

Evolution and similarity evaluation of protein structures in contact map space

Nitin Gupta, Nitin Mangal and Somenath Biswas

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur, Kanpur-208016, India

nitiniitk@yahoo.com, mangal_iitk@yahoo.com, sb@cse.iitk.ac.in

Abstract

Prediction of fold from amino-acid sequence of a protein has been an active area of research in the past few years, but the limited accuracy of existing techniques emphasizes the need to develop newer approaches to tackle this task. In this study, we use contact map prediction as an intermediate step in fold prediction from sequence. Contact map is a reduced graph-theoretic representation of proteins which models the local and global inter-residue contacts in the structure. We start with a population of random contact maps for the protein sequence and “evolve” the population to a “high-feasibility” configuration using a genetic algorithm. A neural network is employed to assess the *feasibility* of contact maps based on their four physically relevant properties. We also introduce five parameters, based on algebraic graph theory and physical considerations, that can be used to judge the structural similarity between proteins through contact maps. To predict the fold of a given amino acid sequence, we predict a contact map that will sufficiently approximate the structure of the corresponding protein. Then we assess the similarity of this contact map with the representative contact map of each fold; the fold that corresponds to the closest match is our predicted fold for the input sequence. We have found that our feasibility measure is able to differentiate between feasible and infeasible contact maps. Further, this novel approach is able to predict the folds from sequences significantly better than a random predictor.

I. INTRODUCTION

Proteins are the basic functional units of the cell which carry out almost all the activities of life. Thus there is great interest in understanding the composition, structure and function of these biological polymers composed of 20 different amino acids. A typical protein has around 50-500 amino-acids. The number of all possible sequences of this length is exponentially high (20^{500}), although the number of sequences actually observed in nature is very small (less than a million). Further, the number of proteins whose structure is known is significantly smaller (20 thousand) [3]. Structure determination is usually done by long and tiring processes like NMR spectroscopy or X-ray crystallography. These processes are very slow, and hence the need arises for in-silico determination of

Corresponding author: Nitin Gupta (nitiniitk@yahoo.com)

protein-structure given the sequence. This problem, often referred to as the protein folding problem, has proven too difficult to have any exact or near exact solution.

Although the number of possible structures that a sequence of 500 amino acids can adopt is exponentially large, given each amino acid has two degrees of freedom (two dihedral angles), the structures observed in nature are very limited. The known structures can be classified into a relatively small number of folds to which most of the known proteins belong [3]. This means that a new sequence with unknown structure is likely to adopt one of the already known folds, though this need not be true in every case. There are publicly available databases which have classified these known folds into an hierarchy (e.g. SCOP, CATH etc.). Each class has some number of associated proteins. If for the sake of simplicity we assume that these folds are the only structures a protein sequence can adopt, then the protein folding problem can be simplified to choosing the most appropriate of these folds for a given sequence. This latter task is known as fold prediction.

A popular method used for fold prediction is threading, in which the known sequence is structurally superimposed on a directory of known templates and its energy is evaluated in these conformations [2]. The conformation with the least energy is likely to be the native fold of the protein. The success of such methods, however, is largely dependent on the formulation of the energy function which is not very well understood. There have been attempts to predict fold without the knowledge of energy functions, based purely on the statistical properties of the sequence. There have also been attempts to incorporate secondary structure information. For instance, Shepherd et al. have used secondary structure prediction for fold recognition using Fourier transforms and neural networks [1].

In this work, we have tried to develop a new fold prediction method to predict fold starting from the sequence of the protein. We use the notion of contact maps, a graph theoretic abstraction of the detailed protein structure which models the contacts between the residues in the protein. Our fold prediction method can be broadly divided into two parts:

- 1) Prediction of the "most feasible" contact map from the amino acid sequence of the protein, using a novel genetic algorithms (GA) approach. We have identified measurable properties of contact maps that can be used to give a numerical value to feasibility or the compatibility of a contact map with the given amino acid sequence. This numerical value of feasibility, which has been computed by using a neural network, is then used as a fitness criterion for selecting the *better* contact maps in the genetic algorithm.
- 2) Predicting the fold of the protein by measuring the structural *similarity* of the predicted contact map with the representative contact map of each fold. For this purpose, we have identified five easy-to-compute parameters of contact maps; our measure of similarity of two contact maps is how close their corresponding parameter values are. An important contribution of this study is the use of results from algebraic graph theory [20] for determining the similarity between two given contact maps.

Figure 1 illustrates our approach towards fold prediction. Section 2 of this paper gives an introduction to contact maps. Section 3 discusses the genetic algorithm that we use to predict contact maps. Section 4 describes a neural network based approach to measure fitness of candidate contact maps considered by our genetic algorithm. Section 5 discusses the method we have used to assess how similar is one contact map to another. Section 6 summarizes

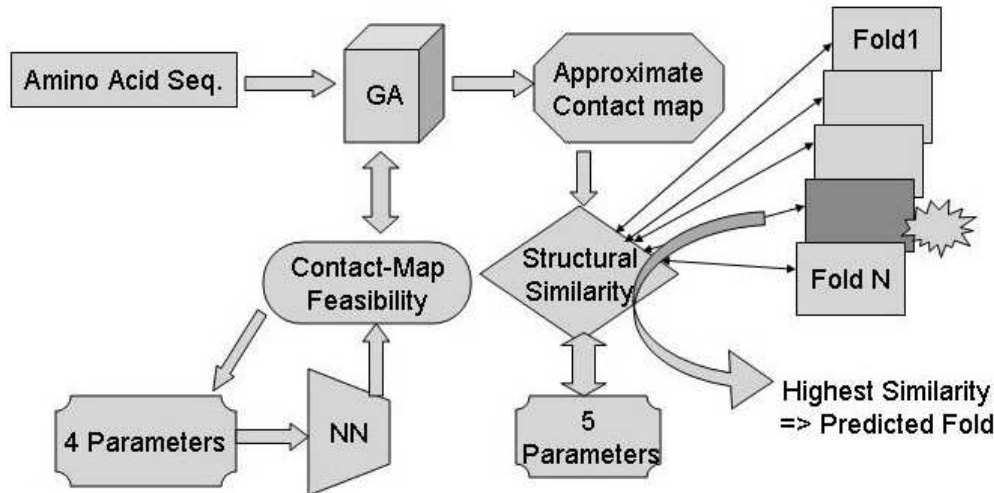


Fig. 1. An overview of our fold prediction strategy. The first step in the pipeline is the prediction of a sufficiently accurate contact map for a given amino acid sequence using GA. This step requires the computation of feasibility of contact maps, which is done using 4 parameters and a neural network. This predicted contact map is then compared for structural similarity against the representative contact maps of each fold in the library of folds using 5 parameters. The fold with the highest score is predicted to be the optimal fold for the input amino acid sequence.

the study, and the next section discusses the limitations of our work and the scope for future work.

