8

The Pros and Cons of Predicting Protein Contact Maps

Lisa Bartoli, Emidio Capriotti, Piero Fariselli, Pier Luigi Martelli, and Rita Casadio

Summary

Is there any reason why we should predict contact maps (CMs)? The question is one of the several 'NP-hard' questions that arise when striving for feasible solutions of the protein folding problem. At some point, theoreticians started thinking that a possible alternative to an unsolvable problem was to predict a simplified version of the protein structure: a CM. In this chapter, we will clarify that whenever problems are difficult they remain at least as difficult in the process of finding approximate solutions or heuristic approaches. However, humans rarely give up, as it is stimulating to find solutions in the face of difficulties. CMs of proteins are an interesting and useful representation of protein structures. These two-dimensional representations capture all the important features of a protein fold. We will review the general characteristics of CMs and the methods developed to study and predict them, and we will highlight some new ideas on how to improve CM predictions.

Key Words: Protein structure prediction; Protein contacts; Small world; Structure reconstruction; Machine learning; Contact map; Protein folding.

1. From Protein Structures to Contact Maps

Proteins structures are described by the coordinates (CO-representation) of the atoms that constitute the macromolecule. For a protein with n atoms we need $3n$ numbers (x, y and z coordinates for each atom) to specify its three-dimensional (3D) structure. An alternative view is to consider the distance matrix (DM), a symmetric matrix that contains the Euclidean distance between each pair of atoms. If the number of atoms is n we need n^2 elements; because

From: *Methods in Molecular Biology, vol. 413: Protein Structure Prediction, Second Edition*
Edited by: M. Zaki and C. Bystroff © Humana Press Inc., Totowa, NJ

the matrix is symmetric (the distance between atoms i and j is the same of that between j and i), the real number of elements is only $n(n-1)/2$. Both representations, namely the coordinates and the DM, are equivalent, that is, we can convert each representation into the other. DM can be computed from the CO-representation simply by evaluating the Euclidean distance between each pair of atoms: values stored in the appropriate DM cell uniquely identify the pair i and j. Conversely, to go from DM to CO is not so trivial. There exists a Lagrange theorem ***(1)*** that states that once that the Gram matrix derived from DM is diagonalized, the three eigenvectors that correspond to the three highest eigenvalues are the atom coordinates in a 3D cartesian reference. Actually, there are two solutions, but the chirality of the molecule routinely can help in selecting the correct one (***1*** and references therein).

DM representation has far more elements than the coordinate-based representation, so why adopt it? The main advantage of DM representation arises when only a part of the data is known (i.e., in low-resolution NMR experiments). Still solutions can be found, thanks to DM properties ***(1)***. Another advantage of DM is that the protein is represented in a framework that automatically incorporates translational and rotational invariance and this in principle is more suitable for learning approaches.

Quite often in order to simplify the protein representation not all protein atoms are taken into account and residues are considered as unique entities. In this case, the DM has a number of rows (and columns) equal to the residue numbers. Each DM entry is then the distance between residue i and j. The distance between two residues can be defined in different ways, such as the following:

- the distance between a specific pair of atoms (i.e., CA–CA or CB–CB),
- the shortest distance among the atoms belonging to i residue and those belonging to residue j, and
- the distance between the centres of mass of the two residues.

Even though these choices are quite different and structurally minimal, they provide enough information to build the protein backbone, or at least the CA trace ***(1,2)***.

Starting from the protein DM and selecting an arbitrary distance cut-off, a further simplified representation can be obtained: the protein contact map (CM). CMs are binary symmetric matrices, whose non-zero elements represent

the contacts between residues (*see* **Fig. 1**). In more details, given a DM and a defined threshold T the corresponding CM can be computed as:

$$\mathrm{CM}[i, j] = 1 \text{ if DM } [i, j] < T$$
$$\mathrm{CM}[i, j] = 0 \text{ if DM } [i, j] \geq T$$

