Protein Engineering vol.14 no.11 pp.835–843, 2001

Prediction of contact maps with neural networks and correlated mutations

Piero Fariselli[1], Osvaldo Olmea[2], Alfonso Valencia[2] and Rita Casadio[1,3]

[1]CIRB and Department of Biology, University of Bologna, via Irnerio 42, Bologna, Italy and [2]Protein Design Group, CNB-CSIC, Cantoblanco, Madrid 28049, Spain

[3]To whom correspondence should be addressed.
E-mail: casadio@alma.unibo.it

Contact maps of proteins are predicted with neural network-based methods, using as input codings of increasing complexity including evolutionary information, sequence conservation, correlated mutations and predicted secondary structures. Neural networks are trained on a data set comprising the contact maps of 173 non-homologous proteins as computed from their well resolved three-dimensional structures. Proteins are selected from the Protein Data Bank database provided that they align with at least 15 similar sequences in the corresponding families. The predictors are trained to learn the association rules between the covalent structure of each protein and its contact map with a standard back propagation algorithm and tested on the same protein set with a cross-validation procedure. Our results indicate that the method can assign protein contacts with an average accuracy of 0.21 and with an improvement over a random predictor of a factor >6, which is higher than that previously obtained with methods only based either on neural networks or on correlated mutations. Furthermore, filtering the network outputs with a procedure based on the residue coordination numbers, the accuracy of predictions increases up to 0.25 for all the proteins, with an 8-fold deviation from a random predictor. These scores are the highest reported so far for predicting protein contact maps.

Keywords: contact maps/correlated mutations/neural networks/protein structure predictions/residue contacts

Introduction

The three-dimensional (3D) architecture of a protein represents the ultimate molecular information and from that springs a wealth of significant scientific results comprising the understanding of the protein-folding mechanism. Anfinsen's thermodynamic hypothesis (Anfinsen, 1973), recently revisited by using lattice models (Govindarajan and Goldstein, 1998), has placed the bases of the today commonly accepted idea that under physiological conditions the covalent structure of a protein is sufficient to funnel the folding of the molecule into its native structure. In order to exploit the content of this 'molecular biology Holy Grail' (Eisenhaber *et al*., 1995), different methods focusing on distances or contacts among residues in the protein have been proposed for predicting the structure starting from the sequence, when a significantly similar homologue with known structure is lacking. In particular, contacts among residues constrain protein folding and characterize different protein structures. Therefore, the prediction of residue contacts in proteins is an interesting problem whose solution may be useful in protein-folding recognition and *de novo* design. Provided that residue contacts are known for a protein sequence, the major features of its 3D structure could be deduced by applying different reconstruction methods (Bohr *et al*., 1993; Vendruscolo *et al*., 1997).

A contact map also constitutes a structural 'fingerprint' of a given protein. The secondary structure, fold topology and side-chain packing patterns (for side-chain contact maps) can be highlighted easily from a contact map. Indeed, each protein can be identified based on its contact map. Furthermore, structural similarity between a pair of proteins can be detected by inspecting their contact maps, without searching for all their possible orientations (Godzik *et al*., 1992). Recently residue contact classification has been correlated to structural patterns (Selbig and Argos, 1998). It follows that a good predictor of protein contacts would play a fundamental role in structural theoretical approaches.

The prediction of contacts between residues requires the study of the protein inter-residue distances related to the specific type of residue pair. The distribution of distances between residue pairs in proteins of known 3D structure has been first computed to derive mean force potentials with the aim of tackling the folding and the inverse folding problem (Sippl, 1990; Maiorov and Crippen, 1992; Huang *et al*., 1995; Mirny and Shaknovich, 1996; Miyazawa and Jernigam, 1999; Zhang and Sung-Hou, 2000). Other methods more specifically address the problem of residue contact prediction using the information derived from the occurrence of correlated mutations in similar proteins (Göbel *et al*., 1994; Shidyalov *et al*., 1994; Taylor and Hatrick, 1994; Thomas *et al*., 1996). Moreover, algorithms have been developed which combine correlated mutations with other properties, such as sequence conservation (as computed from multiple sequence alignment), sequence separation along the chain, alignment stability and residue-specific contact occupancy (as evaluated from the 3D protein structure) (Olmea and Valencia, 1997). Finally, neural network-based methods have been applied to predict whether distances between couples of residues are above or below a given variable threshold (Lund *et al*., 1997; Gorodkin *et al*., 1999) and also contact maps of proteins (Fariselli and Casadio, 1998, 1999).

It has been shown that predictions of contact maps with neural networks score higher than statistical approaches (Fariselli and Casadio, 1999). Even though deviation from a random predictor was equal to six; the efficiency was not sufficient for protein prediction and the development of new predictors was foreseen. Therefore, we thought to take advantage of the high degree of flexibility of the neural network system and in this paper we combine the information derived from different sources, including sequence conservation, correlated mutations and predicted secondary structures, to address the problem of predicting contact maps of proteins.

Table I. The database of proteins used to train and test the predictors

L<100	1c5a	1sco	2sn3	1bkf	1npk	3lzt	1juk	1axn
1a1i_A	1cfh	1spy	2sxl	1bkr_A	1pdn_C	3nul	1kid	1b0m
1a1t_A	1ctj	1sro	3gat_A	1br0	1pkp	5p21	1mml	1bg2
1a68	1cyo	1tbn	3mef_A	1bsn	1poa	7rsa	1mrj	1bgp
1a7i	1fna	1tiv	4mt2	1bv1	1put	*L 170–299*	1nls	1bxo
1acp	1hev	1tle	5pti	1bxa	1ra9	1ad2	1ppn	1dlc
1ah9	1hrz_A	1tsg	*L 100–169*	1c25	1rcf	1akz	1rgs	1irk
1aho	1kbs	1ubi	1a62	1cew_I	1rie	1amm	1rhs	1iso
1aie	1mbh	1uxd	1a6g	1cfe	1skz	1aol	1thv	1kvu
1ail	1mbj	2acy	1acz	1cyx	1tam	1ap8	1vin	1moq
1ajj	1msi	2adx	1asx	1dun	1vsd	1bf8	1xnb	1svb
1aoo	1mzm	2bop_A	1aud_A	1eca	1whi	1bjk	1yub	1uro_A
1ap0	1nxb	2ech	1ax3	1erv	2fsp	1byq_A	1zin	1ysc
1ark	1ocp	2fdn	1b10	1exg	2gdm	1c3d	2baa	2cae
1awd	1opd	2fn2	1bc4	1hfc	2ilk	1cdi	2fha	2dpg
1awj	1pce	2fow	1bd8	1ifc	2lfb	1cne	*L>300*	2pgd
1awo	1plc	2hfh	1bea	1jvr	2pil	1cnv	16pk	3grs
1bbo	1pou	2hoa	1bfe_A	1kpf	2tgi	1csn	1a8e	1arv
1bc8_C	1ppt	2hqi	1bfg	1kte	2ucz	1ezm	1ads	
	1brf	1rof	2lef_A	1bgf	1mak	3chy	1fts	

Protein length (L) is the residue number of the covalent structure.