II. CONTACT MAPS

We believe that prediction of protein structure can be improved if we have a minimalist and yet sufficiently powerful representation of protein structures. A contact map, which is a graph theoretic abstraction of a 3-D structure, is such a representation. The 3D conformation of a protein may be compactly represented in a symmetrical, square, boolean matrix of pairwise, inter-residue contacts. The contact map provides useful information about the protein's secondary structure, and it also captures non-local interactions which give clues to its tertiary structure. Two amino acids in a protein that come into contact give rise to a non-covalent interaction (hydrogen-bonds, hydrophobic effect, etc.). These contacts are the defining feature of any protein structure and can be captured by contact maps.

Formally, the contact map of a polypeptide chain of length N is represented by an $N \times N$ matrix S , which is defined in terms of distances and a given cutoff distance μ (usually taken as 6 to 7 Angstroms).

$$S_{ij} = \begin{cases} 1 & \text{if } \delta(i, j) < \mu, \quad |i - j| \geq 3 \\ 0 & \text{Otherwise} \end{cases} \quad (1)$$

$\delta(i, j)$ above denotes the distance between the residues i and j . Note that we require a minimum sequence distance of 3 to call it a contact, otherwise all the sequence neighbors will trivially become contacts.

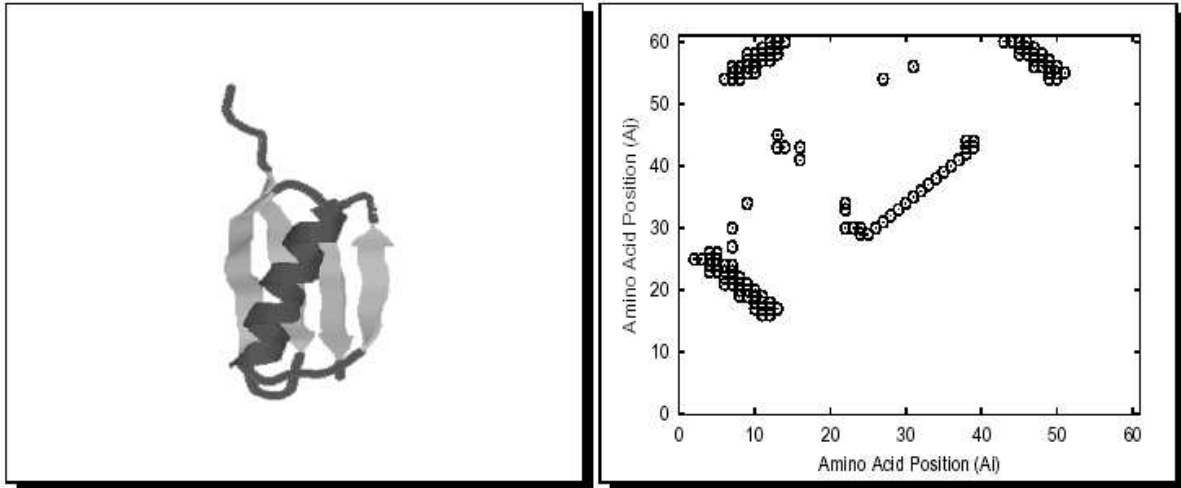


Fig. 2. The figure on left shows the ribbon representation of secondary structures in the protein 2igd. The second figure is the contact map for this protein. We note that secondary structures are distinctly visible: parallel (top-left cluster) and anti-parallel beta sheets (bottom-left and top-right clusters) and alpha-helix (this cluster parallel to main diagonal). Figures adopted from [10].

A. Physical Significance of Contact Maps

The contact map provides a host of useful information. For example, secondary structure can easily be discerned from it. α -helices appear as thick bands along the main diagonal since they involve contacts between one amino acid and its four successors, while β -sheets are thin bands parallel or anti-parallel to the main diagonal, etc. [10] (See Figure 2). Information about the tertiary structure is not, however, easy to infer from a contact map.

The matrix of contact map has some peculiar characteristics that represent physical properties of proteins. Instead of having a uniform distribution of 0's or 1's, they usually have localized regions predominantly containing either 0's or 1's. The "dense" regions in the matrix, which have mostly 1's correspond to local and well-connected substructures in the protein. These substructures might not be very well connected to each other, and we thus get hollow regions in the matrix S . There have also been attempts to develop energy functions for contact maps [12].

III. PREDICTING SUFFICIENTLY INFORMATIVE CONTACT MAPS USING GENETIC ALGORITHM

Given an amino acid sequence, we try to predict a contact map that will sufficiently approximate the structure of the corresponding protein for the task of fold prediction. We note that this task is far simpler than predicting the complete three dimensional structure from the sequence and should thus be more "tractable". Previous works on contact map prediction have employed neural networks [14], statistical techniques based on correlated mutations [13] and association mining of HMM patterns. Recent work by Vendruscolo et al [11] has also shown that it is possible to recover the 3D structure from even corrupted contact maps. Fariselli and Casadio [14] have used a neural network based approach over pairs database, with other contextual information like sequence context windows, amino acid profiles, and hydrophobicity values. Olmea and Valencia [13] used correlated mutations in

multiple sequence alignments for contact map prediction. They added other information like sequence conservation, alignment stability, contact occupancy etc. to improve accuracy. [10] employs a hybrid approach for this task. Firstly the local structural elements are predicted using HMM. Association mining is then applied on top of the HMM states to predict the states that frequently co-occur with contacts which are used for predicting contacts in unseen proteins. The results obtained by these various approaches have reported accuracy less than 30% accuracy, only 5-6 times better than result expected from random prediction.

