8 Protein Contact Map Prediction

Xin Yuan and Christopher Bystroff

8.1 Introduction

Proteins are linear chains that fold into characteristic shapes and features. To understand proteins and protein folding, we try to represent the protein molecule in such a way that its features are easy to see and manipulate. A simple representation facilitates algorithm design for structure prediction. The simplicity of the three-state character string representation of secondary structure is part of the reason for secondary structure prediction receiving so much attention early in the era of computational biology. One-dimensional strings are easily understood, parsed, mined, and manipulated. But secondary structure alone does not tell us enough about the overall shapes and features of a protein. We need a simple way to represent the overall tertiary structure of a protein.

Here we explore a two-dimensional Boolean matrix representation of protein structure, where each dimension is the residue number and each value is true if the residues are spatial neighbors and false otherwise—called a "contact map." A contact map is the simplest representation of a protein that can be faithfully projected back into three dimensions. As such it has received increased attention in recent years from bioinformaticists, who see this as a data structure that is readily amenable to data mining and machine learning.

The first goal of this chapter is to introduce the contact map data structure, how it is calculated from the three-dimensional structure and how it is transformed from two dimensions into three. Then we will explore a series of computational methods that have attempted to predict contact maps directly from the primary sequence, with or without the help of template structures from the protein database. Next, we will discuss the various ways that contact map predictions may be evaluated for accuracy. Finally, we will present some of the other ways contact maps have proven useful.

8.2 Definition of Interresidue Contacts and Contact Maps

Interresidue contacts have been defined in various ways. Routinely, a contact is said to exist when a certain distance is below a threshold. The distance may be that between Cα atoms (Vendruscolo et al., 1997), between the Cβ atoms (Lund et al., 1997; Thomas et al., 1996; Olmea and Valencia, 1997; Fariselli et al., 2001a,b), or it may be the minimum distance between any pair of atoms belonging to the side chain or to the backbone of the two residues (Fariselli and Casadio, 1999; Mirny and Domany, 1996). The following definitions of the interresidue distance D_{ij} have appeared at some time in the literature, each associated with a threshold distance:

1. The distance between alpha carbons (CA)
2. The distance between beta carbons (CB), using alpha carbon for glycine
3. The minimum distance between the van der Waals (vdW) spheres of heavy atoms (HS)
4. The minimum distance between the vdW spheres of backbone heavy atoms (BB)
5. The minimum distance between the vdW spheres of side-chain heavy atoms or alpha carbons (SC)

By one informed account, the best definition for interresidue distance is SC, the minimum distance between side-chain or alpha-carbon atoms (Berrera et al., 2003). Using a cutoff distance of around 1.0 Å between vdW spheres, SC-based contact maps efficiently recognized homologue sequences in a "threading" experiment, where a query sequence is assigned an energy score for every possible alignment of the sequence to a set of template structures. In a threading experiment, the definition of a contact determines which residues in the sequence are used to sum the energy. Using the minimum distance between residues and using a short distance cutoff makes the definition of a contact more energetically realistic, and this makes the sum of amino acid contact potentials a better approximation of the true energy. Contact potentials are energy functions that measure the pairwise side-chain-dependent free energy of residue–residue contacts, irrespective of the side-chain conformations. Contact potentials may be derived from contact maps statistics, as described in a later section.

Simpler, backbone atom-based definitions (CA or CB) with longer distance cutoffs are more readily projected into three dimensions, since the atomic positions used to calculate the distance depend only on the backbone angles. CB is slightly more meaningful than CA in an energetic sense, since side chains that point toward each other, and therefore have a shorter CB distance than CA distance, are more likely to make an actual physical contact. Using the minimum distance measures (HS, BB, or SC) can make projection into three dimensions more difficult because we have not saved the precise identity of the contacting atoms in the Boolean matrix. CB distances with a cutoff of 8 Å were chosen for use in the Critical Assessment of Structure Prediction (CASP) experiments (Moult et al., 2003), and this is the definition that we will discuss here, unless otherwise specified.

Having defined what we mean by the distance D_{ij}, the definition of a contact map is a straightforward distance threshold. For a protein of N amino acids the contact map is an $N \times N$ matrix C whose elements are given, for all $i, j = 1, \ldots, N$, by

$$C_{ij} = \begin{cases} 1 & \text{if } D_{ij} < Dcutoff \\ 0 & \text{otherwise} \end{cases} \tag{8.1}$$

The set of all D_{ij} is commonly referred to as the "distance matrix." Therefore, we can think of C_{ij} as a *thresholded distance matrix*. The mean difference between two distance matrices is sometimes called the “distance matrix error” (DME), as follows:

$$DME\,(a,b) = \frac{\sum\limits_{i=i}^{N-loc} \sum\limits_{j=i+loc}^{N} \left|D_{ij}^{a} - D_{ij}^{b}\right|}{0.5\ (N - loc - 1)\ (N - loc)} \tag{8.2}$$

DME is variously defined as the average of absolute differences or the root-mean-square distance difference, often with a cutoff (*loc*) to exclude local distances. The DME can be shown to correlate with the root-mean-square deviation (RMSD) in atomic positions if both numbers are derived from the same structures:

$$RMSD\,(a,b) = \sqrt{\frac{\sum\limits_{i=1,N} \left(x_i^a - x_i^b\right)^2}{N}} \tag{8.3}$$

By association, since the contact map error (CME) is a crude approximation of the DME, we can say that the sum of differences between two contact maps is a crude approximation of the RMSD between the two proteins they represent:

$$CME\,(a,b) = \frac{\sum\limits_{i=1}^{N-loc} \sum\limits_{j=i+loc}^{N} \left|C_{ij}^{a} - C_{ij}^{b}\right|}{0.5\ (N - loc - 1)\ (N - loc)} \tag{8.4}$$

But this is at best a rough correlation, and then only under the special constraint that each contact map C_{ij} is derived from a 3D structure. As we will discuss later, a simple measure such as CME by itself is usually not a good indicator of structural prediction accuracy. This topic is discussed again in Section 8.5.

8.3 Features of a Contact Map

Contact maps and distance matrices are “internal coordinates,” and as such are independent of the reference frame of the Cartesian atomic coordinates. This frame invariance, plus the Boolean property, makes contact maps attractive to practitioners of machine learning and data mining techniques. Patterns within contact maps are meaningful even when taken out-of-context.

