour may indeed have some predictive value, although it is weaker than we might have hoped.

Acknowledgements

The authors are grateful to Liisa Holm for supplying routines to read PDB files and to calculate contact maps and to Reinhard Schneider for a routine to read HSSP files.

References

Bairoch,A. and Boeckmann,B. (1994) *Nucleic Acids Res.*, **22**, 3578–3580.
Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
Clarke,N.D. (1995) *Protein Sci.*, **4**, 2269–2278.
Göbel,U., Sander,C., Schneider,R. and Valencia,A. (1994) *Proteins: Struct. Funct. Genet.*, **18**, 309–317.
Hendlich,M., Lackner,P., Weitckus,S., Floeckner,H., Froschauer,R., Gottsbacher,K., Casari,G. and Sippl,M.J. (1990) *J. Mol. Biol.*, **216**, 167–180.
Hobohm,U., Scharf,M., Schneider,P. and Sander,C. (1992) *Protein Sci.*, **1**, 409–417.
Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) *Nature*, **358**, 86–89.
Kreisberg,R., Buchner,V. and Arad,D. (1995) *Protein Sci.*, **4**, 2405–2410.
Lüthy,R., Bowie,J.U. and Eisenberg,D. (1992) *Nature*, **356**, 83–85.
Miyazawa,S. and Jernigan,R.L. (1985) *Macromolecules*, **18**, 534–552.
Ouzounis,C., Sander,C., Scharf,M. and Schneider,R. (1993) *J. Mol. Biol.*, **232**, 805–825.
Rost,B. and Sander,C. (1994) *Proteins: Struct. Funct. Genet.*, **19**, 55–72.
Sander,C. and Schneider,R. (1991) *Proteins: Struct. Funct. Genet.*, **9**, 56–68.
Shindyalov,I., Kolchanov,N. and Sander,C. (1994) *Protein Engng*, **7**, 349–358.
Sippl,M.J. (1990) *J. Mol. Biol.*, **213**, 859–883.
Taylor,W.R. and Hatrick,K. (1994) *Protein Engng*, **7**, 341–348.

Received December 20, 1995; revised June 24, 1996; accepted June 26, 1996