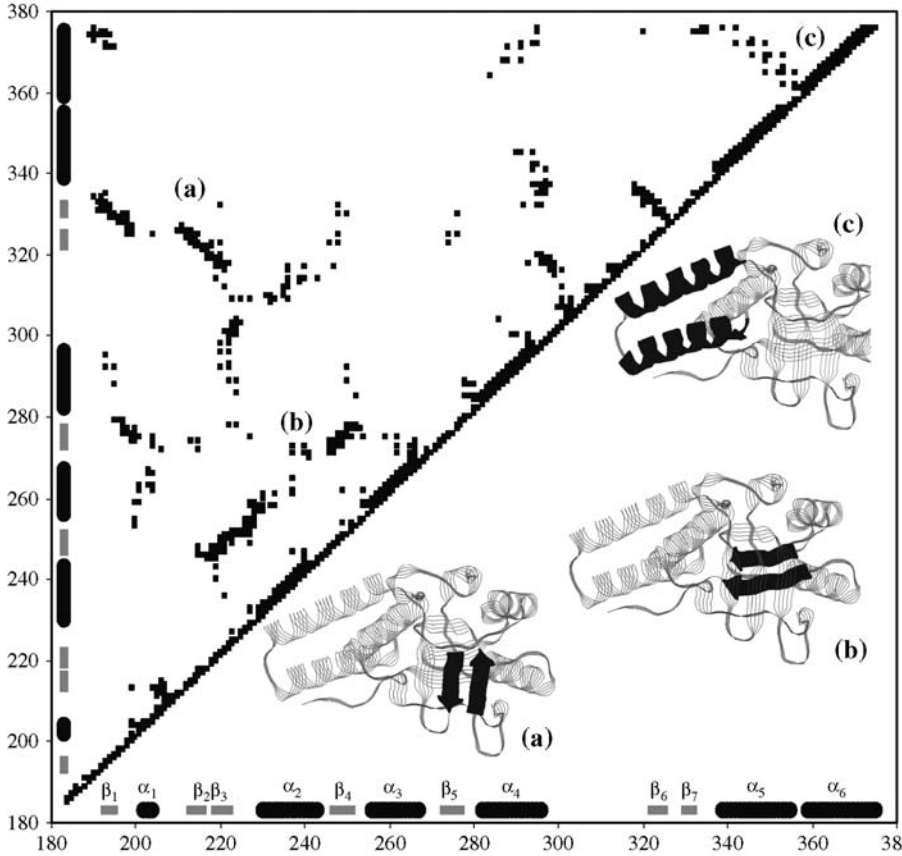

Fig. 1. Contact map of HSP-60 protein fragment (PDB code: 1KID). The secondary structure elements are highlighted along the x axis and y axis. α-helices and β-strands are represented in black and grey, respectively. On the left side of the plot the black dots indicate the contact regions (cut-off radius 8.0 Å centered at CB atoms). On the right side, the structural protein features are shown: (a) Anti-parallel sheet contacts; (b) parallel sheet contacts; (c) contacts between helical regions.

While the problem of reconstructing the protein coordinates from the DM has a well known solution, there are no analogous theorems for CM. However, some empirical applications have been built to address this issue. The results indicate that (at least for the tested proteins) it is possible to reconstruct the CO-representations from CMs ***(2–5)***.

Protein CM representation has some pros and cons.

Pros:

- Unlike other protein representations such as secondary structure, CM conveys strong information about the protein 3D structure.
- The CM representation is translation and rotation invariant and more compact than the DM representation.
- CM is more suited than DM for learning problems. The binary CM nature can be regarded as a classical problem of a two-state classification and this has been thoroughly studied. There are several machine learning methods available to address the problem of the prediction of CM from the protein residue sequence ***(6)***.
- It has been shown that the empirical reconstruction algorithms are quite insensitive to high levels of random noise in CMs, so that for reconstructing the 3D structure of the protein it is not necessary to correctly predict all contacts ***(2,4)***.

Cons:

- There is no theory on CM that can help to define the limits and the strength of this representation. For instance, the effect of the contact threshold on the information content is not theoretically assessable. For this reason, different researchers adopt different protein representations and contact thresholds.
- The problem of CM comparison is very hard, as it is that of a sub-graph isomorphism, which is NP-hard ***(7)***.
- CMs of real proteins are a tiny subset of the possible binary symmetric matrices ***(2)***; however, no simple and fast algorithm has been found to sort out the protein-feasible CM from the others.
- CM prediction is an intrinsically non-local problem. Also, this is a very difficult problem to deal with, as a contact between two residues poses constraints on the feasibilities of all other contacts.
- Although the reconstruction programs are very insensitive to random noise, they are not as robust when the prediction errors are correlated, as is the case with current prediction algorithms.

CMs can be regarded both as symmetric matrices and as graphs. Actually, the CM representation is an adjacency matrix, where the contacts are the edges and the residues are the nodes. It is useful to distinguish between short-range and long-range contacts. The distinction between short-range

(sometimes called 'local') and long-range ('non-local') contacts is not due to the type of interaction, nor the spatial distance, but it is due to the relative sequence separation. Contacts between residues that are separated less than a given number of residues S ($|i-j| \leq S$) are said to be short-range. Conversely, if the sequence separation is greater than S, they are said to be long-range. The choice of S is arbitrary, but it is commonly accepted that $|i-j| \leq 7-10$ represents short-range contacts, while $|i-j| > 7-10$ represents long-range ones.

2. Properties of Protein Contact Maps

When CMs are analyzed, one of the first features is that the number of contacts increases almost linearly with the protein length, independently of the adopted distance measures (CA–CA, CB–CB, etc.) and of the threshold cut-off used ***(8)***. More formally, if L is the protein length and nc is the number of contacts, the real number of contacts can be quite accurately estimated using the linear equation

$$nc = A_T \times L$$

where A_T is a constant that only depends on the contact threshold (T). In practice, a change in the contact threshold T (in a reasonable range) has the only effect of modifying the slope of the line. This finding, together with the fact that the number of possible contacts ***NCM***, which is the number of independent CM elements ($NCM = L(L-1)/2$), increases with the square of the protein length, implies that the contact densities in the map (nc/NCM) decrease as the inverse of the protein length. In other words, long proteins have a lower contact density than short ones ***(8)***.

Protein CMs have also more contacts in the short-sequence separations than those obtained using random graphs with the same number of contacts ***(8)***. This is an indication that protein structures have a high tendency to form contacts with sequence neighbours.

Studying the properties of the CM eigenvectors, it has been found that there is a high correlation between the eigenvector corresponding to the highest eigenvalue (first eigenvector) and the residue coordination numbers ***(5,9)***. The residue coordination number (or contact vector) is the number of contacts of each given residue with all the others in the protein space ***(10)***. This figure can be easily computed from the contact matrix by summing up the rows (or the columns) of CM.

Galaktionov and Marshall ***(5)*** reported that from the knowledge of the real residue coordination numbers, it is possible to reconstruct to some extent (about 4 Å of Root Mean Square Deviation (RMSD)) the 3D structure of the protein.