It is well-known that the number of contacts scales linearly with the chain length (Thomas et al., 1996; Vendruscolo et al., 1997; Fariselli and Casadio, 1999). The slope of the linear dependence depends only on how a contact is defined. Using CB distances and a cutoff distance of 8 Å, and ignoring local contacts with $|i - j| < 3$, the number of contacts in a compact globular protein is approximately 3.0 times the length of the protein, with a relatively small standard deviation of ±0.4. Since every contact involves two residues, this number implies an average of about 6 (±0.8)

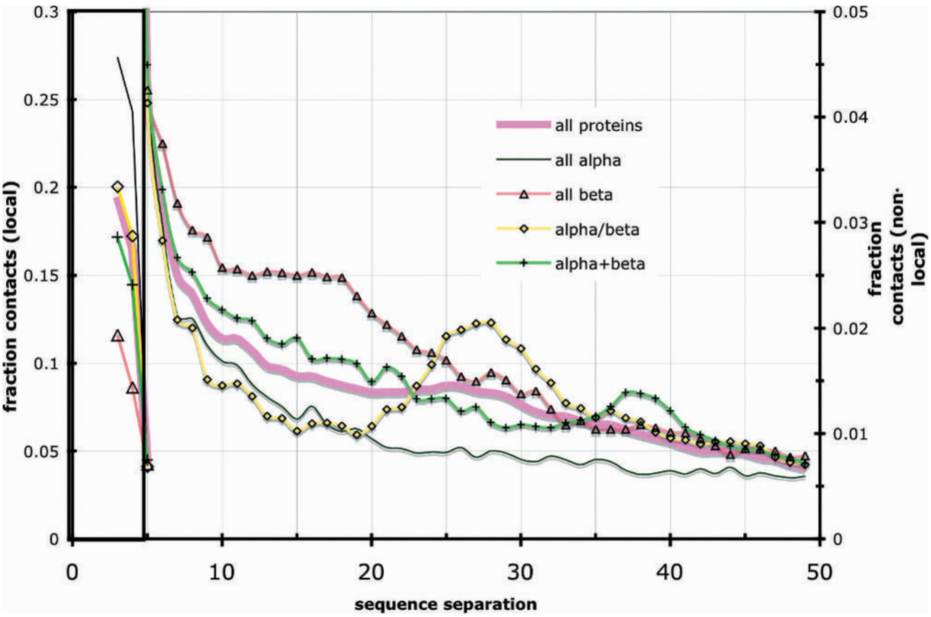

Fig. 8.1 Fraction of all CB contacts with cutoff distance 8.0 Å as a function of sequence separation distance for the four main SCOP classes of proteins. About half of all contacts are local ($3 \leq |i - j| \leq 5$, left axis). Different fold classes have significant differences in the contact profile. The peaks at around 28 in alpha/beta proteins correspond to the sequence distance where parallel strands are separated by one alpha helix, called βαβ-units.

contacts per residue. These numbers are consistent across all fold classes, probably reflecting the invariant packing density and size of amino acids. Parallel α/β proteins deviate most from this average, with an average of 3.3 contacts per residue, but this difference is less than one standard deviation. There are many protein chains with far fewer than three contacts per residue but these are generally not globular domains. Instead they are often parts of larger complexes which, when taken together, also average six contacts per residue.

Most interresidue contacts in proteins are local, and the likelihood of finding a contact drops quickly as the sequence distance between residues increases. There are interesting and obvious class-dependent differences in the sequence separation profile of contacts (Fig. 8.1). This distribution is important to consider when assessing the accuracy of contact map predictions, since local contacts are easier to predict than nonlocal ones. The "contact order" of a protein is defined as the average sequence distance between contacting residues, and this number has been shown to correlate with the folding rate for many small proteins (Plaxco et al., 1998). Some studies use the contact order as a measure of the topological compexity of the fold (Kuznetsov and Rackovsky, 2004; Punta and Rost, 2005). Recently, the notion of contact order has been refined to take nested loop closures into account, giving an "effective contact order" which is probably a much better measure of fold complexity (Chavez

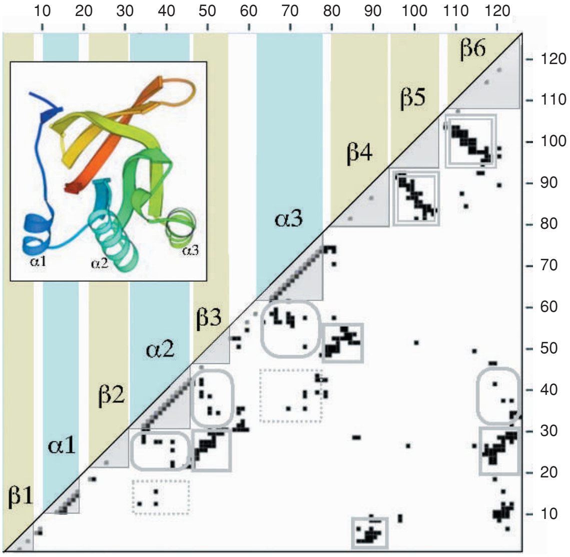

Fig. 8.2 Contact map of glutathione reductase, domain 2 (PDB code 3GRS, residues 166–290). Black boxes are contacts, gray boxes: $i, i+3$ contacts, shaded triangles: contacts within secondary structure elements, gray rectangles: parallel beta-strands, double rectangles: antiparallel beta-strands, dotted rectangles: helix–helix contacts, rounded rectangles: helix–strand contacts. Inset: Molscript (Kraulis, 1991) drawing of 3GRS structure.

et al., 2004). In this study it was understood that the contact order should reflect the configuration entropy lost on the formation of a contact. The effective contact order is the entropy of the closure of a loop that may already contain contacts within it.

A trained eye can identify secondary structure elements in a contact map by looking at the local contacts, i.e., those near the diagonal of the matrix. A helix has an unbroken row of contacts between i, $i \pm 4$ pairs. Extended strands have no local contacts with $3 < |i - j| < 5$, although occasional $i, i+3$ contacts occur in β strands where β-bulges or β-bends occur. Loops have some local contacts but never an unbroken row. Figure 8.2 shows images of common contact patterns that are found between secondary structure elements. Antiparallel and parallel β strands give rise to unbroken rows of contacts in the off-diagonal region. A row of contacts that is perpendicular to the diagonal of the matrix represents a pair of antiparallel strands. These are contacts between residues $i + k$ and $j - k$, where k goes from zero through the length of the strand pairing. Similarly, a row of contacts that is parallel to the diagonal represents a pair of parallel strands, with contacts between $i + k$ and $j + k$. Consequently, β sheets appear as a set of perpendicular or parallel rows of contacts. The strand order can be determined by tracing the pairing interactions (gray rectangles in Fig. 8.2). Contacts between α-helices and other secondary structure elements appear as broken rows or "tire tracks." If the two contacting elements are both helices, then the contacts appear every three or four residues in both directions, following the periodicity of the helix. If one of the elements is a strand, then we see

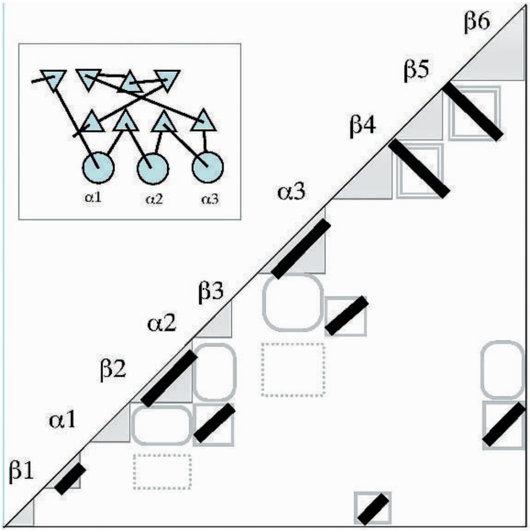

Fig. 8.3 Idealized features in contact maps (thick bars) may be converted to a topological cartoon (Michalopoulos et al., 2004) using simple drawing conventions.

a periodicity of two in the contacts in that direction, since the side chains in a strand alternate sides of the sheet. Domains can be seen as regions of the chain that have dense contacts, since intradomain contacts outnumber interdomain contacts.