We have used the genetic algorithm (GA) paradigm to find, given a sequence, a contact map that will adequately approximate the structure of the corresponding protein for our fold prediction task. To the best of our knowledge, GA paradigm has not been used before for searching in the contact map space. The genetic algorithms, also sometimes referred to as evolutionary algorithms, have previously been used in various engineering applications when the search space is very large, and point-by-point searches are not feasible [18].

We start with the amino acid sequence of a given protein, and the corresponding secondary structure tags for each of the residues. These amino acid sequence of proteins are readily available and the corresponding secondary structure tags can be predicted with the available software. An example is PREDATOR of Frishman and Argos [15]–[17]. They predict the α -helices, β -strands and turns based on recognition of potentially hydrogen-bonded residues in the amino acid sequence and using pairwise alignments. The reported accuracy of the method is 75% which is achieved by accurate secondary structure propensities for individual sequences which account for long-range effects, utilization of homologous information in the form of carefully selected pairwise alignment fragments, and reliance on a much larger collection of protein primary structures. The other frequently used programs are PHD of Rost and Sander [8] and NNPREPDICTION of Kneller et al [9]. From secondary structure prediction, we get the local conformation in some regions of the protein sequence. This helps in putting initial contacts with high confidence. These initial contacts can be used as a framework for the prediction of the remaining contacts.

A genetic algorithm can be used to find a solution in much less time than Monte Carlo simulations or simulated annealing. Although it probably will not find the best solution, it can find a reasonably good solution in less than five minutes on a 1 GHz workstation. The GA involves the following steps as given by the standard GA paradigm [19]:

- 1) Start with an initial *population* of size NPOP containing randomly generated contact maps. Note that the contacts corresponding to the secondary structure are assigned to each individual of the population, and rest of the bits are randomly generated. These secondary structure contacts are preserved during the subsequent operations in our GA. The *chromosome* of each individual is a lower triangular matrix corresponding to the adjacency matrix of each contact map.
- 2) Repeat the following until NPOP offsprings(new contact maps) have been created:
 - Reproduction(tournament selection): select a pair of parent chromosomes from the current population, the probability of selection being an increasing function of fitness. Selection is done ‘with replacement’, therefore a given parent chromosome may be selected more than once.

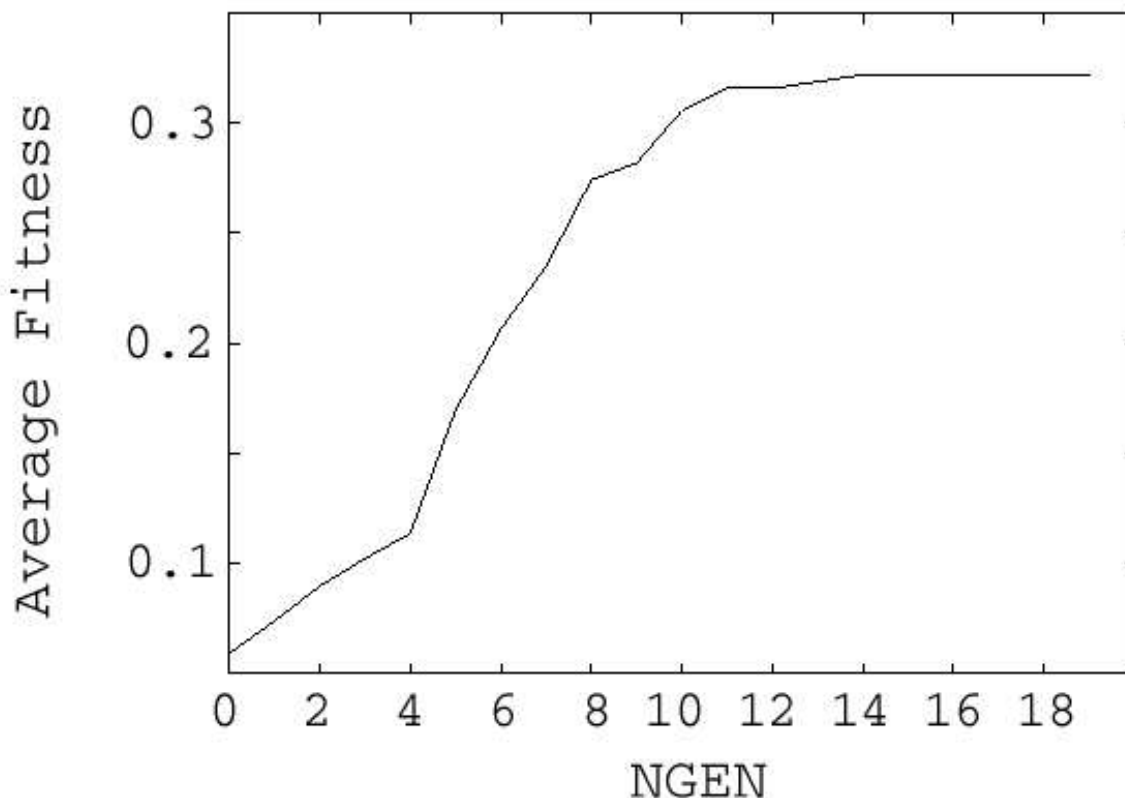


Fig. 3. Shows the increase in average fitness of contact maps in the population with generations/iterations. The average fitness eventually converges to the maximum fitness.

- Crossover (single point): with a probability PCROSS, select a random point along the diagonal of the triangular matrix and swap the parts of the matrices lying above this point.
- Mutation: with a probability PMUT, a bit flip is performed at each position in each of the new chromosomes except those which correspond to the secondary structure.

3) Replace the current population with the new population.

4) Goto 2 unless we have done NGEN generations.