The question still to be answered is then to what extent a neural network system is capable of learning the correlation between the residue covalent structure of a protein and its contact map, as it is computed from its known 3D structure. Specifically, we investigate how all the possible information which can be derived from the sequence alignment of the protein, as previously described (Olmea and Valencia, 1997), affects the predictive performance. The neural networks described in this paper are trained to learn the correlation between protein sequences and the contact maps of a set of non-homologous proteins of known structure, including 173 chains of different length. Proteins are selected with the specific requirement that they have at least 15 sequences in the multiple sequence alignment. The test is performed using a cross-validation procedure. Our results indicate that the neural network-based predictor can reach an accuracy as high as 0.21, with a deviation from a random predictor of a factor >6. Moreover, the inclusion of a filtering procedure previously described and based on residue occupancies (Olmea and Valencia, 1997) improves the average accuracy up to 0.25; in this case the improvement over a random predictor reaches a factor of 8. These scores are higher than that previously obtained by the other methods developed to predict residue contacts (Thomas *et al*., 1996; Olmea and Valencia, 1997; Fariselli and Casadio, 1999) and indicate that the performance of the prediction of contact maps with neural networks can be improved by increasing the amount of information given as input.

Materials and methods

The database

We use a large set of non-homologous proteins of known 3D structure. In Table I we list the proteins sorted by length [proteins are labelled by their Protein Data Bank (PDB) code], since our results require this classification. The list includes all proteins in the PDB select-list of non-sequence-redundant protein structures with percentage of sequence identity lower than 25% (Hobohm and Sander, 1994) and whose chain was not interrupted (822 proteins). Furthermore, the chains are selected provided that alignments with more than 15 sequences were obtained: using this approach, in total, our set includes 173 proteins (Table I).

Contact map definition

The native structure of a protein is approximated by the set of the coordinates listed in its PDB file. If a protein contains N atoms, the corresponding representation requires $3N$ co-ordinates. An alternative view of the protein makes use of a distance matrix, a symmetric square $N \times N$ matrix whose elements are the distances among the atoms in the protein. This representation is obviously redundant: it requires $N(N-1)/2$ degrees of freedom instead of $3N$. However, it is still very important, since it has been demonstrated that the redundancy can help in the reconstruction of the 3D structure of the protein only when some elements of the distance matrix are available (Kuntz *et al*., 1989). This is often the case, especially in NMR spectroscopy. Based on this notion, a correct prediction of a relevant part of a distance matrix could be used to compute *ab initio* the protein 3D structure. Unfortunately, so far no method is available to compute the distance matrix starting from the protein sequence. However, we can try an approximate solution of this problem by predicting the contact map of a protein. A contact map is a binary version of the distance matrix and it is defined as a symmetric matrix consisting of $N \times N$ elements (for a protein of N residues) whose values are set to 1 or 0 according to whether there is or there is no contact between the two residues at issue.

A contact is said to exist between each pair of residues whenever the mutual distance is below a given arbitrary threshold. The distance involved in the different definitions of a contact can be that between the C_α–C_α atoms (Mirny and Domany, 1996), between the C_β–C_β (Thomas *et al*., 1996; Lund *et al*., 1997; Olmea and Valencia, 1997) and the minimal distance between atoms belonging to the side chain or to the backbone of the two residues (Fariselli and Casadio, 1999).

In this paper we adopt the definition of contact used also by Olmea and Valencia (Olmea and Valencia, 1997), i.e. we calculate contacts only using C_β atoms and adopting a threshold value of 8 Å. When it is not stated explicitly, the minimal sequence separation is $|i-j| \leqslant 7$ residues. This choice is done in order to overcome the effect of learning local contacts and

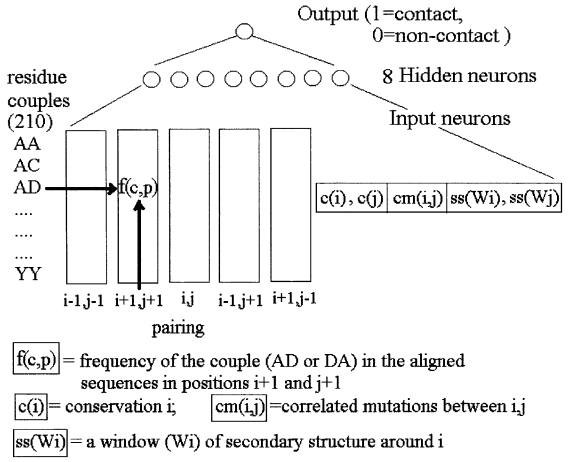

Fig. 1. Neural network architectures implemented in this paper. The different input codings are depicted. The simplest coding includes 210 residue couples (AA, AC, . . .) times five possible pairings (out of the 3-residue-long windows centred in the *i* and *j* paired residues, respectively). Each array (a rectangular box), coding for one of the possible pairing (5), contains the frequency of the couple (c) in the alignment positions (p) [indicated as f(c, p)]. Inputs of increasing complexity includes sequence conservation [c(*i*), c(*j*)], correlated mutations [cm (*i*, *j*)] and predicted secondary structures in a 3-residue-long window (W) centred on each residue of the *i*, *j* couple [ss (W*i*), ss (W*j*)] (see text for further details).

particularly to polarize predictions on the intra turn and intra helix contacts (i, i +4 ; i, i + 5; i, i + 6) and in this, the present approach is different from that previously described (Fariselli and Casadio, 1999).

Computing sequence conservation and correlated mutations

Sequence variability was taken from the HSSP database (Sander and Schneider, 1993). In the HSSP definition (Sander and Schneider, 1991), variability is 0 when positions in the multiple sequence alignments are completely conserved and it increases proportionally to the number of amino acid changes occurring at that position.