If there is additional knowledge to resolve the ambiguity in overall handedness, then the entire molecule can be reconstructed by hand from a contact map. For example, for α/β proteins we can assume that any parallel β-α-β unit has a right-handed crossover (more than 99% of all parallel β-α-β supersecondary structure units are right-handed). If our assumption is right, then we know on which side of the sheet to place the helix. The presence or absence of helix–helix contacts can be used to resolve the placement of any additional helices with respect to the sheet. However, without some external information about either the overall handedness or the handedness of any substructure, two mirror-image reconstructions are possible. Figure 8.3 shows an idealized contact map, the same one shown in Fig. 8.2, and the corresponding protein topology (TOPS) cartoon (Michalopoulos et al. 2004) that can be drawn using only the simplified contact map. Although TOPS cartoons such as this one cannot be accurately projected to three dimensions without additional information such as key contacts, the TOPS graph structures allow the easy visualization of common topological features in proteins.

8.4 From Contact Map Prediction to 3D Structure

Contact maps that are derived from 3D protein structures can be mapped back to their corresponding structures by taking advantage of the known stereochemistry of amino acids and proteinlike backbone angles. But not all square, symmetrical Boolean matrices map to 3D objects, much less to proteinlike objects.

In mathematical terms, a contact map is an undirected graph, where the vertices are the residues and the edges are the residue–residue contacts. But contact maps that have been derived from true protein structures, or from any other set of points in three dimensions, are a special subset of all undirected graphs called "sphere intersection graphs" or "sphere of influence graphs" (SIGs) (Michael and Quint, 1999). In a SIG the edges represent the intersections of fixed-radius spheres. If a graph is a SIG, then at least one solution exists for the positions of the vertices in 3D. The thresholded distances from the solution configuration must correspond to the contacts in the contact map exactly, or the graph is not a SIG. If there is no solution, then the contact map without modification cannot represent a protein, or for that matter, any set of points in 3D! However, there may exist a subset of the contacts that can potentially represent a protein. The problem of mapping a predicted contact map to 3D is the problem of finding the best SIG within a contact map.

Determining whether or not a contact map is a SIG remains an open problem for the general case (Michael and Quint, 1999). But heuristic methods can be applied that use additional information about proteins, including the key facts that (1) adjacent vertices are linked with their distance fixed at 3.8 Å and (2) that all nodes are self-avoiding (i.e., no two nodes can be closer than 3.8 Å). Figure 8.4 illustrates the key constraints on a proteinlike SIG. In addition to these constraints, proteins have

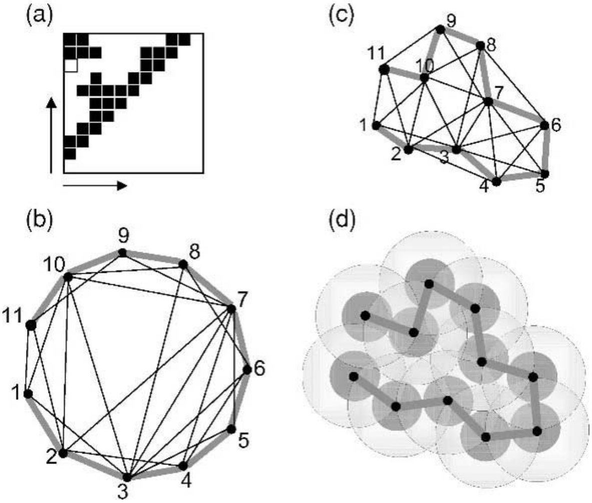

Fig. 8.4 A proteinlike sphere intersection graph (SIG). For a contact map (a) can always be projected to an undirected graph where the vertex positions satisfy nearest neighbor distance constraints and self-avoidance (b). This contact map is a proteinlike SIG because vertex positions are possible (c) such that each edge distance corresponds to a sphere intersection (d, large circles) and all vertices are mutually avoiding (dark circles). The addition of a single contact between 1 and 9 [white box in (a)] breaks the SIG.

characteristic secondary structures and turns that are sequence dependent and restrict the way nodes can be arranged locally along the chain. So while there is still no general solution for finding a SIG within a contact map, the problem of finding a "proteinlike SIG" seems tractable and it is likely it will be solved in the near future.

If the contact map is a proteinlike SIG, then it is possible to reproduce, with considerable accuracy, the 3D structure of the protein's backbone from its contact map (Havel et al., 1979; Saitoh et al., 1993). And at least one heuristic approach has been shown to work in the presence of "noise" contacts, accurately excluding random physically impossible contacts that were added to a true protein contact map (Vendruscolo et al., 1997; Vendruscolo and Domany, 1998). Vendruscolo's method works by minimizing a cost function that contains only geometric constraints, nothing resembling the true energies of the polypeptide chain. The task of predicting the tertiary structure of a protein is split into two steps, making it a crude pathway model. First, a reliable prediction of secondary structure must be realized, then a coarse-grained contact map is used to select contacts between the secondary structure elements. The method succeeds even when up to 10% of the contacts are "noise." Interestingly, it is now possible to reconstruct a contact map from a 1D representation consisting of principal eigenvectors (PE) derived from HS contact maps (Porto et al., 2004). The PE reconstruction of the contact combined with 3D projection using Vendruscolo's method builds models that are typically within RMSD 2.0 Å of the original structure. Unfortunately, there is still a large gap between the prediction accuracy necessary for a good 3D reconstruction and the prediction accuracy possible using today's methods. Worse than that, the distribution of erroneous contact predictions in real cases is probably not random, as this reconstruction algorithm assumes.

