Evaluation of protein surface roughness index using its heat denatured aggregates

Hrishikesh Mishra, Tapobrata Lahiri

Bioinformatics and Allied Sciences Division, Indian Institute of Information Technology, Allahabad, India

Abstract

Recent research works on potential of different protein surface describing parameters to predict protein surface properties gained significance for its possible implication in extracting clues on protein's functional site. In this direction, Surface Roughness Index, a surface topological parameter, showed its potential to predict SCOP-family of protein. The present work stands on the foundation of these works where a semi-empirical method for evaluation of Surface Roughness Index directly from its heat denatured protein aggregates (HDPA) was designed and demonstrated successfully. The steps followed consist, the extraction of a feature, Intensity Level Multifractal Dimension (ILMFD) from the microscopic images of HDPA, followed by the mapping of ILMFD into Surface Roughness Index (SRI) through recurrent backpropagation network (RBPN). Finally SRI for a particular protein was predicted by clustering of decisions obtained through feeding of multiple data into RBPN, to obtain general tendency of decision, as well as to discard the noisy dataset. The cluster centre of the largest cluster was found to be the best match for mapping of Surface Roughness Index of each protein in our study. The semi-empirical approach adopted in this paper, shows a way to evaluate protein's surface property without depending on its already evaluated structure.

Introduction

Structural component of a protein that is responsible for its function is basically localized on its surface. Therefore we may expect that knowledge on protein surface may give us the clue on its functional site. Current practice of drawing insight about protein surface is usually accomplished from the knowledge of its structure. But the difficulty embedded in this procedure is described in the following section.

First let us discuss the difficulty in obtaining a protein structure both by experimental as well as predictive methods. Three-dimensional structure of proteins is considered as sole and key factor in defining their function. Functional form of a protein is generally the tertiary structure or in some cases the quaternary structure that results by folding of amino acid chain into tertiary structure and further arrangement of tertiary structure units. The folding of a protein results in arrangement of the amino acid residues in specific positions in 3D space which form the functional site of that protein. The functional sites are always located on the surface of proteins only[1].

Current approach for studying protein surface requires a pre-evaluated 3D structure of protein. 3D structure of protein can be derived through X-ray crystallography, NMR and in some cases homology modeling and other prediction methods. In spite of the great contribution especially of X-ray crystallography and NMR to contribute to the development of other molecular structure exploring methods (e.g., homology modeling, threading) including ours, where every such methods utilize the molecular structural knowledge gained from them to build the methodology, they have their own pitfalls as constraints like size limit in NMR method[2], requirement of crystal in X-ray crystallography method[3], which pose strong limitation in their applicability for all proteins. In addition, the time and experimental complexity involved in these methods are very high. On the other hand, structure prediction methods like homology modeling depend upon the repository of already evaluated structures which are close to target protein with at least 25% sequence similarity through position specific scoring matrix (PSSM). PSSM is also a stringent criteria that is difficult to fulfill for most of the proteins[4]. Furthermore, accuracy of prediction methods including homology model is questionable under further optimization through energy minimization process (for example, the method adopted in Insight-II) which quite often yields minimum energy structure with very low Ramachandran score[5].

Protein structure can be evaluated through microscopy also although the number so far is very few. According to current PDB statistics 252 structures in PDB were evaluated through electron microscopy (http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=explMethod-em&seqid=100). Cryo-electron microscopy has been used for this purpose that enables the determination of 3D structures of macromolecular complexes and cells from 2 to 100 Angstrom resolution (http://emdatabank.org/)[6,7]. Similarly atomic force microscopy (AFM) was also used to yield 3D structure of proteins[8]. It was shown that AFM gives true atomic resolution in ultra-high vacuum (UHV) and, more recently, in liquid environments[9]. High resolution AFM is comparable in resolution to transmission electron microscopy. But the resolution obtained through microscopy is not comparable with that obtained using x-ray crystallography and NMR, although, electron density maps obtained through electron microscopy can be helpful to increase the quality of models obtained through comparative modeling[10].

Considering the shortcomings of existing methods of protein-structure evaluation, there is a huge sequence-structure gap. The number of structures was 55285 As per PDB statistics of August 25, 2009 (http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html) in comparison to 495880 sequence entries in UniProtKB/Swiss-Prot Release 57.6 of July 28 2009 (http://au.expasy.org/sprot/relnotes/relstat.html). Thus structure is known for only 11.15% of protein sequences existing in sequence database.