We have set NPOP to 20, NGEN to 20, PCROSS to 0.8 and PMUT to 0.2. The aim of the GA is to find the most compatible and representative contact map for the given amino acid sequence and this is achieved by the fitness measure that we use. Thus the fitness assigned to each individual in the population should represent its level of feasibility. We have formulated a feasibility score to evaluate a contact map for a given amino acid sequence based on its physical properties by using a neural network. The next section describes this feasibility measure in detail. Thus, compared to a random contact map, a contact map obtained through GA is more feasible to the protein sequence and more "similar" to the actual contact map of the protein. The findings from the similarity measure

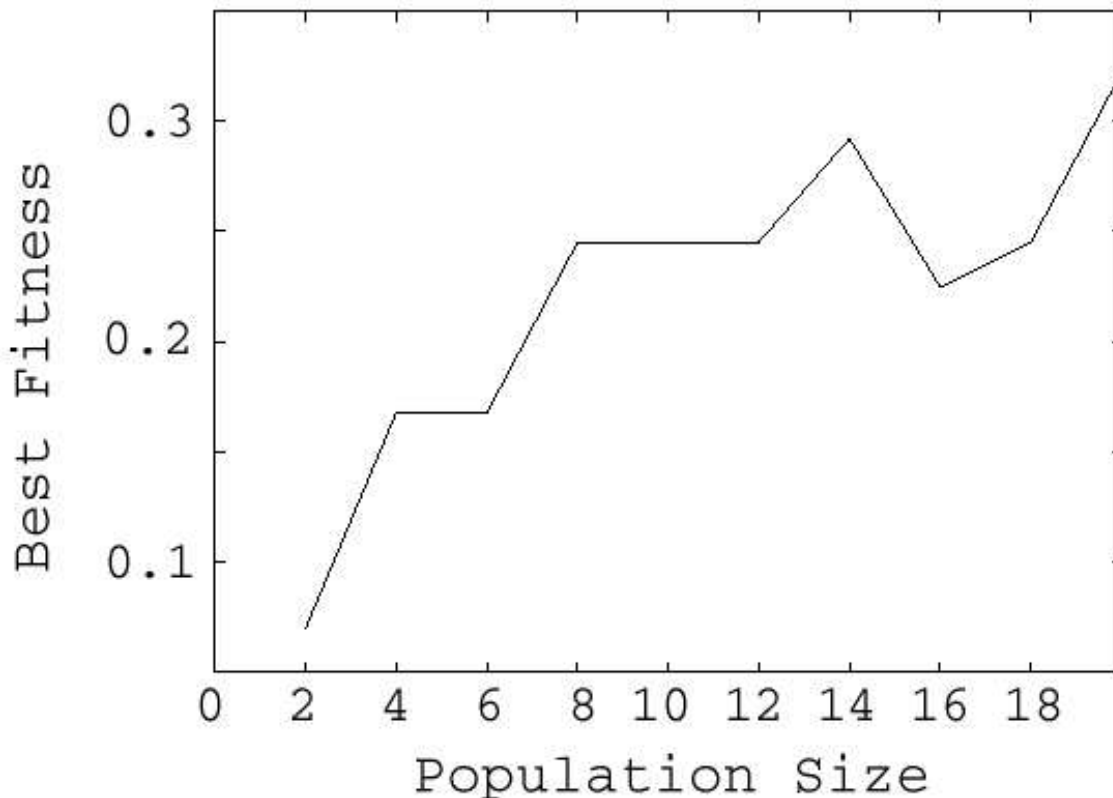


Fig. 4. Effect of population size on the fitness of the predicted contact map. A larger population increases the possibility of searching a better individual.

discussed in section IV confirm this.

Genetic algorithms traverse to an acceptable solution by simulating the way nature uses evolution. GA uses "survival of the fittest" with the different solutions in the population. The good solutions reproduce to form new and hopefully better solutions in the population, while the bad solutions are removed. Thus the average fitness of solutions in the population increases with generations, as shown in Figure 3. Figure 4 indicates that it is also important to start with a big population size to get a good solution.

IV. TESTING FEASIBILITY OF CONTACT MAPS

In this Section, we define the fitness criterion used in our GA. Given a representation of a contact map, i.e., a binary lower triangular matrix, we compute what we call its *feasibility value* which is a measure of how plausible is this representation to be an approximation to the contact map of some protein. The contact map is represented by a two dimensional matrix of 0's and 1's, but not all such matrices would correspond to feasible three dimensional proteins. For example, a matrix of size 100x100 in which all entries are 1 would correspond to a protein of sequence length 100 in which every amino acid is close (within 7 Angstroms) to every other amino acid. Biologists will easily

know that such a protein cannot exist. Thus not all binary square matrices that can be constructed mechanically will have their biological parallel, and we need some method to determine whether a given such matrix approximates the contact map. Also the method should be based on simple parameters that can be computed easily in relatively short times, and do not require manual intervention or actual 3-D reconstruction so that it can be used for high-throughput validation, as required by GA. We find the following four parameters useful in this differentiation, and thus base our method to identify feasible contact maps on these four parameters.

A. Charge

Electrostatic forces, such as those acting within salt-bridges (acid-base pairing) or between polar residues, play an important role in determining the structure of the protein. A stable configuration of a protein would require that these interactions impart an overall negative energy to the system. This will happen when the close/interacting pair of amino acids have opposite charges more often than similar charges. This is intuitive, because like charges repel and unlike charges attract, and thus the pairs of amino acids that are in contact will have opposite charges in general in a feasible contact map.

NC is the number of contacts in the proteins, and P_c is the calculated parameter. Here, c_i represents the charge of amino acid at position i .

$$NC = \sum_{i=1..N, j < i} S(i, j) \quad (2)$$

$$P_c = \left(\sum_{j < i, S(i, j) = 1} c_i c_j \right) / NC \quad (3)$$

Feasible contact maps show negative values for P_c .