8.5 Contact Map Prediction

Contact prediction offers a possible shortcut to predict protein tertiary structure. Over the years, a variety of different approaches have been developed for contact map prediction including neural networks (Fariselli et al., 2001a,b; Pollastri and Baldi, 2002; Lund et al., 1997), support vector machines (Zhao and Karypis, 2003), and association rules (Zaki et al., 2000). Statistical approaches have also been tried, including correlated mutations (Olmea and Valencia, 1997; Thomas et al., 1996; Singer et al., 2002), knowledge-based potentials (Sippl, 1990; Park et al., 2000), and hidden Markov models (Shao and Bystroff, 2003). Statistical pair potentials do not produce sufficiently specific contact predictions. More specific information appears to come from neighboring residues and patterns of mutation, sequence conservation, and predicted secondary structure, all obtainable from multiple sequence alignments. The various features include contacts from patterns of conserved hydrophobic amino acids (Aszodi et al., 1995), sequence profiles derived from multiple sequence alignment (Fariselli et al., 2001a,b; Pollastri and Baldi, 2002; MacCallum, 2004; Hamilton et al., 2004; Shao and Bystroff, 2003), distribution of distances in

Table 8.1 Available servers for contact map predictions

Server	URL	Reference(s)
CORNET	gpcr.biocomp.unibo.it/cgi/predictors/cornet/pred_cmapcgi.cgi	Olmea & Valencia, 1997, Fariselli & Casadio, 1999
PDG	www.pdg.cnb.uam.es:8081/pdg_contact_pred.html	Pazos et al., 1997
HMMSTR	www.bioinfo.rpi.edu/∼bystrc/hmmstr/server.php	Shao & Bystroff, 2003
GPCPRED	sbcweb.pdc.kth.se/cgi-bin/maccallr/gpcpred/submit.pl	MacCallum, 2004
PoCM	foo.acmc.uq.edu.au/∼nick/Protein/contact.html	Hamilton et al., 2004
CMAPpro	www.ics.uci.edu/∼baldig/	Cheng et al., 2005

proteins with known structures (Tanaka and Scheraga, 1976; Wako and Scheraga, 1982; Huang et al., 1995; Mirny and Domany, 1996; Maiorov and Crippen, 1992), correlated mutation and/or combination with other features (Olmea and Valencia, 1997; Fariselli et al., 2001a,b; Pollastri and Baldi, 2002; Hamilton et al., 2004; Göbel et al., 1994; Neher, 1994; Shindyalov et al., 1994), secondary structure information (Shao and Bystroff, 2003; Zaki et al., 2000; Hamilton et al., 2004; Fariselli et al., 2001a,b; Olmea and Valencia, 1997; Zhang and Kim, 2000; Hu et al., 2002). Beyond ones and zeros of a contact map, knowledge-based estimates of residue–residue distance have been used to determine the approximate structure of proteins (Skolnick et al., 1997; Wako and Scheraga, 1982; Monge et al., 1994; Aszodi et al., 1995).

The results in CASP5 (Aloy et al., 2003) and CASP6 (Graña et al., 2005) suggest that there has been at best a very limited improvement for *de novo* contact prediction methods. In the following sections we summarize a few of these approaches to contact map prediction in detail, with an eye toward possible improvements. Table 8.1 lists the currently available web servers for contact map prediction.

8.5.1 Contact Prediction Using Statistical Models

In sequence alignments, some pairs of positions appear to covary in a physicochemically plausible manner, i.e., a "loss of function" point mutation may be rescued by an additional mutation that compensates for the change (Altschuh et al., 1987). Compensating mutations would be most effective if the mutated residues were spatial neighbors; therefore, "correlated mutations" across evolutionary distance should imply spacial proximity. Attempts have been made to quantify this hypothesis and to use it for contact predictions (Neher, 1994; Göbel et al., 1994; Taylor and Hatrick, 1994).

Direct statistical methods require pairwise scoring matrices to compute the contact scores. The scoring matrices are based on *a priori* models of noncovalent residue interactions and/or protein evolution. In various approaches, the matrices have been based on amino acid identity (Shindyalov et al., 1994), amino acid substitution probabilities (Göbel et al., 1994), contact substitution probabilities (Rodionov and Johnson, 1994), biophysical complementarity of electrostatic charge and side chain volume (Neher, 1994), or statistics from evolutionary models (Singer et al., 2002). In the latter case, the energetic value of a contact was estimated as a likelihood matrix, using a large set of proteins of known structure. Mutations are correlated because side-chain interactions have an energetic value, and this energetic value is therefore reflected in the database contact statistics (Fig. 8.5). The predicted target contact energies were calculated by first generating a multiple sequence alignment and then summing the likelihood of all residue pairs in the corresponding columns. The likelihood approach performed better when contacts were local in the sequence, but tended to perform poorly on nonlocal contacts. If combined with other features, the method could give better predictions.

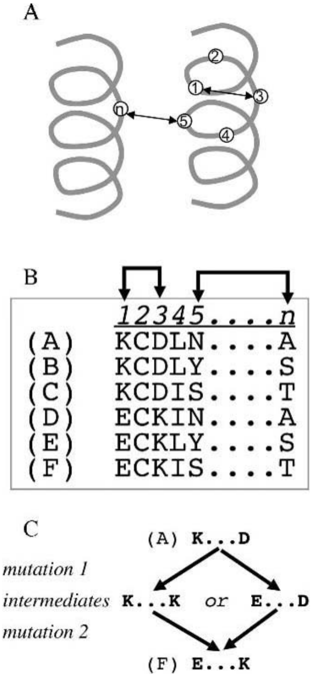

Fig. 8.5 Illustration of correlated mutation theory and application. (A) Several residues are shown in their structure context, in this example, two nearby α-helices. (B) For these, six sequences (A–F) are shown as a multiple alignment. Positions 1 and 3 show correlated substitutions (connected by arrows), as do positions 5 and *n*. (C) The most parsimonious evolutionary pathways are between sequences A and F, for positions 1 and 3. Correlated mutation detects pairs of residue positions that show correlated substitutions without intermediates. The theory is that when a mutation occurs in a structurally important residue (mutation 1), the intermediate has structural instability. Compensatory mutations are then selected (mutation 2) and the structural interaction is restored. Any intermediates are eventually eliminated from the sequence record due to reduced fitness. (Based on a figure from Singer et al., 2002.)

The correlated (or compensatory) mutation information is generally weak. Contact prediction can be improved by combining correlated mutations with other data such as sequence conservation and contact density information (Hamilton et al., 2004). The principle behind contact density is simple. If two nonadjacent residues are in contact, then we expect that the residues adjacent to them will also be in contact with a high probability. Correlated mutations have been combined with other sources of information in some of the methods described in the following sections.

A simpler statistical method is the sequence conservation at single positions. The success of the evolutionary trace method (Lichtarge et al., 1996) in identifying localized side chains based on functional conservation in protein sequence families shows that sequence conservation is both biologically and statistically significant when combined with known structure. In this method, conserved positions are mapped to the surface of a known protein and clustered to find functional sites. In practice, sequence conservation is not used alone but rather as a component of the training data from neural networks, described in the next section.

8.5.2 Contact Maps from Neural Networks

Both the correlated mutation and likelihood approaches performed best on local contacts, but tended to perform poorly on longer sequences where many contacts were nonlocal. Another approach to the problem has been to train neural networks with various encodings of multiple sequence alignments with other inputs such as predicted secondary structure (Fariselli and Casadio, 1999; Fariselli, 2001a,b). These tended to perform better over a wide range of sequence lengths. Fariselli's CORNET predictor claims to have the best contact prediction results to date. It was specifically designed to include evolutionary information in the form of a sequence profile, sequence conservation, correlated mutations, and predicted secondary structures. Sequence conservation was taken from the HSSP database (Dodge et al., 1998). Correlated mutations were calculated as previously described (Olmea and Valencia, 1997; Göbel et al., 1994). This neural network approach involved encoding frequencies of residues in columns of a multiple sequence alignment, as well as having inputs based on predicted secondary structures, length of input sequence, and residue separation. Briefly, each position in the alignment has a distance array that contains the interresidue distances between all of the possible pairs of sequences at that position. The distance between residues is defined using an early amino acid scoring function (McLachlan, 1971). 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. The network was trained by using the back-propagation algorithm, with a single output neuron coding for contact (1) and noncontact (0). Contacts were defined using Cβ atoms (CB) with an 8-Å cutoff, and only those separated by at least six residues were used. The hidden layer consisted of eight neurons. Each residue pair in the sequence was coded as an input vector of 210 elements ($20 \times (20 + 1)/2$), representing all possible pairs of amino acids. CORNET has an average off-diagonal (nonlocal) contact accuracy of 21%. While this result is more than six times better than a chance prediction, it is still far from providing sufficient accuracy for a reliable 3D reconstruction.