A further surprising property of the first eigenvector of CM is the fact that a CM can be reconstructed using only the information contained in this vector coupled along with the information derived from the protein backbone constraints ***(9)***. However, this is not a general property of all binary symmetric matrices, only of the subset comprising single-domain proteins ***(9)***.

3. Reconstructing Protein Structures From Contact Maps

As outlined above, a CM contains a simplified representation of the protein conformation and it is unambiguously computed from the structure by a binary simplification of the DM. It is well known that a protein structure can be reconstructed from its DM by means of the Lagrange theorem ***(1)***. This procedure is unambiguous, except for the ambiguity due to chiral symmetry. The questions are these: is it possible to recover the structure starting from its real CM as well? And from a predicted CM?

Bohr et al. ***(3)*** implemented a method based on the definition of a continuous function that measures the distance of a protein structure from a given CM. By adding some terms for assuring the connectivity and the compactness of the protein structure, a target function was obtained and then minimized using a simple steepest descent algorithm. The optimal computed structure satisfies as many contacts as possible.

At an 8 Å threshold for the distance between two CA atoms, the algorithm recovers the structure starting from the real CM with a RMSD less than 3 Å. It is worth noticing that the threshold value for the contact definition can be chosen within a wide range without greatly affecting the deviation of the recovered structure with respect to the real one. The optimal threshold for the minimization depends on the protein size.

The algorithm is efficient when a real CM is adopted; however, it fails when predicted CMs are considered for defining the target function. When the rate of error on the predicted map is only about 5%, it leads to structures with a RMSD $\geq$ 5 Å. This is due not only to the low quality of the prediction but also to the fact that a physical CM needs to satisfy complex constraints in order to represent a real structure.

When predicting contacts between each pair of residues in a sequence, the computation is independent of the other assigned contacts and then the resulting map is likely to be non-physical. In these cases, the recovering algorithm has to deal with the noise introduced by the inconsistency of the predicted

contacts. This issue was thoroughly discussed by Vendruscolo and Domany ***(2)*** who implemented a stochastic algorithm for building a structure satisfying the protein CM. The algorithm builds a structure adding residues one at a time, trying different random conformations and then randomly adapting the preceding portion of the chain. In each step, the number of fulfilled contact constraints is the objective function for selecting the best conformations. By this, starting from the real map with a threshold distance value equal to 9 Å, the protein structure is reconstructed with a RMSD between 1 and 2 Å. The authors introduce noise in the physical map by flipping randomly chosen positions in the map and their algorithm results more robust than that of Bohr et al. ***(3)***. Indeed even when about 20% of the map is randomly inverted, the algorithm reconstructs structures with a 4 Å RMSD to the real protein. However, this kind of non-physical CMs are likely to contain much more information than the predicted ones, as the randomness of the flipping conserves most of the original protein structure representation. Unfortunately, in a predicted map, errors are often more correlated and then recovering of the 3D structure is far more difficult.

In short, the implemented algorithms to reconstruct protein structure starting from CM prove that for a wide range of distance cut-offs, the CM is a good representation of the protein backbone conformation. It is possible to reconstruct the structure in the best cases with a deviation of less than 3 Å. Nevertheless, it should be considered that presently it is still impossible to deal with predicted maps, as in this case the level of noise is too high.

4. The Prediction of Protein Contact Maps

In these years, several researchers have been predicting CMs starting from protein sequence information. This interest grew after it was shown that it is possible to reconstruct protein structures from their CMs (*see* Section 3). Among the first attempts to predict residue contacts in proteins, there are methods based on correlated mutations ***(11,12)***. In this case, the basic idea is that the maintenance of protein functions constrains the evolution of residue sequences. This fact can be exploited to interpret correlated mutations, observed in a sequence family, as an indication of a probable physical contact in 3D. On this basis, if a given residue mutates in a position, it is likely that a residue in contact with it will mutate too, in order to compensate the previous change. Also, strong hydrophobic conserved residues have a high probability of being in contact ***(11)***.

An alternative approach is to learn the correlation between sequence and CM using machine learning tools. In this respect, several methods have been

introduced: neural networks that exploit multiple sequence alignments ***(4,8, 13,14)***, hidden Markov models ***(15)***, support vector machines ***(16)***, genetic programming ***(17)*** and recurrent neural networks ***(18)***. Neural network-based methods incorporate several sequence features related to the local environment of two residues for their prediction of being or not in contact, including in some cases correlated mutations and residue conservation ***(4,8)***. More recently, Punta and Rost have improved the neural network prediction accuracy by adding information relative to the segment that connects the two residues undergoing prediction. This is done by coding also the sequence environment of the residue that falls exactly in the middle between the two residues considered. More precisely, if the contact propensity for the pair i, j $(j > i)$ is predicted, they also code the environment for position $k = (j+i)/2$. This information seems to improve the neural network prediction accuracy up to 32% when sequence separation is six residues long ***(14)***, and this is the highest score reported so far. Similar to other predictors, this accuracy is obtained using a number of predicted contacts equal to half of the protein length ***(14)***.

Another method codes the protein underlying grammar for hidden Markov models to find residue contact patterns among different pairs of segments by adopting an approach that can be regarded as an extension of threading methods ***(15)***.

Recently, machine learning methods have tried to incorporate information relative to the geometric properties of CMs. It seems that the introduction of the information relative to the prediction of the first eigenvector density components helps the prediction of the final CM ***(19,20)***.