Secondly, there was also a difficulty in extracting usable surface information from protein structure as reported by the existing methods. Even if the structure for a protein can somehow be evaluated with the difficulty described above, the next level of difficulty commonly faced by researchers is the extraction of information on protein's functional site from its structure. This is a hard to solve problem till today giving approximately 0.5% success only over the whole set of evaluated structure for enzymatic proteins reported in Pfam database[11]. The reason behind this difficulty can be explained in the paradigm of protein surface where, obtaining surface from structure is easy; however the usability of that surface remains an issue over decades because of the absence of proper surface characterizing parameter. In this light, reports of the following investigations are listed as recent example of works.

Research on protein surface showed how surface active incidents of proteins provide clue to individual properties of proteins or their families. The nature of binding sites on protein surface differs according to binding molecule with which they interact. Binding molecule may be a small molecule or a drug, another protein or a nucleic acid. Binding sites on protein-surface are basically part of the surface that may be either more rough or clefts for binding small molecules and drugs, or flat area for binding another protein or convex area for binding nucleic acids[12,13]. There is no one to one relationship between surface pockets and the ligands because one ligand may have affinity towards more than one protein surfaces and same protein surface may bind to different ligands[14].

In an important approach of study on protein surface roughness, Bowie and Pettit, showed a correlation between protein surface roughness and small molecular binding sites[15]. By studying Smoothed Atomic Fractal Dimension (SAFD) for each atom, they concluded that binding sites for small ligands are rougher in comparison to the binding sites for large ligands like DNA. It is so because the binding sites for larger ligands like DNA or proteins have large surface area greater than 600 $Å^2$. In these sites binding occurs by hydrophobic interactions. But for smaller binding sites additional binding interactions are required. Roughness of these small binding sites reflects the complex local shapes required for binding interactions.

Study done by Varadwaj et al., (2005) represented the Surface property of protein molecule by Surface Roughness Index (SRI) which is an orientation independent surface parameter of a protein molecule[12]. The surface parameter SRI served as representative of surface roughness at eight different surface locations of the concerned protein, whereas different surface locations were interpreted as octants of the internal orientation invariant coordinate system (ICS) of the concerned protein. Surface roughness index was measured as a set of eight indices whereas each index was measured as standard deviation of the distances of residues residing at each of the octants, from the ICS origin. Thus SRI of a protein can be represented as $S = \{S_i\}_{i=1}^{N}$ where

$$S_i = \sqrt{\frac{1}{M} \sum_{j=1}^{M} \{x_j - \bar{x}\}^2}$$

and $\{x_j\}$ is the set of distances between origin of ICS and M number of residues within ith octant and $\bar{x}$ is the mean of $\{x_j\}$. Application potential of SRI was shown by Singha et al., (2006) to predict SCOP class of protein especially the Fold of protein using machine learning approach[16].

The example of these works predicate the fact that carefully designed surface parameters of proteins can be utilized to extract and identify surface properties of proteins (an ambitious example is the functional site of protein). Outputs of these works also predicate the need of derivation of these surface parameters from a simple method, rather than to utilize the existing rigorous structure evaluating methods. We think following recent examples of researches give important requisites to start such an investigation.

In another interesting set of research reports the specificity of protein surface-active phenomenon to individual proteins was shown. Recently Lahiri et al., showed that an protein aggregate image (ordinary microscopic) based parameter, Intensity Level based Multi-Fractal Dimension (ILMFD) can be utilized to discriminate individual proteins[17,18]. In this study universally applicable aggregation method of heat denaturation was used to produce protein aggregates. Phase contrast microscopic images of heat denatured protein aggregates (HDPA) were processed and analyzed to derive the ILMFD. In the subsequent step ILMFDs were further used to differentiate the proteins selected for the study by a neuro-GA classifier which was customized for this purpose only.

The aim of the present study however stemmed from the complexity and difficulty in finding protein surface property without using its evaluated structure. As discussed above, considering the limitations of current techniques available for evaluation of protein structure, it appears to be difficult to get information about surface properties especially active site without its structure data. In this direction report of several recent studies can be listed which are related to development of novel and reliable approaches for identification of active site on protein surface[19,20]. Among these the most significant publications are regarding relation of surface roughness to small molecule binding sites of proteins[1,15,21]. Examples of these works point out to the fact that accurate prediction of SRI by a simple and fast method is necessary to further utilize it to explore protein surface properties. We put our effort exactly in this direction, where we have utilized protein aggregate level parameter ILMFD to map it to its individual level parameter, SRI.