B. Neighborhood Hydrophobicity

Hydrophobicity in a region of protein indicates the affinity in a part of the protein to repel water. A region of high hydrophobicity will be energetically stabilized if it is in proximity to another region of high hydrophobicity, rather than being close to a hydrophilic (polar) region. Similar arguments can be given for regions of low hydrophobicity too. Thus we expect that amino acids that are in contact in the protein lie in neighborhoods of similar hydrophobicity. We use the Kyte-Doolittle scale of hydrophobicity [7] which rates the 20 amino acids on a numerical scale of -4.5 to $+4.5$.

N_i denotes the set of residues that are in contact with residue i . H_i is the hydrophobicity of residue i . $\langle H_i \rangle$ denotes the neighborhood hydrophobicity for residue at position i . We use a scaling factor μ (between 0 and 1) to determine the relative weightage of residue i in determining the strength of hydrophobicity around itself. The value of μ used in our calculations is 0.66. The idea is that the residue itself will affect the hydrophobicity around itself more significantly than other residues.

$$N_i = \{j | S(i, j) = 1\} \quad (4)$$

$$\langle H_i \rangle = \left((\mu) H_i + (1 - \mu) \sum_{j \in N_i} H_j \right) / (|N_i| + 1) \quad (5)$$

$$P_h = \left(\sum_{j < i, S(i,j)=1} \langle H_i \rangle \cdot \langle H_j \rangle \right) / NC \quad (6)$$

P_h is the parameter for hydrophobicity, and feasible contact maps show higher values for this parameter as compared to randomly generated contact maps.

C. Sequence Distance

Another parameter of interest is the average sequence distance between the residues that are in contact in a protein. The contacts within the local conformations of a protein will have small sequence distances while the contacts resulting from global conformations will have larger sequence distances. Thus P_s will give an idea about relative frequency of local and global contacts and this simple feature might be useful in distinguishing between feasible and infeasible contact maps.

If d_{ij} denotes the sequence distance between residues i and j , then the sequence distance parameter P_s can be formulated as:

$$P_s = \left(\sum_{j < i, S(i,j)=1} d_{ij} \right) / NC \quad (7)$$

D. Degree of Vertices

The degree of a vertex in the contact map corresponds to the number of contacts that a residue forms in the protein. Thus degree gives a measure of crowding around a residue. The higher the degree, the more crowded is the residue and thus the less is its affinity to form another contact.

$$\delta_i = |N_i| \quad (8)$$

$$\delta_{ij} = (\delta_i + \delta_j) / (2 * \delta_{max}) \quad (9)$$

$$P_v = \left(\sum_{j < i, S(i,j)=1} \delta_{ij} \right) / NC \quad (10)$$

We first compute δ_i as the degree of residue i , the number of contacts that it forms. δ_{ij} indicates the crowdedness of contact between residue i and j , and is used in the calculation of parameter P_v . δ_{max} is the maximum degree of any residue in the given protein.

	True Positive	True Negative	False Positive	False Negative
P_c	14	16	8	10
P_h	0	24	0	24
P_s	21	23	1	3
P_v	17	17	7	7
All Parameters	21	24	0	3

TABLE I

PERFORMANCE OF DIFFERENT PARAMETERS IN JUDGING THE FEASIBILITY OF CONTACT MAPS. THE TRAINING DATASET CONTAINS 200 PROTEINS, WHILE THE TEST DATASET COMPRISES OF 24 DIFFERENT PROTEINS, 8 EACH BELONGING TO 3 DIFFERENT FOLDS. THE POSITIVE EXAMPLES ARE THE ACTUAL CONTACT MAPS OF THESE PROTEINS, AND THE NEGATIVE EXAMPLES ARE OBTAINED BY RANDOMIZING THESE CONTACT MAPS. A CONTACT MAP IS HERE CALLED A FEASIBLE MAP IF ITS PREDICTED VALUE IS ABOVE 0.5.

E. Combining all parameters

The above mentioned four parameters are independently useful in distinguishing feasible and infeasible contact maps to a certain extent. We now combine all the four into a single parameter, with appropriate weightages assigned to each of the parameters. To learn the optimal weights, we use a two layered feed forward backpropagation neural network. The input to the network are four values, one corresponding to each of the parameters and scaled to the interval $[0, 1]$. The output is a single value indicating how feasible is the given contact map. The training is done using a dataset of proteins. The examples of feasible maps are the contact maps built from these proteins, and the examples of infeasible maps are given by randomizing the above contact maps. The underlying assumption is that the randomization of a given contact map (that is, arbitrarily shuffling the contacts) results in an infeasible contact map. This is reasonable because we expect the space of infeasible maps to be much larger than the feasible ones. Note that the contacts corresponding to the secondary structure are not altered in these randomizations. Thus even the randomized contact maps we have generated have this information, indicating that feasibility estimate is not dependent upon the secondary structure which is computed by external programs.

Table I shows the performance of various parameters in judging the feasibility of contact maps. It is obvious that the performance increases significantly when all parameters are used together, instead of using any of them alone. We also note that the parameter P_h , gives all negatives, and is not useful when used alone. But it might be playing the useful role of a filter when used in conjunction with the other parameters for preventing false positives. Figure 5 shows the values obtained for each of the test proteins. We note that there are three false negatives. A close examination of these three proteins(1eytA, 1c9fA and 7msiA) reveals that they perform poorly on P_s , owing to a larger percentage of short range contacts than expected. Ideally, we would expect the neural network to assign weightages to the parameters such that poor performance on one index is optimally countered by good performance on others. In fact, many proteins that are false positives or false negatives on individual parameters in Table I get corrected in the final score when all parameters are included, but the three proteins remain falsely identified. Such