During the last Critical Assessment of Technique for Protein Structure Prediction (CASP6), some methods and servers were mainly evaluated on long-range contact predictions for a set of about 10 proteins belonging to the new fold targets ***(21)***. The assessors found that three approaches, including PROFcon ***(14)***, with similar levels of accuracy and coverage performed a little better than others ***(14,17,21)***. Comparisons of the predictions of the three best methods with those of CASP5/CAFASP3 suggested some improvement, although there were not enough targets in the comparison set to make this statistically significant. Irrespective of the CM prediction accuracy, they are still better than constraints from the best de novo 3D prediction methods ***(20)***.

How a predicted CM looks like? As an example, in **Fig. 2**, we show the prediction of an all-alpha protein. For this specific protein, accuracy is 44%, a quite satisfactory value when it is considered that this protein structural type is the most difficult to be predicted. Prediction in this case was computed with an updated version of our CORNET method ***(8)***.

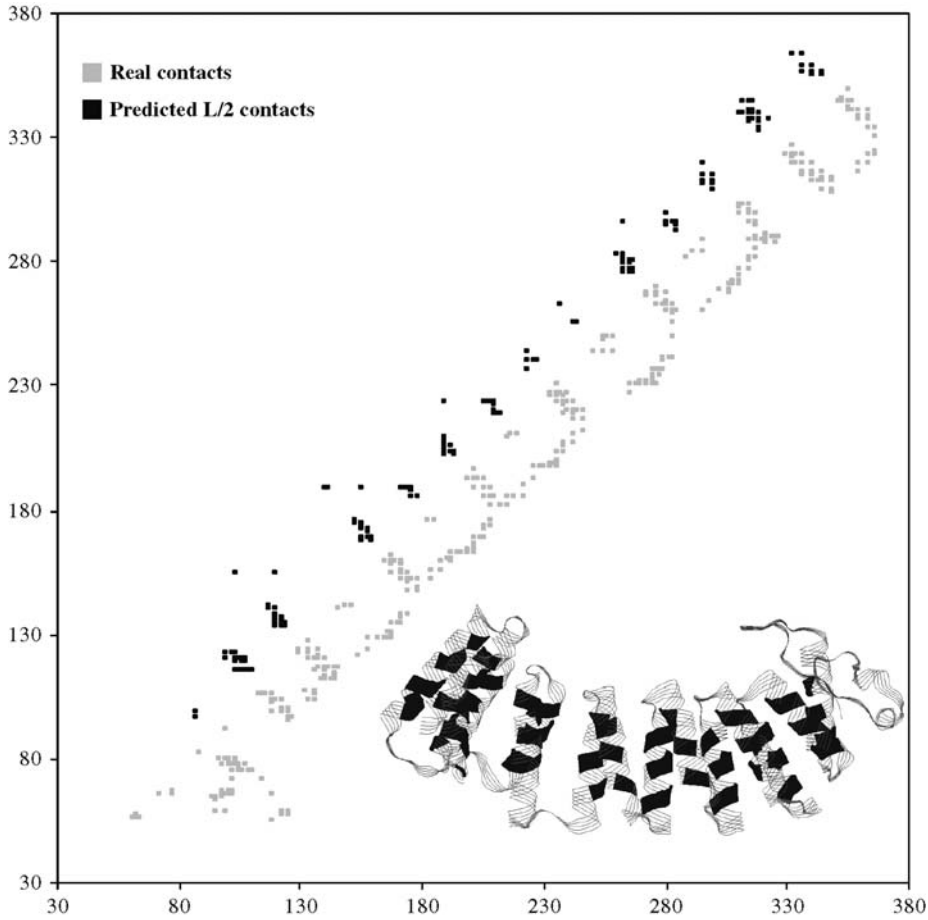

Fig. 2. Real versus predicted contact map of the α-subunit of the human Farnesyltransferase (PDB code: 1LD8 chain A). On the left side of the plot the black dots indicate the predicted L/2 contact residues. On the right side, in grey, the real residue contacts are shown (cut-off radius is 8.0 Å centered at the CB atoms and sequence separation ≥ 6). In the corner on the right, the protein structure is shown, highlighted in black, the correctly predicted contacts. On this protein, our neural network-based predictor reaches an accuracy equal to 44%.

5. Small World and Contact Maps

Unfortunately we use CMs, we predict them, but we are still unhappy. How do we improve our methods and our prediction? The solution is still to be found. In the meantime, we suggest another perspective in the following sections.

5.1. Small World

To overview some recent literature on proteins, we should introduce a few concepts explaining what 'small world' is and how it has been used to highlight protein folding properties.

In the mid-1990s, Duncan Watts, while studying for his PhD in Applied Mathematics, was invited to study a very particular problem: how crickets synchronize their singing ***(22)***. He was convinced that, to deeply understand this problem, he had to observe the way the crickets pay attention to each other. This is the starting point of the study of networks under a different perspective than that of random networks that were previously introduced by Erdós and Rényi (***(22)*** and references therein). Watts started his study on social networks trying to answer to a simple question: how many probabilities are there that two persons, both my friends, know each other? With his Professor Steven Strogatz, he found that social networks were clustered and not randomly distributed and that the same paradigm could model dynamical relations in many different systems ***(22)***.