In GIOHMM (Pollastri and Baldi, 2002), a new neural net architecture was introduced. The contact matrix was represented as a 2D graph. It is implemented in two steps. The first step is the construction of a statistical graphical model (Bayesian network) for contact maps, where the states are arranged in one input plane, one output plane, and four hidden planes. The parameters of the Bayesian network are the local conditional probability distributions. The second step is the reparameterization of the graphical model using artificial recurrent neural networks. In the training of the neural net, the input includes the information for the contact, secondary structure, and solvent accessibility. The authors cite a prediction accuracy of 60.5% for CB contacts with an 8-Å cutoff and 45% for CB contacts with a 10-Å cutoff, but only local contacts were considered ($|i - j| < 7$). While intriguing, these numbers cannot be compared directly with those mentioned above. Prediction of local contacts is intermediate between secondary structure prediction, for which the highest three-state prediction accuracies average 75–80% (Jones, 1999), and nonlocal contact map prediction, for which a highest accuracy of 21% has been reported. The same group (P. Baldi) has recently released a new contact map predictor, CMAPpro, as part of a battery of tools for protein feature prediction (Cheng et al., 2005). The innovation in this neural net architecture is a heirarchical scheme where the output of local contact predictions is used as the input for predicting nonlocal contacts.

Lund et al. combined two independent data driven methods (Lund et al., 1997). The first used statistically derived probability distributions of the pairwise distance between two residues, similar to the knowledge-based pair potentials of Sippl (1990). The second consisted of a neural network with a single hidden layer connected to two three-residue windows a defined distance apart on the sequence. For both of these functions, the underlying physical determinants of the statistics are the various chemical affinities between short sequence patterns of amino acid side chains. Nonpolar side chains attract through the hydrophobic effect, polar side chains through hydrogen bonds and salt bridges. This affinity alone does not determine the likelihood of a contact but is combined with sequence separation distance, since the polypeptide chain has a certain degree of stiffness that limits the ways the side chains can come together when the loop is short. Their results showed that prediction by neural networks is more accurate than predictions by probability density functions. The accuracy of the prediction can be increased by using sequence profiles instead of single sequences.

As mentioned earlier, patterns of contacts form when an α-helix is in contact with a strand, a helix with a helix, or when two strands are paired in a β sheet. A recent study used a neural network approach to find patterns of correlated mutations (Hamilton et al., 2004). The main input to the neural network was a matrix of 25 mutational correlation values for a pair of five-residue windows centered on the residues of interest. Each entry in the matrix is the correlation between two residues (Göbel et al., 1994). This information was combined with other inputs such as predicted secondary structure using Psi-Pred (Jones, 1999; McGuffin et al., 2000), the type of amino acids, and the input sequence length. Using this method an average prediction accuracy of 21.7% was obtained. The accuracy was found to be relatively consistent across different sequence lengths, but to vary widely with the secondary structure. As with previous studies, contact predictions were found to be particularly difficult for α-helical proteins (Fariselli and Casadio, 1999, Fariselli et al., 2001b). Fariselli suggested that the poor predictions from their methods on this subset of proteins might be a result of the underrepresentation of α-type proteins in the training set. But in Hamilton et al., even if they trained the model on proteins of α-type to predict an α-type protein, no improvement in prediction accuracy was obtained. It might indicate that the patterns of contact are less locally defined in α-helical proteins and may require the window size to be larger. Alternatively, the predictions could be improved by finding a better measure of correlated mutations, and perhaps by applying the contact occupancy filtering as described in Olmea and Valencia (1997).

8.5.3 A Genetic Algorithm for β-Strand Contacts

MacCallum has noted that protein architectures impose regularities in local sequence environments (MacCallum, 2004). Based on the fact that many proteins have pairs of neighboring strands with similar sequence patterns, the GPCPRED algorithm used only sequence profile and residue separation information as input to a genetic programming approach to contact prediction. Sequence profiles are classified using a self-organizing map algorithm (SOM), and the new classes reveal a distinctive "striping" pattern across facing strand pairs. The predictions were equal to or better than existing automated contact predictors that use more fitting parameters. Predictions of sets of "$L/10$" contacts (i.e., number of contacts predicted equals length of protein over 10), each between positions separated by at least eight residues, were 27% correct for proteins up to length $L = 400$. As they suggest, the predictions could be improved if they included additional information such as sequence conservation and correlated mutations. As good as they are, the predictions cannot be uniquely mapped to three dimensions, but with an additional postprocessing step based on the packing rules, this could be remedied.

8.5.4 Contact Prediction Using Support Vector Machine

A support vector machine (SVM) is a method for binary classification in an arbitrary feature space and as such is well-suited for the contact map problem. In one study (Zhao and Karypis, 2003), contact and noncontact residue pairs were treated as positive and negative instances in a feature space comprised of position-dependent information for amino acid content, physicochemical environment, secondary structure, and evolutionary correlation. SVM was used to define an optimal multidimensional hyperplane for dividing contacts and noncontacts in the space of the features. The model was trained on all classes of protein structure in the CATH database (Orengo et al., 1997). The results indicated that the secondary structure feature is most helpful for contact prediction in proteins containing β strands. On the other hand, correlated mutations and sequence profile methods performed the best for proteins containing α-helices. Models learned separately for different protein classes might result in better performance in contact prediction.

8.5.5 Prediction Using Association Rules

Data mining was used to extract valuable information from true contact maps (Hu et al., 2002; Zaki et al., 2000) in the form of recurrent nonlocal contact patterns and sequence–contact association rules. Zaki et al. developed a string encoding and hashing technique to extract all of the nonlocal contact patterns for a sliding window across all contact maps of existing structures. The contact patterns were clustered based on their similarities, and sequence-to-contact relationships were expressed as logical statements, or association rules. By applying association rules to the output of the hidden Markov model HMMSTR (Bystroff et al., 2000), their contact map predictions had about 20% accuracy for $L/2$ contacts with $|i - j| \geq 4$, corresponding to about 20% coverage. Even with low coverage the predictions contained physically impossible combinations of contacts (see Fig. 8.6). To make the predictions more meaningful, there is a need to filter out the physically impossible contacts.