Correlated mutations are calculated as previously described (Göbel *et al*., 1994). Briefly, a distance array is used to codify each position in the alignment. This position-specific array contains all the residue–residue distances between all the possible pairs of sequences at that position. The correlation value between each pair of positions in the alignment is computed as the correlation of the two arrays for each possible residue pair. Corresponding elements in the arrays contain the distance between the same two sequences in the two positions under comparison. The scoring matrix of McLachlan (McLachlan, 1971) defines the distances between residues. Positions with a percentage of gaps >10% are set at a correlation value of –1 and completely conserved positions are set at a correlation value of 0. The similarity value of gaps is set to a dummy value of 0.

Neural network architecture and input codings

Neural networks have been proved to be one of the most successful methods for prediction of contact maps of proteins (Fariselli and Casadio, 1999). In this work we implement a neural network architecture which is similar to that described before. This topology, which was found to be the best performing one with the problem at hand, is depicted in Figure 1. A single output neuron codes for contact (output value close to 1) and non-contact (output value close to 0). The hidden layer consists of eight hidden neurons. A new type of input coding was previously introduced (Fariselli and Casadio, 1999) and it is also used here. Each residue pair in the protein sequence is coded as an input vector containing 210 elements ($20\times(20 + 1)/2$), representing all the possible ordered couples of residues (considering that each residue couple and its symmetric are coded in the same way). This is done in order to reduce the number of weight junctions. When a single sequence is used, the input neuron coding for the ordered couple of amino acidic residues at positions i and j is set to 1, while the remaining 209 are set to 0. In order to take into account the sequence neighbours we use a 3-residue-long input window, considering both parallel and anti-parallel pairing of the two segments centred at positions i and j, respectively. This leads to the coding of the couples formed by the residues in positions $\{i - 1, j - 1\}$, $\{i, j\}$, $\{i + 1, j + 1\}$ (parallel pairing) and $\{i - 1, j+1\}$, $\{i, j\}$, $\{i + 1, j - 1\}$ (anti-parallel pairing) ending up with five possible combinations ($\{i - 1, j - 1\}$, $\{i, j\}$, $\{i + 1, j + 1\}$, $\{i -1, j + 1\}$, $\{i + 1, j - 1\}$) of the ordered couples. This procedure requires 1050 (210×5) input neurons. When multiple sequence information is used (as is always the case in this paper), this binary input code is changed to a frequency-based one. This is done by considering the alignment from the corresponding HSSP files (Sander and Schneider, 1991) and taking all the possible couples generated by residues in positions i and j of the different aligned sequences. After normalization to the number of sequences, the frequencies of occurrence in the alignment of each couple are used in the corresponding position of the 210 element input vector representing all the possible ordered couples. Using this approach, the 210 element vector may have more than one component activated (Fariselli and Casadio, 1999).

A major characteristic of the networks described in this paper is that they use different input codings, characterized by increasing complexity. We implement neural networks taking into account each position in the sequence alignment information derived from sequence conservation and correlated mutations as previously computed (Olmea and Valencia, 1997). Moreover, the predicted secondary structures are also added as input to the networks.

To solve our specific problem five neural networks of different complexity are implemented (Figure 1):

(i) NET is the neural network with the simplest input, consisting only of the pairing couples taken from the multiple sequence alignments with a 3-residue-long window centred at the residue pair to be predicted. This network, which was also used in a previous paper (Fariselli and Casadio, 1999), is implemented here for the sake of comparison with our previously published results (see below).

(ii) NETC is a neural network which contains three input nodes more than NET. Two nodes code for the conservation of each residue in position i and j, respectively, and a third one codes for the correlated mutations of the i and j positions.

(iii) NETCW is similar to NETC, with six input neurons coding for the sequence conservation (instead of the two of NETC). This is done in order to also take into consideration the sequence conservation value of the nearest neighbouring residues. The number of input neurons coding for the correlated mutations is increased from one in NETC (referring to the i, j couple) to nine (each combination in a window of three residues, Figure 2).

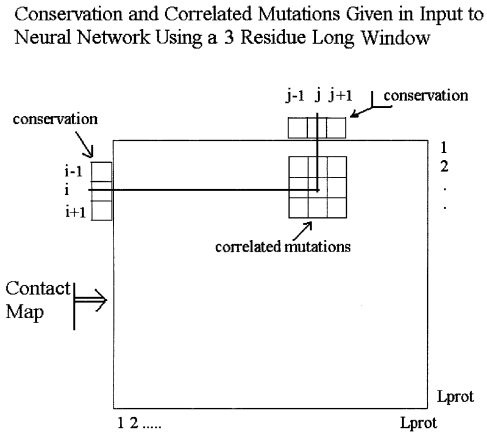

Fig. 2. A graphical view of the windows presented to the network (NETCW) including sequence conservation and correlated mutations.

(iv) NETCSEP is NETC with one more input neuron that codes for the length of the sequence separation between residues to be predicted. The length of the protein under inspection is used as a normalization factor.

(v) NETCSS adds to NETC 18 input neurons in order to include also the predicted secondary structure. For each residue couple (i, j) a 3-residue-long input window, that contains a node for each discriminated secondary structure type (α, β, coil), is considered. Predictions are obtained with a neural network-based predictor, implemented 'in house' and trained using a cross-validation procedure on 822 non-homologous proteins (25% sequence identity). Its overall three-state accuracy reaches 0.74 (Jacoboni *et al.*, 2000).

All the five networks are trained using a cross-validation procedure. This is performed by splitting the set comprising 173 proteins into three subsets including an approximate equal number of proteins. The networks are trained using the back-propagation algorithm and a balancing procedure (Fariselli *et al.*, 1993) in order to avoid deterioration of the performance due to the large imbalance between contacts (less abundant) and non-contacts (more abundant). The weight junctions are randomly initialized in the range [–0.01, 0.01]; the learning rate and the momentum term are set to 0.1 and 0.9, respectively.

In order to compare all the implemented networks together and with the previously described methods based on correlated mutations, the number of predicted contacts is set equal to half of the protein length $L_p/2$ (following the procedure described in Olmea and Valencia, 1997). A sorting procedure based on the network output values is adopted. Contacts are defined as the highest $L_p/2$ prediction values for a protein of length equal to L_p and are routinely characterized by output activation values $>$0.75–0.80.

The filtering procedure

To avoid contact overprediction, the predicted pairs are filtered taking into account the amount of contacts that each residue type can make (Olmea and Valencia, 1997). The filtering procedure is based on the occupancy data (or residue-coordination numbers) of each residue. This value is statistically derived from the set of protein structures of the database and takes into account the secondary structure type and the solvent exposition of each residue. Using this, the number of predicted contacts of a residue becomes a function of its structural environment. Therefore, the occupancy can be considered an estimate of the maximal number of contacts that each residue can make and is used to limit the number of contacts predicted for each residue.