To explain the omnipresence of clustering in real world networks, Watts and Strogatz ***(23)*** proposed a new connection topology called a 'small world' network, showing that it can be interpolated between regular and random networks with a random rewiring procedure. According to this model, small world systems can be highly clustered, like regular graphs, and at the same time they are endowed with a small average path length, as it is for random networks.

Watts and Strogatz ***(23)*** introduced two numbers to describe the characteristics of small world networks: the characteristic path length L and the clustering coefficient C. L is given by the number of edges in the shortest path between two vertices, averaged over all pairs of vertices:

$$L = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} L_{ij},$$

where L_{ij} is the shortest path length between vertices i and j.

Supposing that a vertex k has N_k neighbours, then at the most $N_k(N_k-1)/2$ edges can exist between them. If n_k is the actual number of edges among the neighbours, then C is defined as:

$$C = \frac{1}{N} \sum_{k=1}^{N} \frac{n_k}{N_k(N_k-1)/2}.$$

L measures the typical separation between two vertices in the graph (a global property) and C is a measure of local clustering or cliquishness of a typical neighbourhood (a local property) ***(23)***.

5.2. Small World and Protein Structures

The extension of the small world view to proteins was straightforward. Vendruscolo et al. ***(24)*** showed that protein structures have small world topology. The small world behaviour of protein structures is reflected by the presence in their graph of a relatively small number of vertices with many connections ***(24)***. Two residues are considered as connected if the distance between their CA atoms is less than a threshold distance fixed at 8.5 Å. By analysing a data set of 978 representative proteins, it was found that the average value of L is 4.1 ± 0.9 and that of C is 0.58 ± 0.04. These values were compared with those obtained for random and regular graphs. By assuming that K is the average number of links in the graph (the average number of contacts in a protein) and N is the number of vertices (protein residues), then $L_{\text{random}} \sim \ln N / \ln K$ and $C_{\text{random}} \sim K/N$; $L_{\text{regular}} \sim N(N+K-2)/2K(N-1)$ and $C_{\text{regular}} \sim 3(K-2)/4(K-1)$ ***(25)***. Values of 2.4 ± 0.3 and 0.08 ± 0.06 were reported for L_{random} and C_{random} respectively; L_{regular} and C_{regular} were 10.4 ± 7.0 and 0.67 ± 0.04, respectively ***(24)***.

In this chapter for sake of clarity and with the specific aim of relating the small world representation to CMs (see below), we perform the same type of analysis on a new and a more selected data set of non-redundant mono-domain proteins (497 proteins) (*see* **Fig. 3**). We reached similar conclusions as before ***(24)***, obtaining L and C equal to 3.9 ± 0.9 and 0.57 ± 0.03, respectively. For our data set, L_{random} is 2.1 ± 0.2, C_{random} is 0.08 ± 0.04, L_{regular} is 8.7 ± 4.2 and C_{regular} is 0.67 ± 0.01, confirming again that $L_{\text{random}} < L < L_{\text{regular}}$ and that $C_{\text{random}} < C < C_{\text{regular}}$, a key conclusion for resorting small world behaviour.

Small world view was adopted also for homopolymers obtained with a CM dynamics ***(26)*** and for atomic clusters obtained with Lennard–Jones interactions with a Monte Carlo method ***(27)***. In both cases, the values of C and L were found similar to those of proteins, indicating a small world topology also for these systems. It was therefore concluded that protein chain connectivity plays a minor role in the small world behaviour and that for a globular protein the small world character would mainly arise from the overall geometry (surface to volume ratio) ***(24)***.

What we did in house was substantially to add to these concepts by analysing other properties of our non-redundant protein set that have been related to small world behaviour. Another tendency that shows this property is that L increases linearly with log N (as a measure of the protein length) and that the slope is higher than the random reference case (*see* **Fig. 4**). This type of plot is frequent in the pertinent literature ***(28,29)***. In our case, we add to the conclusion by analysing a non-redundant set of mono-domain proteins.

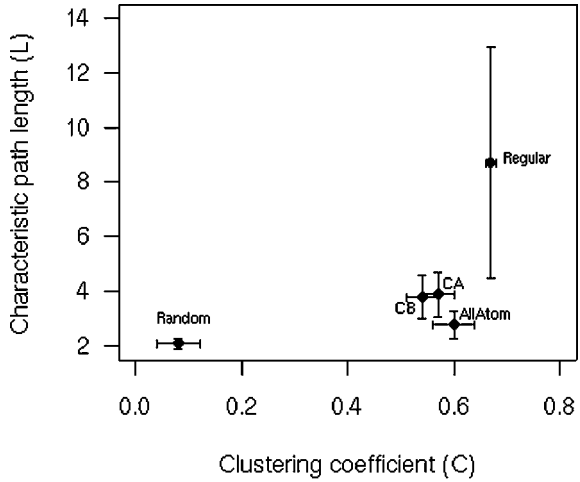

Fig. 3. Plot of the average path length versus the clustering coefficient computed on our data set of non-redundant set of mono-domain proteins (comprising 497 protein chains with sequence identity < 25%). Average values are reported with the associated standard deviation. Proteins are represented by CA, CB and all-atom (cut-off radius is 8.5 Å). Random: corresponding random graphs; Regular: corresponding regular graphs. See text for details.

As observed in the work of Atilgan et al. ***(28)***, the average value of *C* remains nearly constant with increasing protein size. We found the same trend on our data set (*see* **Fig. 5**). It should be however noticed that for each protein the tendency is that *C* decreases at increasing protein size. This fact is viewed as indicative of the modular nature of the small world networks. When globular and fibrous proteins are compared, no relevant difference arises, and a general belief is that 'small worldness' persists irrespectively of structural differences ***(28–30)***.