8.5.6 Prediction Using Pathway Models

A contact prediction method that makes use of sequence profiles, fragment templates, and pathway models was used for the first time in the CASP5 experiment (Shao and Bystroff, 2003), with accuracies comparable to or higher than previous approaches, depending on how accuracy is measured. In this prediction method, the first step is to assign a probability to each potential contact. The probability in this case is the database-derived likelihood of contact between any two local structure motifs.

Local structure motifs were predicted probabilistically as Markov states from the HMMSTR model (Bystroff et al., 2000). A matrix γ expresses the probability of each motif at each sequence position, solved using the Forward/Backward algorithm (Rabiner, 1989):

$$\gamma\,(i, q) = P(q|i) \tag{8.5}$$

Then the contact potential $G(p, q, s)$ between any two HMMSTR states p and q, given a sequence separation s, was calculated as the negative log of the sum over all joint probabilities as follows:

$$G(p, q, s) = -\log \frac{\sum_{CATH} \sum_{i \ni D_{i,i+s} < 8\mathring{A}}}{\gamma(i, p)\gamma(i + s, q)} \sum_{CATH} \sum_{i} \gamma(i, p)\gamma(i + s, q) \tag{8.6}$$

In the numerator, the sum is over all residue pairs $(i, i+s)$ that are in contact in all CATH proteins. In the denominator, the sum is over all residue pairs $(i, i+s)$. To predict contacts, we first calculate the contact potential E_{ij}, by summing $G(p, q, s)$ over all states p at i and all states q at j. E_{ij} may be thresholded to give a contact map prediction.

This algorithm implies that the local structure motif folds first, followed by motif–motif condensation to form larger units. Rule-based filtering techniques were applied to remove contacts that were impossible given a previously defined set of contacts. "Common sense" rules were applied. For example, any one β-strand residue may pair with at most two other β strands, not three of course. Other rules enforced the physically possible density of contacts and mutual contacts, and the triangle inequality. In addition, contacts were assigned only if they had an effective sequence separation of 8 or less after "loop closure," similar to the effective contact order (Chavez et al., 2004). This gave local contacts opportunity to form before assigning nonlocal contacts (Fig. 8.6). This simple folding pathway model was sufficient to extract the correct set of contacts for some but not all of the CASP5 targets (Shao and Bystroff, 2003; Bystroff and Shao, 2003). The most common error was the wrong

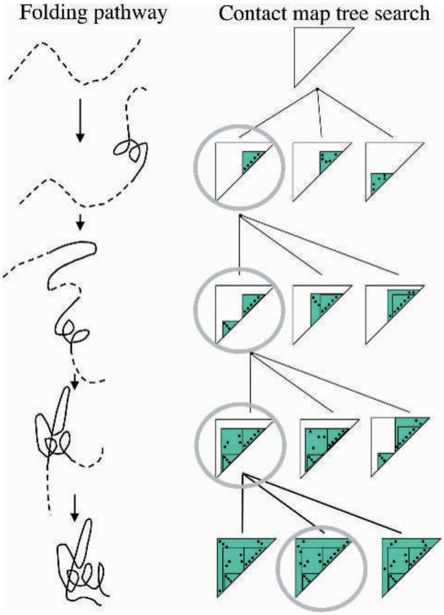

Fig. 8.6 In the HMMSTRCM ("hamster CM") method, a folding pathway is expressed as a tree search in contact map space where each branch represents the addition of new contacts to the previous set of contacts (shaded triangles, thick lines). An energy function may be applied to select among alternative sets of contacts. Local contacts (shaded triangles) form first. These come together (larger shaded triangles), subject to a set of simple rules. HMMSTRCM succeeded in cases where the initial contacts were correctly assigned, but could not overcome bad initial assignments.

choice of the nucleation site, since early errors propagated to further errors. A better means to choose the folding nucleation site would remedy this problem.

8.6 Evaluation of Contact Map Predictions

The current evaluation criteria used for contact map predictions include (Graña et al., 2005; Aloy et al., 2003; Koh et al., 2003): (1) accuracy: the number of correctly predicted contacts divided by the total number of predicted contacts; (2) coverage: the number of correctly predicted contacts divided by the total number of contacts; (3) improvement over random: the calculated accuracy divided by the random accuracy; and (4) the delta evaluation: the percentage of correctly predicted contacts that are within a certain number (delta) of residues of the experimental contact, measured along the sequence.

Another useful measure is the distance distribution of predicted contacts, X_d:

$$X_d = \sum_{i=1}^{15} \frac{P_{ip} - P_{ia}}{15 d_i} \tag{8.7}$$

where the sum runs over 15 distance bins covering the range from 0 to 60 Å. d_i is the distance representing each bin. P_{ip} is the percentage of predicted contacts whose true distance is in bin i. P_{ia} is the same percentage but for all of the residue pairs, not just contacts. Defined in this way, $X_d > 0$ indicates that more of the predicted contacts are either true contacts or close to being true contacts. $X_d \leq 0$ indicates that the contacts are random (Pazos et al. 1997).

Each of these criteria is well-behaved when the prediction is close to perfect, but they diverge to different extents from good behavior when the prediction strays from the path of perfection. A good predicted contact map should map uniquely to the correct 3D structure. Contact maps may be divided into blocks representing contacts between secondary structure elements (Figs. 8.2 and 8.3). If a prediction identifies most of the contact blocks but the overall accuracy is low, we may still recognize it as a good prediction that maps to a single correct structure. On the other hand, if the accuracy is high but correct contacts concentrated in the local region or in one block of the molecule, then it is less meaningful as a 3D structure. In our opinion, we might have included the "block count" as another evaluation criterion. It would be defined as the total number of counts of true contact blocks, with one count for each block. The higher the block count, the better the prediction is. Since the nonlocal contacts are harder to predict than the local ones, we might give more weight to the nonlocal contacts.

If we look at the parallels between contact map prediction and the solution of protein structures by NMR, we can see a more logical way of defining accuracy in contact map prediction. NMR structures are solved by applying distance geometry methods while minimizing a cost function that seeks to satisfy as many of the experimental distance constraints as possible. The result of distance geometry is an ensemble of possible structures where each structure is a local minimum of the distance geometry cost function and each structure satisfies the distance constraints to about the same extent. If the distance constraints from the NMR experiment are more self-consistent and mutually re-enforcing, then the ensemble is more tightly clustered, and the average pairwise RMSD between members of the ensemble is small. Some of the best NMR structures have ensembles with RMSDs around 1.0 Å, the worst have very high RMSDs approaching random. Most often this reflects the disorder in the polypeptide rather than the quality of the NMR data.

A contact map prediction represents an ensemble of states in the same way and for the same reason as a set of NMR distance constraints represents an ensemble of states. Therefore, it makes sense to measure the quality of a contact map prediction in the same way as we measure the quality of an NMR structure, by sampling an ensemble of 3D solutions and then measuring the diversity of the ensemble. If the ensemble is mostly disordered, then the DME metric might make more sense, or the size of the largest fragment with an average RMSD below a cutoff. In any case, a measure of contact map accuracy in 3D would alleviate the problems associated with 2D accuracy assessments, and the accuracy would better correlate with the usefulness of the prediction.