Evaluation of the efficiency

To score the efficiency of the predictors we use three statistical indices. The first and the most used index to evaluate the predictor accuracy is defined as:

$$A = N_{cp}^*/N_{cp} \quad (1)$$

where N_{cp}^* and N_{cp} are the number of correctly assigned contacts and that of total predicted contacts, respectively. Routinely the accuracy is evaluated for each protein and then averaged over the protein set under consideration ($<A>$).

The improvement over a random predictor is evaluated by computing the ratio between A (Equation 1) and the accuracy of a random predictor (N_c / N_p):

$$R = A/(N_c/N_p) \quad (2)$$

where N_c is the number of real contacts in the protein of length L_p, and N_p are all the possible contacts. In this paper in order to limit the prediction of local contacts (clustered along the main diagonal of the contact map) we set to 7 the minimum length of the sequence separation between residues in contact. Since the contact map S is symmetric and residues whose sequence gap is <7 are not included, N_p is computed to be equal to $(L_p - 7)(L_p - 6)$. In a previous paper in which we used neural networks having as input only evolution information, the minimum sequence gap between residues in contact was set to 4. Therefore, the R values computed in this paper cannot be directly compared to those previously published since the accuracy of the random predictor is differently evaluated (Fariselli and Casadio, 1999). Also for the R value an averaging procedure is adopted ($<R>$).

A third and useful index measures the difference in the distribution of the inter-residue distances in the 3D structure for predicted pairs compared with all pair distances in the structure (Pazos *et al.*, 1997). This index is defined as:

$$X_d = \Sigma_{i=1,n} (P_{ic} - P_{ia})/nd_i \quad (3)$$

where n is the number of bins of the distance distribution (15 equally distributed bins from 4 to 60 Å cluster all the possible distances of residue pairs observed in the protein structure); d_i is the upper limit (normalized to 60 Å) for each bin, e.g. 8 Å for the 4–8 Å bin; P_{ic} and P_{ia} are the percentage of predicted contact pairs (with distance between d_i and d_{i-1}) and that of all possible pairs, respectively. By definition, values of $X_d = 0$ indicate no separation between the two distance populations, meaning that the predicted contacts are randomly distributed; values of $X_d > 0$ indicate positive cases, when the population of the distances between predicted contact pairs is shifted to smaller values with respect to the population of the distances of all residue pairs in the protein. Since contact distances have an upper limit of 8 Å, the larger and positive X_d, the more efficient prediction of contacts is. Similarly to the other two indexes, X_d is also averaged on the protein sets ($<X_d>$).

Results

Contact distribution

Irrespective of the contact definition, it has been shown that the number of contacts (N_c) in the proteins increases almost

linearly with the protein sequence length, while the number of 'non-contacts' (N_{nc}) increases with the square of the sequence length (Vendruscolo *et al.*, 1997). On average N_c/N_{nc} is approximately 1/60 (Fariselli and Casadio, 1999). In order to compensate for this disproportion a balancing procedure is used during the training phase of the networks and a sorting procedure based on the network output values is adopted (see Materials and methods, *Neural network architecture and input codings*).

However, a crucial point to take into consideration is the distribution of protein contacts with respect to the length of the sequence separation between pairs. The sequence separation between two residues in position i and j in the protein sequence is computed as $|j - i|$. It is obvious that a well performing predictor of contacts should not be polarized by contact predictions among residues that are nearest neighbours in the sequence and covalently bounded. It can be evaluated that the frequency distribution of the contacts decreases at increasing length of the sequence separation between residue pairs (Thomas *et al.*, 1996; Fariselli and Casadio, 1999).

The theoretical distribution of contacts in a protein can be computed by assuming that:

(i) For the protein i, the number of contacts $N_c(i)$ increases linearly with the protein length L_i. The phenomenological estimate $N_c(i) = a \times L_i$ is consistent with previous evaluations (Vendruscolo *et al.*, 1997; Fariselli and Casadio, 1999).
(ii) In a contact map, contacts are randomly distributed (*a priori*).

From these assumptions it follows that the *a priori* probability of finding a contact (c) in a map i is:

$$P(c, L_i) = aL_i/L_i^2 = a/L_i \quad (4)$$

The *a priori* number of contacts at a given length of sequence separation (s) for a protein of length L_i is:

$$N_c(c, s, L_i) = (a/L_i)(L_i - s) \quad (5)$$

From this it follows that the probability $P(c, s)$ decreases linearly at increasing length of sequence separation. However, it must also be considered that the distribution is computed on proteins of different size, and that not all of the proteins can equally contribute to all sequence separations since the constraint $s < L_i$ must hold. Taking this into account, the probability of contacts for a given length of sequence separation (s) becomes:

$$P(c, s) = N_c(s)/N_{tot} = \Sigma_{Li > s}[N(L_i)N_c(c, s, L_i)]/N_{tot} \quad (6)$$

where the sum is carried out over all protein length, $N(L_i)$ is the number of proteins of length L_i and N_{tot} is the sum over all the sequence separations.

The theoretical frequency distribution for a *a priori* random assignment of contacts in a protein (Equation 6) is computed using the $N(L_i)$ value from our database. In Figure 3 the theoretical frequency distribution of contacts is compared with the real one evaluated using our database. It is evident that the pattern of the *a priori* distribution is characterized by an exponential-like behaviour, indicating that most of the contacts (68%) in proteins are between residues with a sequence separation ranging from 7 to 100. The real distribution deviates from that of a *a priori* random assignment of contacts. It is remarkable that the real frequency distribution of contacts contains in the same region of sequence separation 81% of the population. This indicates that contacts in proteins are not randomly distributed and occur predominantly between residues with a sequence separation spanning from 7 to 100 residues.

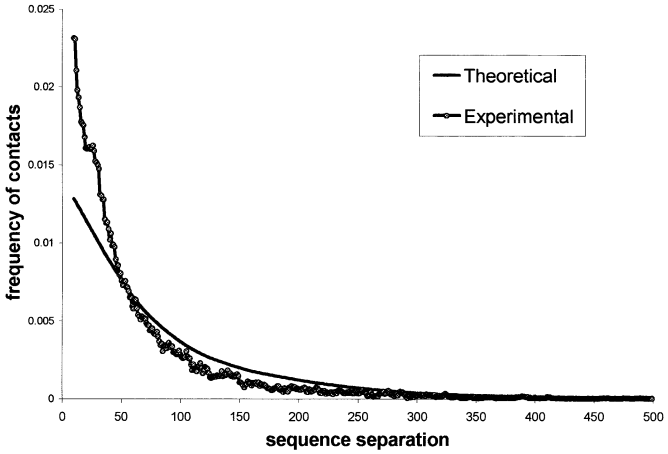

Fig. 3. Frequency distributions of the real and hypothetical contacts as a function of sequence separations. Hypothetical contacts are computed according to Equation 6. The minimum value of the length of sequence separation is 7 (residues).