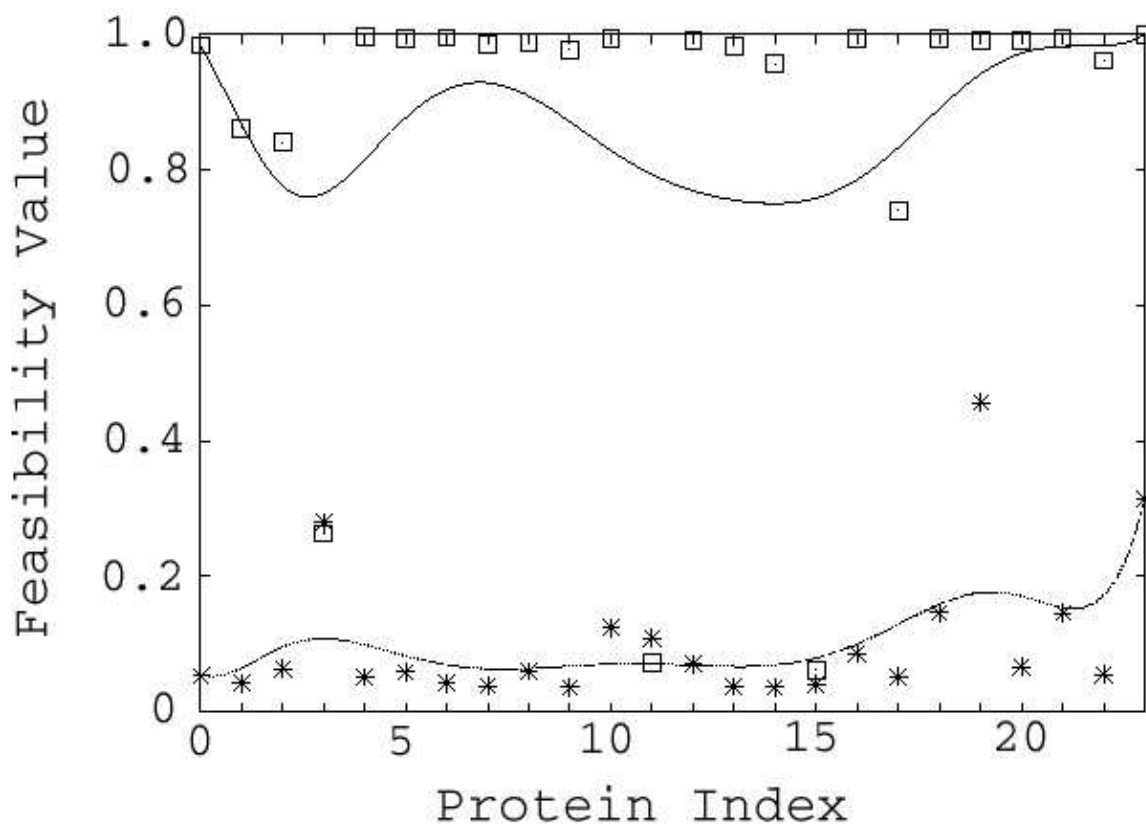


Fig. 5. Performance of neural network in testing feasibility of contact maps after combining all parameters. The boxes represent the actual contact maps while stars represent the randomized contact maps. Corresponding bezier curves are also shown. The figure shows the marked difference in feasibility values obtained for actual contact maps and the randomized contact maps.

limitations are common in machine learning approaches, and could be attributed to various reasons such as limited training set size. Also there might be other parameters besides the four studied in this work that should be included to represent the feasibility of a contact map more accurately.

V. STRUCTURAL SIMILARITY THROUGH CONTACT MAPS

The contact map that our GA outputs for a protein sequence with unknown structure cannot be used directly to know the structure. Can it be used to compare structural similarity with the proteins of known structures? The basic assumption behind our study is as follows:

- 1) The proteins within the same fold have more structural similarity than proteins of different folds, i.e. in fold classification of proteins, the ones with the same group are structurally more similar.
- 2) A sufficiently accurate contact map of a protein, in spite of being a reduced representation, does incorporate some useful information about its structure which can be used to evaluate its structural similarities with other proteins.

The underlying problem is thus to define a similarity measure for contact maps that can utilize this information present in the contact maps, and is consistent with the fold classification of proteins. Our ability to find such a measure justifies the assumption stated above. Contact map overlap is a commonly considered measure of structural similarity, and is computed from a monotonic one-to-one mapping between two sub-sets of the two sequences that maximizes the number of "contact" pairs that are mapped to each other. A contact map (n, E) is an undirected graph $G = (V, E)$ such that the set of vertices $V = \{1, 2, 3, \dots, n\}$ is linearly ordered. The contact map overlap problem is the following optimization problem: Given two contact maps (n, E) and (m, E') , find two subsets $S \subseteq \{1, 2, \dots, n\}$ and $S' \subseteq \{1, 2, \dots, m\}$ with $|S| = |S'|$ such that the cardinality $|\{[u, v] \in E : u, v \in S, [f(u), f(v)] \in E'\}|$ is as large as possible, where f is an order-preserving bijection between S and S' . It has been proven that calculating contact map overlap is NP-hard, although can be solved for several special cases in polynomial time. Also it is doubtful if such a rigorous approach would be useful in case of real proteins in which local structures are rigid in some parts (like active sites) and flexible in others. We are, therefore, trying to develop a similarity measure that is easily computable, captures the important structural properties, and is not too sensitive to slight fluctuations.

A metric for structural similarities of contact maps can be very useful in fold prediction when we have an algorithm for obtaining acceptably approximate contact maps from sequences. The advantage of this approach over the usual homology based techniques is that it does not rely on sequence alignments. As more structures are being discovered, it is becoming evident that proteins can adopt similar overall structures inspite of completely different amino acid sequences. Our approach is discussed below.

A. Parameterized Approach for Similarity

We have devised five parameters, based on algebraic-graph theory and also on physical properties of proteins, which are helpful in comparing the structural similarity of two given contact maps. The parameters are based on properties that can characterize the structure of a protein. The parameters give a high score for contact maps with high structural similarity, and a low score otherwise.

1) *Distinct Eigen Values:* Algebraic graph theory deals with the use of algebraic techniques in the study of graphs. The properties of graphs are translated into algebraic properties and then, using the results and methods of algebra, theorems are deduced about graphs [20].