Atilgan et al. ***(28)*** studied 595 proteins with sequence homology < 25%, a set described before ***(13)***. The protein core local organization (residues residing at depths greater than 4 Å) is the same even if the size of the protein is different. Beyond a depth of approximately 4 Å from the protein surface, the clustering coefficient approaches a fixed value of approximately 0.35, irrespective of the size of the protein at hand. The same small world organization seems therefore to live throughout the protein, despite the heterogeneous density distribution that it may be found in different folds pertaining to different proteins.

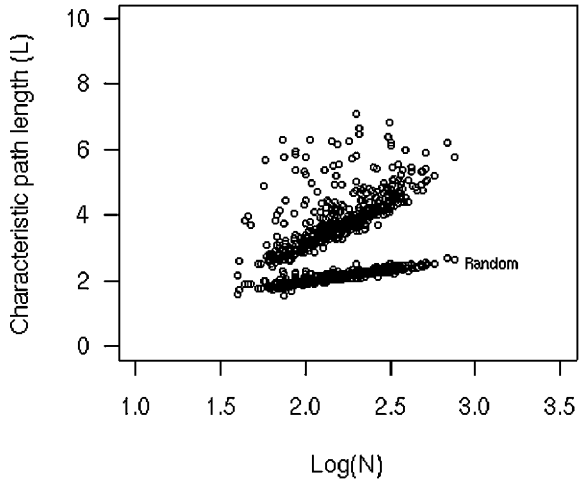

Fig. 4. Characteristic path length as a function of logarithm of the protein length [Log (N)]. L is shown for each protein of our data set. Real protein values cluster above those of corresponding random networks.

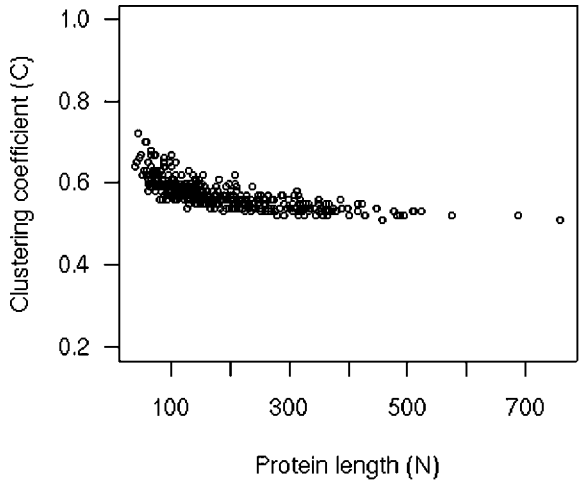

Fig. 5. Clustering coefficients of the different proteins as a function of the protein length (see text for details).

5.3. Local Versus Global Contacts

Greene and Higman ***(30)*** adopted an all-atom representation of the proteins instead of the less informative CA simplified representation. A contact was allowed between two residues when at least one pair of their atoms is within 5 Å from each other. By this, multiple links between residues are allowed. The small world property was analysed on a set of 65 non-redundant proteins divided into nine highly populated fold types representing the four SCOP protein classes: all-α, all-β, α/β, α + β (http://scop.mrc-lmb.cam.ac.uk/scop/). Interestingly Greene and Higman ***(30)*** found a difference of the behaviour between what they called networks of short-range and long-range contacts. Interactions are considered short range or long range if they occur between residues that are ≤ 10 and more than 10 residues apart in the protein sequence, respectively. A long-range interaction graph does not differ from a random graph; however, when also short-range contacts are taken into consideration the small world behaviour emerges. By following the short-range and long-range contact distinction, we compute C and L values for our protein set. The results are shown in **Fig. 6**, confirming that long-range contacts

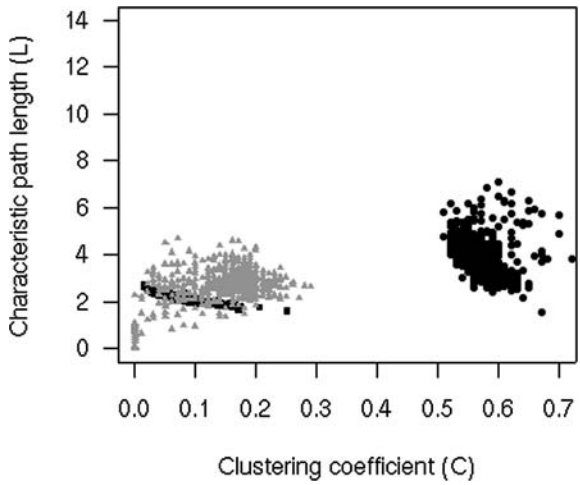

Fig. 6. The characteristic path length versus the clustering coefficient for each protein in the data set considering long-range contacts and complete contact maps. Black circles: complete protein contact maps. Grey triangles: long-range contacts. Black squares: random networks. Apparently, long-range contacts overlap with corresponding random networks.

can be modelled by a random graph and that small world properties emerge only when the whole CM is considered.

5.4. All-α Versus All-β Contacts

Several authors inspected how small world behaviour is dependent on the protein structural type, routinely following the SCOP classification ***(28–30)***. A thorough investigation study reveals a marginal but consistent difference in the C index value of all-α and all-β proteins. We show our results in **Fig. 7**. When considering the average C values, we find that they are 0.597 for all-α and 0.551 for all-β proteins, respectively. These values confirm the difference previously reported ***(29)***. This difference may be due to the larger geometrical compactness of α-helices as compared to β-sheets. Our data set contains 113 all-α proteins and 110 all-β proteins.