The efficiency of the predictors

The implemented networks are tested on the selected database using a cross-validation procedure. The results are compared with those obtained using a previous method based only on correlated mutations (Olmea and Valencia, 1997) (Table II). In order to highlight the dependency of the results on the protein dimension, the accuracy values are listed after grouping the proteins also by sequence length. It is evident that on average for all the proteins neural networks (NET) give a very good performance with respect to the method based only on correlated mutation (Corr). For the sake of comparison with a neural network-based predictor previously described (Fariselli and Casadio, 1999) we notice that the average accuracy is the same as that published before (0.16). However, since the sequence separation value is set now to 7 (instead of 4 residues, as before), deviation from random is 1.16 points lower than before. This simply reflects the fact that with the current procedure we are narrowing the range of contact predictions accepted as correct.

When the information relative to correlated mutations is introduced as input to the network (NETC), the accuracy increases by 1% and the X_d value becomes 8.5 instead of 7.83 (Table II). This indicates that although the neural network captures most information, the pre-processing computation of the sequence conservation and correlated mutations helps the system in the classification task. This is not the case when the method based only on correlated mutations is adopted (Corr).

When information on sequence conservation and correlated mutations is extended to a 3-residue-long window (NETCW), the accuracy value increases of one percentage point; however, deviation from random (R) and X_d values are not modified with respect to NETC, indicating that the efficiency of the prediction is rather similar to that obtained with NETC. If the prediction efficiency of the subsets of proteins of different size is considered, it can be concluded that this result is mainly due to the decrease in prediction accuracy of proteins of large size (>300 residues).

Table II. Efficiency of the different methods used to predict contact maps

Method	All proteins (173)[a]			$L < 100$ (65)[a]			$100 \leqslant L < 170$ (57)[a]			$170 \leqslant L < 300$ (30)[a]			$L \geqslant 300$ (21)[a]		
	$<A>$	$<R>$	$<X_d>$	$<A>$	$<R>$	$<X_d>$	$<A>$	$<R>$	$<X_d>$	$<A>$	$<R>$	$<X_d>$	$<A>$	$<R>$	$<X_d>$
Corr	0.09	2.95	4.31	0.11	2.06	3.69	0.1	2.99	4.41	0.08	3.93	5.36	0.05	4.02	4.43
NET	0.16	4.84	7.83	0.23	4.21	8.9	0.16	4.98	8.14	0.09	4.94	6.16	0.07	6.24	6.02
NETC	0.17	5.08	8.48	0.24	4.44	9.4	0.17	5.20	8.88	0.1	5.36	7.04	0.07	6.30	6.57
NETCW	0.18	5.10	8.35	0.25	4.52	9.55	0.17	5.50	9.07	0.1	5.28	6.2	0.06	5.60	5.73
NETCsep	0.19	6.12	9.78	0.24	4.39	9.52	0.18	5.50	9.38	0.14	7.44	10.07	0.13	11.3	11.3
NETCSS	0.21	6.43	9.95	0.26	4.78	10.07	0.21	6.45	10.74	0.15	7.77	9.62	0.11	9.55	7.93

Corr, correlated mutations (Olmea and Valencia, 1997); NET, neural network trained with multiple sequence alignment; NETC, network with correlated mutations and sequence conservation as input only for residues *i* and *j*; NETCW, NetC with a 3-residue-long window centred on the couple to be predicted comprising correlated mutations and sequence conservation (see also Figure 2); NETCsep, NETC with sequence separation; NETCSS, NETC with a window of predicted secondary structures in input; $<A>$, number of correctly predicted over that of total predicted contacts; $<R>$, improvement with respect to a random predictor; $<X_d>$, distribution accuracy of the predicted contacts (see text for further explanation).
[a]Number of proteins in the subset; protein length (L) is defined as in Table I.

If sequence separation of the contact residue pairs is taken into account (NETCsep), an increase of the accuracy by two percentage points with respect to NETC is noticed. In this case all the other indices also increase. NETCsep reaches an average deviation from random (R) higher than 6 and an X_d value of 9.8. Noticeably, this predictor performs well also on proteins of large size (>300 residues).

The predictor which also implements the protein-predicted secondary structures (NETCSS) gives the highest score (Table II). As compared to NETC, which only considers correlated mutations, accuracy is increased 2-fold. With the increase of protein length the performance of NETCSS decreases (Table II). For large proteins the accuracy is lower than that obtained with NETsep. Nevertheless the overall performance of NETCSS is still the highest for proteins of small and medium size (<300 residues), indicating that this method is the best predictor of contact maps for proteins of average dimensions.

In Figure 4, the accuracy of the contact prediction is plotted as a function of the protein length using the different methods (Corr, FNET and NETCSS). It is evident that the accuracy of the prediction is dependent on the length of the protein. It is indeed easier to predict the contacts of short sequences, since the contact density is higher in small than in large proteins (Vendruscolo *et al*., 1997). Neural network-based methods (the NET predictors) are out-performing that based only on correlated mutations (Corr); however, the most significant differences in the efficiency of the prediction between the two approaches are to be found in the contact prediction of proteins of small and medium size.

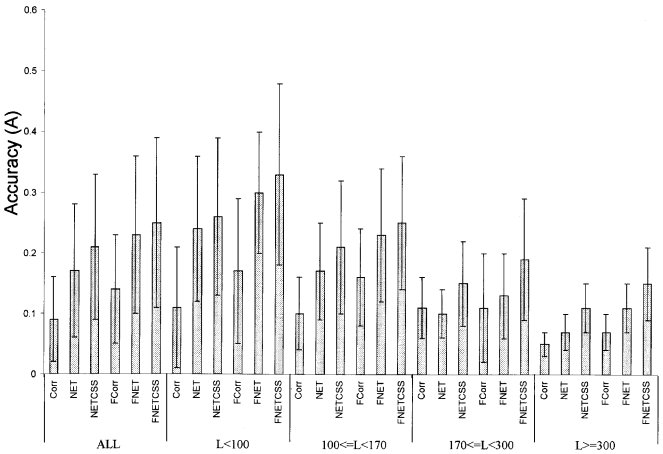

Fig. 4. Bar graph showing a comparison of the accuracy (A) of the contact predictions obtained with the different methods used in this paper. The effect of filtering (F) is also included. Accuracy of each method is averaged over the set of all proteins and also on the four different subsets including proteins of small, medium and large size, as indicated. For each protein set, six different sets of predictions are considered. Contact predictions are obtained with: Corr, correlated mutations; NET, the network trained with multiple sequence alignment; and NETCSS, the best performing network; the results of the predictors after filtering (FCorr, FNET, FNETCSS) are also shown. The height of the bars corresponds to the average accuracy of each method and the lines to the standard deviation.