8.7 Other Applications of Contact Maps

Up to this point, we have confined the discussion to contact prediction in globular proteins, but the prediction of membrane protein structures is potentially much more valuable. Membrane protein structures are harder to characterize experimentally, since membranes interfere with both crystallization and NMR experiments. Electron microscopy, sometimes using monoclonal antibodies, and fluorescent resonance energy transfer (Eisenhawer et al., 2001) have been used with some success to obtain the gross layout of the transmembrane parts of membrane proteins, but in general these experiments are not sufficient to build a detailed model. Molecular simulations have been used successfully to refine the structure of the transmembrane regions (Enosh et al., 2004). One way contact map predictions can potentially be used is to assign a contact or noncontact value to residues in the soluble part of a membrane protein and use that information to predict contacts within the membrane. This would work because transmembrane regions are either helices or strands, and these generally pass directly through the membrane without any turns. Thus, contacts on the membrane surface imply contacts within the membrane and on the other side.

Contact maps have been used to align sequences to structures and to align structures to structure, even nonsequentially. The correlated mutation metric described in Fig. 8.5 ignores the identity of the amino acids involved. A more specific model for correlated mutations has been constructed in which a score for each of the possible amino acid substitution pairs was stored in a 400 × 400 contact substitution matrix called CAO (Contact Accepted mutatiOn) (Lin et al., 2003; Kleinjung et al., 2004). Each matrix element expresses the degree, positive or negative, to which the two mutations were observed in tandem in known structures. For example, the score for F_Y:I_V would be positive if a contact between an F and a Y was frequently observed to mutate F → I and Y → V as compared to random chance. CAO is not used, like correlated mutations, to predict contacts directly, but instead it is used to score sequence alignments to proteins of known structure. The greater sensitivity of this method allowed the authors to assign functional annotations to previously uncharacterized sequences with improved confidence.

We have used contact maps to align structures nonsequentially, and have applied the contact map alignment to search for conserved packing arrangements in protein cores (Yuan and Bystroff, 2005). The program SCALI (Structural Core ALIgnment) assembles a nonsequential alignment from a pairs-list of short gapless local alignments. Each of these short alignments relates some of the contacts in the target to some contacts in the template. The contact map score determines which segments to keep in a search through alignment space. The resulting alignment is nonsequential if the aligned segments are ordered differently in the two proteins. Nonsequential alignments do not imply homology, but may be used to find structural motifs. We used nonsequential alignments to find recurrent multibody interactions in protein cores.

8.8 Conclusions

Contact maps represent a useful and easily manipulated data structure for protein structure prediction by statistical, machine learning, and simulation methods. Progress is being made toward building predictive models that use this data structure, and new insights are being discovered about the nature of protein folding. Contact maps are bridging the gap between accurate 1D structure predictions and 3D structure predictions, but much work remains to be done. Here is a short list of open problems in contact map prediction as discussed in this chapter.

- *Scoring and error correction.* The impact of this representation on the field of protein structure prediction depends on advances in methods for correcting imperfect contact map predictions.
- *Nonlocal contacts.* Most methods are far more accurate on local contacts, but the global topology is defined by the nonlocal, or long range, contacts.
- *Proteinlike SIG recognition.* A general solution for the problem of recognizing a sphere intersection graph given proteinlike constraints remains an open problem.
- *Evaluation.* As in all areas of structure prediction, methods for evaluating success in contact map prediction need to correlate with usefulness, otherwise interative training of any sort will not converge on the truth.
- *Reconstruction of HS and SC contact maps.* Contact maps based on side chains work the best for fold recognition, but projecting maps into 3D is problematic.

If a fool-proof method can be established for converting side-chain contacts to a 3D ensemble, it will eventually unleash the power of machine learning methods to attack the protein folding problem.

Recommended Reading

Baker, D. 2000. A surprising simplicity to protein folding. *Nature* 405:39–42.

Vendruscolo, M., Najmanovich, R., and Domany, E. 1999. Protein folding in contact map space. *Phys. Rev. Lett.* 82:656–659.

References

Aloy, P., Stark, A., Hadley, C., and Russell, R.B. 2003. Predictions without templates: New folds, secondary structure, and contacts in CASP5. *Proteins* 53 (Suppl. 6):436–456.

Altschuh, D., Lesk, A.M., Bloomer, A.C., and Klug, A. 1987. Correlation of coordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J. Mol. Biol.* 193:693–707.

Aszodi, A., Gradwell, M.J., and Taylor, W.R. 1995. Global fold determination from a small number of distance restraints. *J. Mol. Biol.* 251:308–326.

Berrera, M., Molinari, H., and Fogolari, F. 2003. Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. *BMC Bioinformatics* 4:8.

Bystroff, C., and Shao, Y. 2003. Modeling protein folding pathways. In *Practical Bioinformatics* (J.M. Bujnicki, Ed.). Berlin, Springer-Verlag.

Bystroff, C., Thorsson, V., and Baker, D. 2000. HMMSTR: A hidden Markov model for local sequence–structure correlations in proteins. *J. Mol. Biol.* 301:173–190.

Chavez, L.L., Onuchic, J.N., and Clementi, C. 2004. Quantifying the roughness on the free energy landscape: Entropic bottlenecks and protein folding rates. *J. Am. Chem. Soc.* 126:8426–8432.

Cheng, J., Randall, A., Sweredoski, M., and Baldi, P. 2005. SCRATCH: A protein structure and structural feature prediction server. *Nucleic Acids Res.* 33: 72–76.

Dodge, C., Schneider, R., and Sander, C. 1998. The HSSP database of protein structure–sequence alignments and family profiles. *Nucleic Acids Res.* 26:313–315.

Dosztanyi, Z., Fiser, A., and Simon, I. 1997. Stabilization centers in proteins: Identification, characterization and predictions. *J. Mol. Biol.* 272:597–612.

Eisenhawer, M., Cattarinussi, S., Kuhn, A., and Vogel, H. 2001. Fluorescence resonance energy transfer shows a close helix–helix distance in the transmembrane M13 procoat protein. *Biochemistry* 40:12321–12328.

Enosh, A., Fleishman, S.J., Ben-Tal, N., and Halperin, D. 2004. Assigning transmembrane segments to helices in intermediate-resolution structures. *Bioinformatics* 20 (Suppl. 1):I122–I129.

Fariselli, P., and Casadio, R. 1999. A neural network based predictor of residue contacts in proteins. *Protein Eng.* 12:15–21.

Fariselli, P., Olmea, O., Valencia, A., and Casadio, R. 2001a. Prediction of contact maps with neural networks and correlated mutations. *Protein Eng.* 14:835–843.

Fariselli, P., Olmea, O., Valencia, A., and Casadio, R. 2001b. Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins* Suppl. 5:157–62.

Göbel, U., Sander, C., Schneider, R., and Valencia, A. 1994. Correlated mutations and residue contacts in proteins. *Proteins* 18:309–317.