Here we use one such result from algebraic graph theory and see its application in determining similarity in contact maps which can be seen also as undirected graphs. The result states that a connected graph with n vertices and diameter d has atleast $d + 1$, and at most n , distinct eigen values [20]. Note that in graph theory, the diameter of any graph represents the maximum distance between any two vertices in the graph, in terms of the number of intermediate vertices that need to be traversed to go from one vertex to the other. This suggests that the number of distinct eigen values has some relation with the diameter and thus with structure of the protein. We expect proteins with similar structures to have similar values of η_e/N , where η_e is the number of distinct eigen values, and N is the length of the protein. Two contact maps are assigned a higher similarity score, S_e , if their values of η_e/N are closer.

2) *Fiedler Eigen Value*: The second smallest eigen value in the adjacency matrix of a graph has been found to be related with structural properties of the graph. This value, often known as Fiedler value [21], is a measure of the compactness of the associated tree graph. An application of this value is found in [22], where it is used for RNA folding prediction.

We determine a similarity parameter, S_f , based on the similarity in the Fiedler eigen values of the two given contact maps. S_f is higher if the Fiedler values are closer.

3) *Partial Overlap*: This is equivalent to finding the most similar fractions of proteins, then judging the extent of their similarity. We fix one contact map, and slide the other one on it to find the biggest overlapping region and with the least Hamming distance in the overlapping bits. This similarity measure, S_o , is very useful when two proteins have some part similar, and some parts varying, or if one of the proteins is itself a part of the other bigger protein. Partial overlap can find out the parts that are similar, and give a corresponding high score.

4) *Helix Component*: The percentage of residues that take part in forming alpha-helix is a characteristic of protein structure. Two contact maps that have the same proportion of helix are more likely to have a similar structure than the ones which have a marked difference in their helix composition. The parameter, S_h is higher if the given contact maps have a similar percentage of helix in their structure.

5) *Surface Hydrophobicity Fraction*: We have developed an algorithm to distinguish between the surface residues and core residues of a protein, given its contact map. This is an example of the utility of contact maps when complete structures are not available. Firstly, a list of residues is formed which have a limited number of contacts. The residues in the core of a protein are more densely packed and thus form a larger number of contacts as compared to surface residues. We have imposed an upper limit of 7 contacts that a surface residue can form in our calculations. For the selected residues, we compute the average neighborhood hydrophobicity and reject those residues whose hydrophobicity exceeds a threshold. This ensures that the residues selected for surface lie in a hydrophilic environment, which is available on the surface.

The knowledge of surface residues in a contact map can be used in measuring similarity across maps of different proteins. We compute the total hydrophobicity of surface residues as a fraction of the total hydrophobicity of the protein. We observe this surface hydrophobicity fraction to be similar in proteins with similar structures. Thus the similarity parameter S_s is given a higher value if the surface hydrophobicity fractions are similar in the given contact maps.

B. Application to Fold Prediction

The five similarity parameters can be combined to form a composite similarity measure for contact maps. Thus, given two contact maps and their corresponding similarity scores based on above parameters, the composite similarity is given by S_{all} .

$$S_{all} = \phi_e S_e + \phi_f S_f + \phi_o S_o + \phi_h S_h + \phi_s S_s \quad (11)$$

where $\phi_e, \phi_f, \phi_o, \phi_h, \phi_s$ are relative weights of each of the parameters and all are currently set to 0.2 (thereby giving equal weightage to all the parameters in similarity testing).

	F1-F1	F2-F2	F3-F3	F1-F2	F1-F3	F2-F3
S_e	0.97	0.95	0.65	0.94	0.62	0.62
S_f	0.76	0.59	0.57	0.67	0.50	0.43
S_o	0.97	0.97	0.96	0.97	0.97	0.97
S_h	0.74	0.88	0.64	0.59	0.60	0.68
S_s	0.97	0.92	0.80	0.95	0.80	0.77
S_{all}	0.88	0.86	0.72	0.82	0.69	0.69

TABLE II

AVERAGE SIMILARITY AMONG THREE FOLDS, DENOTED BY F1, F2 AND F3. WE OBSERVE THAT INTRA-FOLD SIMILARITY IS HIGHER THAN INTER-FOLD SIMILARITY FOR S_{all} , I.E. WHEN ALL THE SIMILARITY PARAMETERS ARE USED.

This similarity measure can be used for predicting the fold of a predicted contact map. For this, we also create contact maps of proteins with known structures and fold affiliations. These contact maps serve as templates to which the predicted map can be matched. The template with which the predicted contact map gives the highest similarity corresponds to the closest structure, and thus the predicted map is likely to have the same fold as this template.

C. Results

We validated our method using a dataset of 24 proteins. These belong to three folds or architectures, as defined in CATH [5]. These folds are 1.20, 3.10 and 4.10, and each had corresponding 8 arbitrarily selected proteins. We start with only the sequence of a protein, predict its contact map, and compare it with the contact maps derived from known structures of all the 24 proteins. The average similarity values obtained between proteins of two different folds are summarized in Table II. The first three columns show the average similarity within the folds, while the last three columns show inter-fold structural similarity. We expect a fold to show more similarity to itself than others. This is indeed what we observe in Table II.

Table II also shows the performance of individual similarity parameters. The overall performance of the method is just the equal-weighted sum of the five parameters, and appears better than the individual ones. We also observe that though the average similarity of a fold with itself is higher than its similarity with other folds, the similarity values are not remarkably different. This can be improved upon by properly scaling the similarity scale. The values obtained for S_{all} are between 0.69 and 0.88 and this interval can be mapped to the interval [0,1]. This will widen the differences between the intra-fold and inter-fold similarities.

VI. SUMMARY

In this study, we have proposed new approaches for contact map prediction, measuring the feasibility of contact maps and evaluating the structural similarity in proteins through their contact maps. We have combined these novel

approaches into an algorithm for predicting which fold the amino acid sequence is likely to assume from a number of given folds.

Contact maps are graph theoretic abstractions that contain partial information about the structure of a protein. We use genetic algorithms for predicting a sufficiently accurate contact map for a sequence with unknown structure. We have noted some physical constraints and properties of proteins that are also visible in the reduced contact map representation, and have developed a method based on these properties to assess the feasibility of generated contact maps. We have also identified parameters that are helpful in finding the similarity between any two contact maps. These parameters are based on algebraic graph theory and on those physical properties of proteins that are relevant to the structure.