Accuracy after the filtering procedure

When the filtering procedure is adopted, even though the number of predicted contacts is reduced, the accuracy level of the rest is of higher quality (Table III). All the networks increase their efficiency as measured by the different indices while keeping the same order of performance listed in Table II. After filtering, it also becomes evident that there is no improvement when the 3-residue-long window coding for conservation and correlated mutations is used; indeed NETC and NETCW show the same accuracy values. The former performs better on proteins of large size, while the latter performs better on small ones.

The most remarkable result shown in Table III is the accuracy reached by FNETCSS. The efficiency of FNETCSS on the whole set of proteins is demonstrated by the accuracy value equal to 0.25. Furthermore, the improvement over a random predictor is increasing up to 8 and the X_d value is close to 12. The other relevant observation is that this predictor still maintains its accuracy on proteins of medium and large size. This indicates that after filtering, contacts are predicted with a high efficiency, as concluded by considering the improvement over random is 9.47 and 12.71 times, respectively, on medium and large size proteins.

Network performance in different regions of sequence separation

All the contact predictions described above were performed considering a sequence separation of at least 7 residues. In this section we explore the effect of increasing this minimal sequence separation length on the prediction. Increasing the minimal sequence separations causes a severe reduction of the number of real contacts taken into consideration (Figure 3). Since the greatest amount of contacts in a protein is found in

Table III. Comparison of the performances of the different predictors after the filtering procedure

Filter method	All proteins (173)[a]			$L < 100$ (65)[a]			$100 \leqslant L < 170$ (57)[a]			$170 \leqslant L < 300$ (30)[a]			$L \geqslant 300$ (21)[a]		
	$<A>$	$<R>$	$<X_d>$	$<A>$	$<R>$	$<X_d>$	$<A>$	$<R>$	$<X_d>$	$<A>$	$<R>$	$<X_d>$	$<A>$	$<R>$	$<X_d>$
FCorr	0.14	4.76	7.01	0.18	3.57	7.53	0.16	4.62	7.39	0.09	4.90	6.44	0.07	6.45	6.10
FNET	0.21	6.41	10.02	0.30	6.02	11.93	0.21	6.06	10.38	0.12	6.19	7.67	0.10	8.78	7.06
FNETC	0.23	6.80	10.67	0.31	5.97	12.46	0.23	6.64	11.07	0.13	6.88	8.51	0.11	9.43	7.61
FNETCW	0.23	6.87	10.61	0.33	6.18	12.69	0.24	6.97	11.16	0.14	7.47	8.35	0.09	7.61	6.58
FNETCsep	0.24	7.70	11.86	0.31	5.87	12.19	0.23	6.75	11.45	0.17	8.77	11.57	0.16	13.93	12.45
FNETCSS	0.25	8.05	11.87	0.33	6.38	12.91	0.25	7.33	12.14	0.19	9.47	10.78	0.15	12.71	9.77

For notations see Table II.

Table IV. The efficiency of the best performing method (NETCSS) with and without filtering as a function of sequence separation

Minimal sequence separation	All proteins (173)[a]			$L < 100$ (65)[a]			$100 \leqslant L < 170$ (57)[a]			$170 \leqslant L < 300$ (30)[a]			$L \geqslant 300$ (21)[a]		
	$<A>$	$<R>$	$<X_d>$	$<A>$	$<R>$	$<X_d>$	$<A>$	$<R>$	$<X_d>$	$<A>$	$<R>$	$<X_d>$	$<A>$	$<R>$	$<X_d>$
7	0.21	6.43	9.95	0.26	4.78	10.07	0.21	6.45	10.74	0.15	7.77	9.62	0.11	9.55	7.93
7 (filter)	0.25	8.05	11.87	0.33	6.38	12.91	0.25	7.33	12.14	0.19	9.47	10.78	0.15	12.71	9.77
20	0.18	6.60	9.88	0.23	4.92	10.26	0.19	6.32	10.67	0.13	7.66	9.04	0.10	10.25	7.93
20 (filter)	0.21	8.28	11.41	0.29	5.84	12.51	0.23	7.27	12.15	0.16	9.34	10.02	0.13	9.27	9.66
40	0.15	7.62	9.21	0.20	4.18	7.97	0.18	7.12	10.75	0.11	8.25	8.47	0.08	10.44	7.35
40 (filter)	0.17	9.48	10.67	0.20	4.97	10.53	0.23	8.24	12.68	0.13	9.95	9.25	0.10	12.89	8.68
80	0.07	5.83	6.58	–	–	–	0.12	3.76	9.59	0.07	5.63	6.64	0.04	7.18	4.94
80 (filter)	0.09	9.07	7.68	–	–	–	0.50	4.92	15.31	0.10	7.75	7.87	0.07	10.84	7.09

(Filter), NETCSS outputs are filtered (Olmea and Valencia, 1997). Other notations are as in the legend of Table II.

the region of small sequence separation (see above), increasing this value is equivalent to reducing the expectation of randomly locating contacts during the prediction. In Table IV we analyse the results obtained using our best performing method (NETCSS), with and without the filtering procedure, tested on the protein sets using four different minimal sequence separations (7, 20, 40 and 80 residues). As expected, the accuracy index A diminishes as the minimal sequence separation increases. The X_d value decrease less than the accuracy value, indicating that the predicted contacts are anyway quite close in space (Olmea and Valencia, 1997). This is due to the fact that in the regions of large sequence separations contacts are not only less abundant but also more spot-like and less grouped together. This explanation is also confirmed by the behaviour of the index evaluating the improvement over random prediction (R).

In the set of experiments shown in Table IV, the values of R are nearly constant, even when the accuracy (A) is decreased. This indicates that when the minimal sequence separation increases, the improvement over a random predictor is almost constant due to the fact that the probability of randomly getting correct answers is also decreasing. Therefore, it can be concluded that quality of contact predictions is independent of the sequence separation above a given threshold, whose extent depends on the protein length (using a sequence separation >80 residues, only those proteins whose length is >170 residues give significant results).