Graña, O., Baker, D., Maccallum, R.M., Meiler, J., Punta, M., Rost, B., Tress, M.L., and Valencia, A. 2005. CASP6 assessment of contact prediction. *Proteins* [Epub 26 Sep 2005].

Hamilton, N., Burrage, K., Ragan, M.A., and Huber, T. 2004. Protein contact prediction using patterns of correlation. *Proteins* 56:679–684.

Havel, T.F., Crippen, G.M., and Kuntz, I.D. 1979. Effects of distance constraints on macromolecular conformation. II. Simulation of experimental results and theoretical predictions. *Biopolymers* 18:73–81.

Hu, J., Shen, X., Shao, Y., Bystroff, C., and Zaki, M.J. 2002. Mining protein contact maps. *BIOKDD 2002*, Edmonton, Canada.

Huang, E.S., Subbiah, S., and Levitt, M. 1995. Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.* 252:709–720.

Jones, D.T. 1999. Protein secondary structure prediction based on position-specific scoring matrices, *J. Mol. Biol.* 292:195–202.

Kleinjung, J., Romein, J., Lin, K., and Heringa, J. 2004. Contact-based sequence alignment. *Nucleic Acids Res.* 32:2464–2473.

Koh, I.Y., Eyrich, V.A., Marti-Renom, M.A., Przybylski, D., Madhusudhan, M.S., Eswar, N., Grana, O., Pazos, F., Valencia, A., Sali, A., and Rost, B. 2003. EVA: Evaluation of protein structure prediction servers. *Nucleic Acids Res.* 31:3311–3315.

Kraulis, P.J. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. App. Crystallogr.* 24:946–950.

Kuznetsov, I.B., and Rackovsky, S. 2004. Class-specific correlations between protein folding rate, structure-derived, and sequence-derived descriptors. *Proteins* 54:333–334.

Lichtarge, O., Bourne, H.R., and Cohen, F.E. 1996. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* 257:342–358.

Lin, K., Kleinjung, J., Taylor, W., and Heringa, J. 2003. Testing homology with CAO: A contact-based Markov model of protein evolution. *Comp. Biol. Chem.* 27:93–102.

Lund, O., Frimand, K., Gorodkin, J., Bohr, H., Bohr, J., Hansen, J., and Brunak, S. 1997. Protein distance constraints predicted by neural networks and probability density functions. *Protein Eng.* 10:1241–1248.

MacCallum, R.M. 2004. Striped sheets and protein contact prediction. *Bioinformatics* 20(Suppl. 1):I224–I231.

Maiorov, V.N., and Crippen, G.M. 1992. Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* 227:876–888.

McGuffin, L.J., Bryson, K., and Jones, D.T. 2000. The PSIPRED protein structure prediction server. *Bioinformatics* 16:404–405.

McLachlan, A.D. 1971. Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c 551. *J. Mol. Biol.* 61:409–424.

Michael, T.S., and Quint, T. 1999. Sphere of influence graphs in general metric spaces. *Math. Comput. Model.* 29:45–53.

Michalopoulos, I., Torrance, G.M., Gilbert, D.R., and Westhead, D.R. 2004. TOPS: An enhanced database of protein structural topology. *Nucleic Acids Res.* 32:D251–D254.

Mirny, L., and Domany, E. 1996. Protein fold recognition and dynamics in the space of contact maps. *Proteins* 26:391–410.

Monge, A., Friesner, R.A., and Honig, B. 1994. An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure. *Proc. Natl. Acad. Sci. USA* 91:5027–5029.

Moult, J., Fidelis, K., Zemla, A., and Hubbard, T. 2003. Critical assessment of methods of protein structure prediction (CASP)–round V. *Proteins* 53 (Suppl. 6):334–339.

Neher, E. 1994. How frequent are correlated changes in families of protein sequences? *Proc. Natl. Acad. Sci. USA* 91:98–102.

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

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. 1997. CATH—A hierarchic classification of protein domain structures. *Structure* 5:1093–1108.

Park, K., Vendruscolo, M., and Domany, E. 2000. Toward an energy function for the contact map representation of proteins. *Proteins* 40:237–248.

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

Plaxco, K.W., Simons, K.T., and Baker, D. 1998. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* 277:985–994.

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

Porto, M., Bastolla, U., Roman, H.E., and Vendruscolo, M. 2004. Reconstruction of protein structures from a vectorial representation. *Phys. Rev. Lett.* 92:218101–218104.

Punta, M., and Rost, B. 2005. Protein folding rates estimated from contact predictions. *J. Mol. Biol.* 348:507–512.

Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77:257–286.

Rodionov, M.A., and Johnson, M.S. 1994. Residue–residue contact substitution probabilities derived from aligned three-dimensional structures and the identification of common folds. *Protein Sci.* 3:2366–2377.

Saitoh, S., Nakai, T., and Nishikawa, K. 1993. A geometrical constraint approach for reproducing the native backbone conformation of a protein. *Proteins* 15:191–204.

Shao, Y., and Bystroff, C. 2003. Predicting interresidue contacts using templates and pathways. *Proteins* 53(Suppl. 6):497–502.

Shindyalov, I.N., Kolchanov, N.A., and Sander, C. 1994. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng.* 7:349–358.

Singer, M.S., Vriend, G., and Bywater, R.P. 2002. Prediction of protein residue contacts with a PDB-derived likelihood matrix. *Protein Eng.* 15:721–725.

Sippl, M.J. 1990. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* 213:859–883.

Skolnick, J., Kolinski, A., and Ortiz, A.R. 1997. MONSSTER: A method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.* 265:217–241.

Tanaka, S., and Scheraga, H.A. 1976. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules* 9:945–950.

Taylor, W.R., and Hatrick, K. 1994. Compensating changes in protein multiple sequence alignments. *Protein Eng.* 7:341–348.

Thomas, D.J., Casari, G., and Sander, C. 1996. The prediction of protein contacts from multiple sequence alignments. *Protein Eng.* 9:941–948.

Vendruscolo, M., and Domany, E. 1998. Efficient dynamics in the space of contact maps. *Fold Des.* 3:329–336.

Vendruscolo, M., Kussell, E., and Domany, E. 1997. Recovery of protein structure from contact maps. *Fold Des.* 2:295–306.

Wako, H., and Scheraga, H.A. 1982. Visualization of the nature of protein folding by a study of a distance constraint approach in two-dimensional models. *Biopolymers* 21:611–632.

Yuan, X., and Bystroff, C. 2005. Non-sequential structure-based alignments reveal topology-independent core packing arrangements in proteins. *Bioinformatics* 27:1010–1019.

Zaki, M.J., Shan, J., and Bystroff, C. 2000. Mining residue contacts in proteins using local structure predictions. *Proceedings IEEE International Symposium on Bio-Informatics and Biomedical Engineering*, Arlington, VA.

Zhang, C., and Kim, S.H. 2000. Environment-dependent residue contact energies for proteins. *Proc. Natl. Acad. Sci. USA* 97:2550–2555.

Zhao, Y., and Karypis, G. 2003. Prediction of contact maps using support vector machines. *BIBE 2003*, Bethesda, MD. IEEE Computer Society, pp. 26–36.