5.5. Scale-Free Networks and Contact Maps

Scale-free networks are small world; however, small world networks are not necessarily scale-free ***(31)***. In the protein world, CMs are not scale-free networks. A scale-free connectivity follows a power law $p(k) \sim k^{-\gamma}$ (where k is the number of links of a node and p is the probability of a node to have

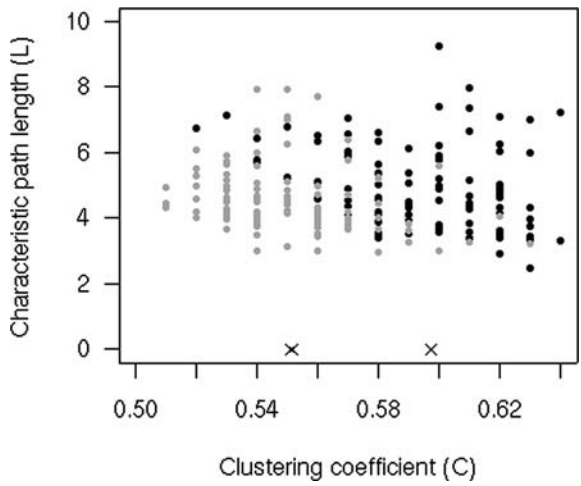

Fig. 7. The characteristic path length versus the clustering coefficient for 113 all-α (black dots) and 110 all-β proteins (grey dots). The two crosses indicate the average C values for the two groups: 0.597 and 0.551 for all-α and all-β proteins, respectively (see text for details).

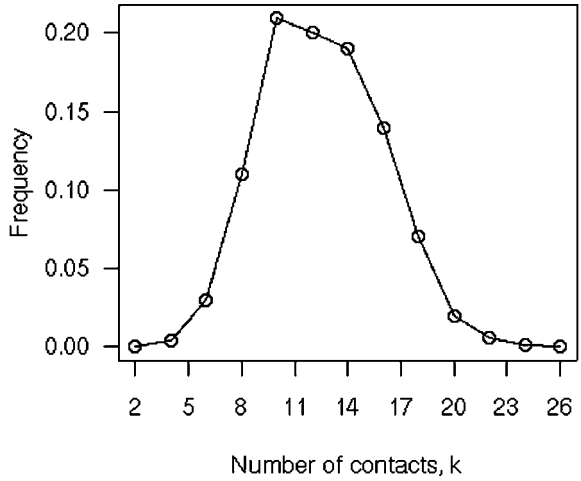

Fig. 8. Small world networks are not scale-free: frequency of residues (vertices) as a function of the number of contact per residue (k) in our protein data set.

k links). In a typical scale-free network $2 \leq \gamma \leq 3$. The distribution of both long-range and short-range contacts reveals a tendency to a bell-shaped Poisson curve which is typical of random networks and not of scale-free ones ***(30)***. The plot shown in **Fig. 8** is the result of a study on our data set of complete CMs, confirming the non-scale-free behaviour of contact distribution in our protein set.

6. Exploiting Small World Properties of Contact Maps

In Section 5, we showed that protein CMs are peculiar graphs that exhibit small world properties. The question arises whether predicted CMs behave similarly. Thus, we predicted some 100 mono-domain proteins using PROFcon ***(14)*** that has been demonstrated to be one of the best performing available methods ***(21)***. However, PROFcon assigns predictions only to pair of residues that are more than five residues apart, and therefore, in order to compare the predicted CMs with the observed ones, we also added the trivial connectivity to the predictions (which consists of the CM diagonals i, $i + 1$ and i, $i + 2$). The trivial contacts are due to the backbone connectivity when a CB threshold is set to 8 Å (as was in this case). The results are reported in **Fig. 9**, where it is evident that also the predicted CMs generate graphs with small world behavior. Nevertheless, the predicted CMs have lower values of both characteristic path length (L) and clustering coefficient (C) with respect to real proteins. Prediction

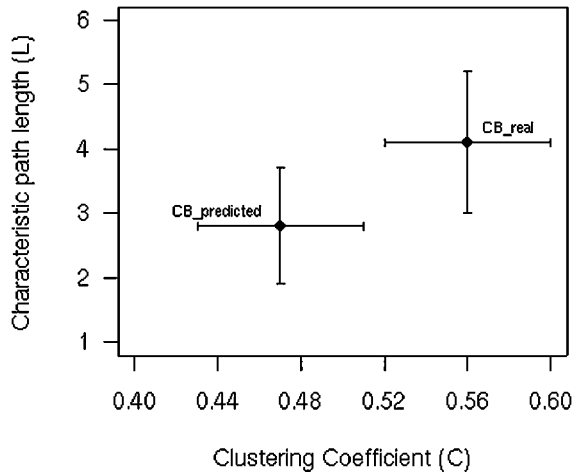

Fig. 9. Plot of the average characteristic path length versus the average clustering coefficient computed on 113 contact maps of all-α proteins predicted with the PROFcon prediction method ***(14)*** (CB predicted) compared to physical ones (CB real). Predicted contact maps are non-random but still different from real contact maps.

therefore generates CMs that are different from random but still far from the real proteins. Eventually, this perspective may help in filtering out spurious assignments.