CASP3 common targets

Both groups (Olmea and Valencia, 1997; Fariselli and Casadio, 1999) sent predictions to the Critical Assessment of techniques for protein Structure Prediction (CASP3) competition (Orengo *et al.*, 1999). At that time there was not a common criterion for the prediction of contact maps, in terms of number of predicted contacts, minimal sequence separation and contact definition. Then, in order to compare the accuracy we have re-calculated the predictions in a consistent way. This is done by scoring the old submissions using a number of predicted contacts equal to half of the length of the protein, a minimal sequence separation of 7 residues and a contact definition based only on C_β with a threshold value of 8 Å. Among the submitted targets only three of them were common to the predictions of both groups, namely T0067, T0077 and T0079 (Table V). The new method (NETCSS with filter) for two of the three targets predicts with an accuracy level much higher than the two single methods separately (Table V). Interestingly, for the target T0079 the method based on correlated mutations (and filtered) ranks higher than neural networks. However, the predictor described in this paper scores on the targets with a very high efficiency, which is even greater than the average values obtained on our database. The predictions obtained with NETCSS on the three targets are highlighted in Figures 5–7, where all the predicted contacts are shown using a 'sticks' representation (the correct contacts are depicted in black).

Conclusions

It has been previously shown that neural networks are efficient tools in solving several kinds of problems, including predictions of secondary structures (Rost and Sander, 1995; Jones, 1999), transmembrane helices (Rost *et al.*, 1996) and folding initiation sites (Compiani *et al.*, 1998). Neural networks have been used in pioneer work by Bohr *et al.* (Bohr *et al.*, 1990) to predict distances between amino acids of homologous protein sequences and also to represent knowledge based potentials (Grossman *et al.*, 1995). In all this work it is clearly demonstrated that neural network-based predictors perform with a better efficiency when more information is given in input to

Table V. Comparison of FNETCSS with previously described methods on the same CASP3 targets

Target		Method								
Name	Protein length	Olmea and Valencia (1997)			Fariselli and Cassadio (1999)			FNETCSS (filter)		
		$<A>$	$<R>$	$<X_d>$	$<A>$	$<R>$	$<X_d>$	$<A>$	$<R>$	$<X_d>$
T0067	187	0.14	4.7	7.4	0.18	6.0	8.9	0.38	12.7	17.5
T0077	105	0.06	1.2	0.4	0.15	3.1	7.7	0.43	8.8	15.7
T0079	129	0.25	12.5	6.9	0.09	4.3	8.1	0.12	5.7	7.6
Average		0.15	6.1	4.9	0.14	4.5	8.2	0.33	9.1	13.6

For notations see Table II. T0067, phosphatidylethanolamine-binding protein, *Homo sapiens*, PDB code: 1BD9; T0077, ribosomal protein L30, *Saccharomyces cerevisiae*; T0079, MarA protein, *Escherichia coli*, PDB code: 1BL0.

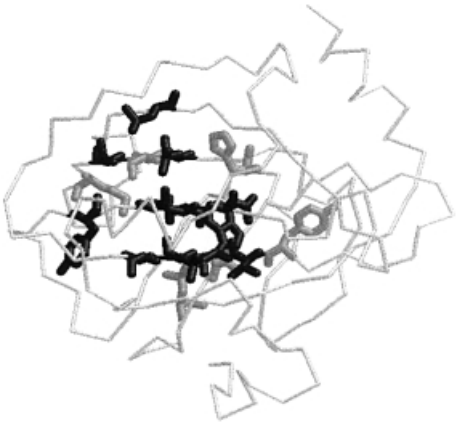

Fig. 5. Localization of the predicted contacts of the CASP3 target T0067 (phosphatidylethanolamine-binding protein, 1BD9) with FNETCSS. The prediction accuracy for this protein is $A = 0.38$, $R = 12.7$, $X_d = 17.5$. Side chains of correctly and wrongly predicted contact residues are highlighted in black and grey, respectively.

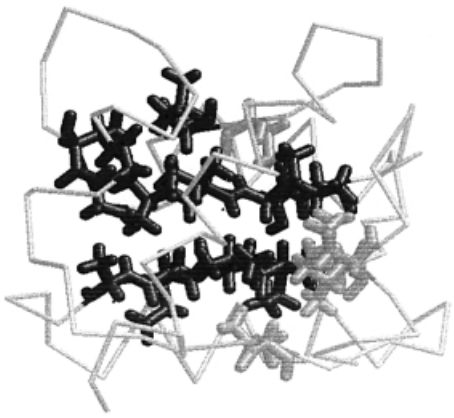

Fig. 6. Localization of the predicted contacts of CASP3 target T0077 (ribosomal protein L30) with FNETCSS. The prediction accuracy for this protein is $A = 0.43$, $R = 8.8$, $X_d = 15.7$. Side chain details are as in Figure 5.

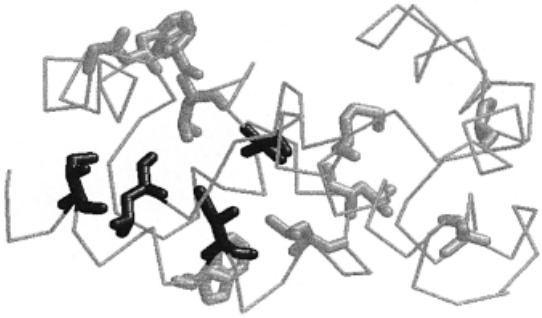

Fig. 7. Localization of the predicted contacts of CASP3 target T0079 (MarA protein, 1BL0) with FNETCSS. The prediction accuracy for this protein is $A = 0.12$, $R = 5.7$, $X_d = 7.6$. Side chain details are as in Figure 5.

the system. Therefore, we took advantage of the adaptability of neural networks to re-address the problem of predicting protein contact maps. The purpose of the present work is to use as input to the networks all the 'orthogonal' information left behind in our first approach in which neural networks were trained only with evolutionary information (Fariselli and Casadio, 1999).

From the CASP3 results of contact map predictions (Orengo *et al.*, 1999), it turned out that the method based on correlated mutations (Olmea and Valencia, 1997) was the best performing of the predicted contacts with an accuracy equal to 0.15. However, the number of contacts predicted was less than that obtained with other methods. On the other hand, the method based on neural networks (Fariselli and Casadio, 1999) performed best when the statistical significance of the predictions was evaluated with a self-threading method (Orengo *et al.*, 1999) aimed at recognizing the correct threading of the protein sequence onto its 3D structure from all the alternative threadings.