Contact map prediction and similarity measure can together be used for fold prediction of a protein sequence. To determine the fold of a protein starting from its sequence, we first predict the contact map of the protein, and then find the similarity of this contact map with one template contact map from each fold. The maximum scoring template determines the fold of the given protein. The results obtained are in accordance with our expectations.

We have divided fold prediction into two steps: firstly predicting a sufficiently accurate contact map and secondly deciding the fold of the protein by evaluating the similarity of the predicted contact map with representatives of each fold. An advantage of this two-tier approach is that each step can be improved upon independently of the other in the long run, and that each has its own utility in isolation. Contact map prediction, if done with reasonable accuracy, can be used to determine various properties of a protein whose structure is not available.

VII. LIMITATIONS AND SCOPE FOR FUTURE WORK

This study is novel in many aspects, but there is certainly scope for improvement. Some probable areas where further experimentation can be done to improve the performance of our method are:

- This study uses genetic algorithms for the first time for predicting contact maps. The GA that we have used is very simple, and many advanced GA techniques and operators can be tried out to see if they improve the performance.
- We have used four parameters for testing the feasibility of a contact map. There may be many more possible parameters that can be used for the purpose. The challenge is, thus, to think of new and independent properties and test their performance.
- We have attempted to use graph theoretic similarity measures in this study, like number of distinct eigen values and the Fiedler eigen value. An attempt can be made to devise new parameters that can be helpful in assessing similarity. One such possibility is the notion of "complexity" of a graph [20], but the problem with this measure is that it is non-linearly dependent upon the length of the input sequence. If one is able to remove length-dependence, this parameter might be useful.
- In this study, we have given equal weightages to the different similarity parameters while evaluating the overall similarity measure S_{all} . This is not an ideal scenario and it is very likely that the results can be improved to some extent by judging the optimal relative importance of these parameters. For this purpose, a machine

learning technique like neural networks or support vector machines may be employed. Our purpose was simply to see which parameters might be useful in similarity testing, rather than giving a final optimal formula of contact map similarity. The latter task is left open for future studies.

- Future work in this direction may use a larger training and validation datasets for more accurate results.

VIII. ACKNOWLEDGEMENT

This study was done as a *B. Tech. Project* by NG and NM under the supervision of SB. We wish to thank the members of our evaluation committee at IIT Kanpur for giving us constructive suggestions during the intermediate project reviews. We also thank Dr. R. Sankararamkrishnan (BSBE, IIT Kanpur) for providing us with the protein data files.

REFERENCES

- [1] Shepherd, A. J., Gorse, D. and Thornton, J. M. (2003) A Novel approach to the recognition of protein architecture from sequence using Fourier Analysis and Neural Networks. *Proteins: structure, function and genetics*, 50, 290-302.
- [2] Jones, D. T., Taylor, W. and Thornton, J. M. (1992) A new approach to protein fold recognition. *Nature* 358,86-89.
- [3] Murzin, A. G., Brenner, S. E., Hubbard T. and Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540.
- [4] Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S., Ramaswamy, R. (1997) Prediction of probable genes by Fourier Analysis of genomic sequences. *Comput Appl Biosci* Jun 13(3), 263-270.
- [5] 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*. Vol 5. No 8. p.1093-1108.
- [6] C. Branden and J. Tooze, *Introduction to Protein Structure* (Garland Publishing, New York, 1991).
- [7] Kyte, J., Doolittle, R.F. (1982). A simple method for displaying the hydrophatic character of a protein. *J. Mol. Biol.* 157, 105-132.
- [8] Rost, B., and Sander, C. (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl. Acad. Sci. U.S.A.* 90, 7558-7562.
- [9] Kneller, D. G., Cohen, F. E., and Langridge R. (1990). Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.* 214, 171-182.
- [10] Hu, J., Shen, X., Shao, Y., Bystroff, C. and Zaki, M. J. (2002) Mining Protein Contact Maps. In *Proc. of BIODDD02: Workshop on Data Mining in Bioinformatics, with SIGKDD02 Conference*.
- [11] Venduscolo, M., Kussell, E., Domany, E. (1997). Recovery of protein structures through contact maps. *Folding and Design*. 2(5):295-306.
- [12] Park, K., Vendruscolo, M., Domany, E. (2000). Towards an energy function for the contact map representation of proteins. *Proteins: Structure, Function and Genetics* 40, 237-248.
- [13] Olmea, O., Valencia, A. (1997) Improving contact predictions by combining of correlated mutations and other sources of sequence information. *Folding & Design*, 2, s25-s32.
- [14] Fariselli, P., Casadio, R. (1999). A neural network based predictor of residue contacts in proteins. *Protein Engineering*, 12(1), 15-21.
- [15] Frishman, D. & Argos, P. (1996) Incorporation of long-distance interactions into a secondary structure prediction algorithm. *Protein Engineering*, 9, 133-142.
- [16] Frishman, D. & Argos, P. (1997) 75% accuracy in protein secondary structure prediction. *Proteins*, 27, 329-335.
- [17] Frishman, D & Argos, P. (1995) Knowledge-based secondary structure assignment. *Proteins: structure, function and genetics*, 23, 566-579.
- [18] Deb, K. (1995). Optimization for engineering design: Algorithms and Examples. *New Delhi: Prentice Hall*
- [19] Goldberg, D.E. (1989). Genetic Algorithms for Search, Optimization and Machine Learning. *Reading, MA: Addison-Wesley*.
- [20] Biggs, N. (1974). Algebraic Graph Theory. *Cambridge University Press*.
- [21] Fiedler, M. (1973). Algebraic Connectivity of graphs. *Czechoslovak Mathematical Journal, Vol.23, pp298*.

- [22] Barash, D. (2003) Spectral decomposition of the Laplacian Matrix Applied to RNA Folding Prediction. *Proceedings of the Computational Systems Bioinformatics (CSB'03), IEEE, 0-7695-2000-6/03*