7. Conclusions

Writing a review article is always an effort, especially when piled up results in a field are still promising results. In this chapter, we hope to have addressed the old and present problems in CM predictions and highlighted why we are still willing to devote our effort to this field. Also, we have suggested that possibly by merging small world view of proteins and CMs, new optimization algorithms may be developed to reduce signal-to-noise ratio. This will eventually help us also in finally reconstructing the 3D protein structure from predicted CMs.

Acknowledgments

We thank MIUR for the following grants: PNR-2003 grant delivered to PF, a PNR 2001–2003 (FIRB art.8) and PNR 2003 projects (FIRB art.8) on Bioinformatics for Genomics and Proteomics and LIBI-Laboratorio Internazionale di BioInformatica, both delivered to RC. This work was also supported by the

Biosapiens Network of Excellence project (a grant of the European Unions VI Framework Programme).

References

1. Havel, T. F. (1998). Distance geometry: theory, algorithms, and chemical applications. *Encyclopedia of Computational Chemistry*. John Wiley & Sons, New York.
2. Vendruscolo, M. and Domany, E. (1999). Protein folding using contact maps. *arXiv* cond-mat, 9901215.
3. Bohr, J., Bohr, H., Brunak, S., Cotterill, R. M., Fredholm, H., Lautrup, B. and Petersen, S. B. (1993). Protein structures from distance inequalities. *Journal of Molecular Biology* **231**, 861–869.
4. Fariselli, P., Olmea, O., Valencia, A. and Casadio, R. (2001). Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins* **Suppl 5**, 157–162.
5. Galaktionov, S. G. and Marshall, G. R. (1994). *27th Annual Hawaii International Conference on System Sciences (HICSS-27)*, Maui, Hawaii.
6. Baldi, P. and Brunak S. (2001). *Bioinformatics: The Machine Learning Approach, A Bradford Book,* Second edition. MIT Press, Cambridge.
7. Goldman, D., Istrail, S. and Papadimitriou, C. (1999). Algorithmic aspects of protein structure similarity. *Proceedings of the 40th IEEE Symposium on Foundations of Computer Science*, New York, (USA), 512–522.
8. Fariselli, P., Olmea, O., Valencia, A. and Casadio, R. (2001). Prediction of contact maps with neural networks and correlated mutations. *Protein Engineering* **14**, 835–843.
9. Porto, M., Bastolla, U., Roman, H. E. and Vendruscolo, M. (2004). Reconstruction of protein structures from a vectorial representation. *Physical Review Letters* **92**, 218101.
10. Pollastri, G., Baldi, P., Fariselli, P. and Casadio, R. (2002). Prediction of coordination number and relative solvent accessibility in proteins. *Proteins* **47(2)**, 142–153.
11. Goebel, U., Sander, C., Schneider, R. and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins* **18**, 309–317.
12. Olmea, O. and Valencia, A. (1997). Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Folding & Design* **2**, S25–S32.
13. Fariselli, P. and Casadio, R. (1999). A neural network based predictor of residue contacts in proteins. *Protein Engineering* **12**, 15–21.
14. Punta, M. and Rost, B. (2005). PROFcon: novel prediction of long-range contacts. *Bioinformatics* **21**, 2960–2968.

15. Bystroff, C. and Shao, Y. (2002). Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA. *Bioinformatics* **18 Suppl 1**, S54–S61.
16. Zhao, Y. and Karypis, G. (2003). *3rd IEEE International Conference on Bioinformatics and Bioengineering (BIBE).*
17. MacCallum, R. M. (2004). Striped sheets and protein contact prediction. *Bioinformatics* **20 Suppl 1**, I224–I231.
18. 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.
19. Vullo, A., Walsh, I. and Pollastri, G. (2006). A two-stage approach for improved prediction of residue contact maps. *BMC Bioinformatics* **7**, 180.
20. Eyrich, V. A., Przybylski, D., Koh, I. Y., Grana, O., Pazos, F., Valencia, A. and Rost, B. (2003). CAFASP3 in the spotlight of EVA. *Proteins* **53 Suppl 6**, 548–560.
21. Grana, 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* **61 Suppl 7**, 214–224.
22. Barabasi, A. L. (2003). *Linked: The New Science of Networks*, Perseus Publishing, Cambridge, Massachusetts.
23. Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442.
24. Vendruscolo, M., Dokholyan, N. V., Paci, E. and Karplus, M. (2002). Small-world view of the amino acids that play a key role in protein folding. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* **65**, 061910.
25. Watts, D. J. (1999). *Small Worlds. The Dynamics of Networks Between Order and Randomness*, Princeton University Press, Princeton, New Jersey.
26. Vendruscolo, M. and Domany, E. (1998). Efficient dynamics in the space of contact maps. *Folding & Design* **3**, 329–336.
27. Andricioaei, I., Voter, A. F. and Straub, J. E. (2001). Smart Darting Monte Carlo. *The Journal of Chemical Physics* **114**, 6994–7000.
28. Atilgan, A. R., Akan, P. and Baysal, C. (2004). Small-world communication of residues and significance for protein dynamics. *Biophysical Journal* **86**, 85–91.
29. Bagler, G. and Sinha, S. (2005). Network properties of protein structures. *Physica A* **346**, 27–33.
30. Greene, L. H. and Higman, V. A. (2003). Uncovering network systems within protein structures. *Journal of Molecular Biology* **334**, 781–791.
31. Barabasi, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* **286**, 509–512.