In this paper we combine both methods and show that neural networks can predict protein contacts with an accuracy of 0.21 starting from the residue sequence and including residue conservation, correlated mutations and predicted secondary structures. With this approach, prediction is improved as compared to that of a random predictor by a factor >6, previously obtained with neural networks using as input only evolutionary information (Fariselli and Casadio, 1999). Our results are even more important if we consider that the score is independent of the information on sequence separation, making the predictor of a more general use than before.

At the CASP3 competition it has been shown that the information on sequence separation also strongly affected the prediction of distances between residues made by other neural network-based predictors (Lund *et al.*, 1997).

A substantial improvement of the prediction efficiency is obtained by filtering the network outputs using the filtering procedure previously introduced by Olmea and Valencia (Olmea and Valencia, 1997) and based on the residue-coordination numbers. In this case, the neural network-based predictor increases its accuracy up to 0.25 for all proteins, with a deviation from a random predictor of a factor of 8. Even though this level of accuracy is still not sufficient for folding a protein, we believe that this is a tangible step forward in predicting contact maps of proteins.

Acknowledgements

Financial support to this work was provided to R.C. by a grant from the Ministero della Università e della Ricerca Scientifica e Tecnologica (MURST) to the project 'Structural, Functional and Applicative Prospects of Proteins from Thermophiles' and by a grant for a target project in biotechnology from the Italian Centro Nazionale delle Ricerche (CNR). We thank the Italian Ministero della Università e della Ricerca Scientifica e Tecnologica and the Spanish Minister of Research for supporting the joint collaboration between Italy and Spain.

References

Anfinsen,C.B. (1973) *Science*, **181**, 223–230.
Bohr,H., Bohr,J. Brunak,S. Cotterill,R. Fredholm,H. Lautrup,B. and Petersen,S.B. (1990) *FEBS Lett.*, **261**, 43–46.
Bohr,J., Bohr,H., Brunak,S., Cotterill,R.M.J., Fredholm,H., Lautrup,B. and Petersen,S.B. (1993*) J. Mol. Biol.*, **231**, 861–869.
Compiani,M., Fariselli,P., Martelli,P. and Casadio,R. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 9290–9294.
Eisenhaber,F., Persson,B. and Argos,P. (1995) *Crit. Rev. Biochem. Mol. Biol.*, **30**, 1–94.
Fariselli,P. and Casadio,R. (1998) In *Proceedings of World Multiconference on Systemics*, *Cybernetics and Informatics (SCI'98*, *Orlando USA)*, Vol. 1, pp. 527–533.
Fariselli,P. and Casadio,R. (1999) *Protein Eng.*, **12**, 15–21.
Fariselli P, Compiani M. and Casadio R. (1993) *Eur. Biophys. J.*, **22**, 41–51.
Kuntz,I.D., Thomason,J.F. and Oshiro,C.M. (1989) *Methods Enzymol.*, **177**, 159–205.
Govindarajan,S. and Goldstein,R. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 5545–5549.
Göbel,U., Sander,C., Schneider,R. and Valencia,A. (1994) *Proteins*, **18**, 309–317.
Godzik,A., Skolnick, J., and Kolinski,A. (1992) *J. Mol. Biol.*, **227**, 227–238.
Gorodkin,J., Lund,O. Andersen,C.A. and Brunak,S. (1999) *Proc. Int. Conf. Intell. Sys.*, **6**, 95–105.
Grossman,T., Farber,R. and Lapedes,A. (1995) *Mol. Biol.*, **3**, 154–161.
Hobohm,U. and Sander,C. (1994) *Protein Sci.*, **3**, 522–524.
Huang,E.S., Subbiah,S. and Levitt,M. (1995) *J. Mol. Biol.*, **252**, 709–720.
Jacoboni,I., Martelli,P.L., Fariselli,P., Compiani,M. and Casadio,R. (2000) *Proteins*, **41**,535–544
Jones,D.T. (1999) *J. Mol. Biol.*, **292**, 195–202.
Lund,O., Frimand,K., Gorodkin,J., Bohr,H., Bohr,J., Hansen,J. and Brunak,S. (1997) *Protein Eng.*, **10**, 1241–1248.
Maiorov,V.N. and Crippen,G.M. (1992) *J. Mol. Biol.*, **227**, 876–888.
McLachlan,A. (1971) *J. Mol. Biol.*, **61**, 409–424.
Miyazawa,S. and Jernigan,R.L. (1999) *Proteins*, **36**, 357–369.
Mirny,L. and Domany,E. (1996) *Proteins*, **26**, 391–410.
Mirny,L.A. and Shaknovich,E.I. (1996) *J. Mol. Biol.*, **264**, 1164–1179.
Orengo,C.A., Bray,J.E., Hubbard,T., LoConte,L. and Sillitoe,I. (1999) *Proteins*, **37**, 149–170.
Olmea,O. and Valencia,A. (1997) *Fold. Des.*, **2**, S25–S32.
Pazos,F., Helmer Citterich,M., Ausiello,G. and Valencia,A. (1997) *J. Mol. Biol.*, **271**, 511–523.
Rost,B. and Sander,C. (1995) *Proteins*, **3**, 295–300.
Rost,B., Fariselli,P. and Casadio,R. (1996) *Protein Sci.*, **5**, 1704–1718.
Sander,C. and Schneider,R. (1991) *Proteins*, **9**, 56–68.
Sander,C. and Schneider,R. (1993) *Nucleic Acids Res.*, **21**, 3105–3109.
Selbig,J. and Argos,P. (1998) *Proteins*, **31**, 172–185.
Shindyalov,I.,N., Kolchanov,N.A. and Sander,C. (1994) *Protein Eng.*, **7**, 49–358.
Sippl,M.J. (1990) *J. Mol. Biol.*, **213**, 859–883.
Taylor,W.R. and Hatrick,K. (1994) *Protein Eng.*, **7**, 341–348.
Thomas,D.J., Casari,G. & Sander,C. (1996) *Protein Eng.*, **9**, 941–948.
Vendruscolo,M., Kussell,E. and Domany,E. (1997) *Fold Des.*, **2**, 295–306.
Zhang,C. and Sung-Hou,K. (2000) *Proc. Natl Acad. Sci. USA*, **97**, 2550–2555.

Received November 30, 2000; revised June 19, 2001; accepted July 10, 2001