Result

The aim of the study taken up in this work is to find alternative means to extract information on structural properties of protein whereas i) it can avoid methodological complexity and inapplicability of x-ray crystallography and NMR (however accurate it can be), as well as ii) it can overcome the problem of inaccuracy of mere theoretical predictive exercises by, systematically incorporating simple experimental clues generated at different layers or stages of the whole methodological process. Whereas, our methodology showed a general approach to solve complex biological problem, we have taken up the issue of a part of protein structure problem (i.e., to find surface property of protein) as a case study to test the efficiency of the given semi-empirical design. Thus the result section showcased both the simple experimental and computational predictive outputs and also the information extracted through the combination of them.

Result of the calculated values of SRI for proteins of our study

Table 1 shows the SRI values calculated from the PDB structure i.e., from coordinate information of corresponding protein. Table 1 also shows high similarity of SRI values between the proteins cytochrome c and ferritin, and high difference of SRI values of albumin from both of these proteins.

In table 2 the difference of SRI values between pairs of proteins were quantified by adopting calculation similar to ME calculations. However, table 2 also represented the high similarity in SRI values between cytochrome c and ferritin.

Result of mapping efficiency

Our recurrent network was built to utilize it as a mapping function. It showed average mapping efficiency MEF of data-set as 88.879% as shown in table 8 with its comparison with the improved efficiency obtained by clustering of decisions.

Result of grouping of decision observed through different intervals of frequency histogram of ME

In the frequency histogram of ME of each protein majority of mapped SRI (SRI^{pred}) outputs generated through the recurrent network were found to be grouped (on the basis of their corresponding mapping error) in first interval (first bin) having highest frequency and minimum ME as shown in table 3. The result of grouping of decisions is obtained indirectly from their correspondence to their associated mapping error.

Result of clustering of decision by k-means clustering

K-means clustering gives direct results of clustering of decision that is comparable to grouping of decision by frequency histogram of ME. For example, the centers of largest clusters were found to have minimum deviation from original SRI of corresponding proteins. Results of K-means clustering for all proteins are given in following tables.

Table 9 presented the matrix of MEs calculated for protein taking SRI of other proteins as expected SRI. This exercise was done to show the specificity of mapping. The diagonal elements of the matrix showed smallest mapping error and thus proved the efficacy of mapping protocol.

Discussion

The target of the study described in this pilot work is mainly to find out a fast and simple protocol to obtain broadly the structural component of protein and specifically its surface property, surface roughness index (SRI), by systematic incorporation of information generated from simple experiment or experiments. In this direction, we have designed a semi-empirical protocol and applied it with limited number of proteins. The basis of selection of materials and methodology adopted by us was discussed in the following paragraphs.

Selection of proteins of diverse functionality

All the proteins used are functionally specific. Albumin acts as binding protein for several substances like drugs in blood circulation. Hemoglobin functions in oxygen transport from lungs to all the body parts and in carbon dioxide transport from body tissues to lungs. Ferritin is a storage protein used for storing iron in the liver cells. Cytochrome c is enzymatic protein acting in various metabolic reactions in body.

Requirement of adopting single universal method to get protein aggregation

Although getting an aggregate of native protein appears to be the most suitable starting point for our experimentation, in practice, it is nearly impossible to get it by applying a single protocol that is universally applicable for all the existing proteins. On the other hand, we find that getting aggregated form of protein is easy if we consider its prior denaturation. Also, the fact that protein denaturation is strongly sensitive to a particular denaturing method[22], encouraged us to find one such denaturation protocol which is universally applicable to get aggregation of all types of proteins. Among many examples of protein denaturation methods which include denaturation by changing pH, salt concentration, heat, or adding urea or mercaptoethanol to the protein solution, we have selected heat denaturation method of aggregation that could be used for all the proteins in our experiment and other proteins in future[23].

Reason for using recurrent backpropagation network

Elman network, which is a type of recurrent network, was used for mapping ILMFD to SRI. Recurrent neural network was chosen because it has superiority over simple feed-forward neural networks in its capability of auto-association like human brain. Recurrent neural networks give better performance even in the presence of corrupted or incomplete data which was very much common in case of our data. Moreover Elman network can learn to recognize and generate both temporal and spatial patterns[24].

Reason for getting High efficiency in predicting SRI

In our layer based methodology, we actually tried to systematically build a hierarchical graph, whose nodes were either an experimental model or a computational model and whose edges were link parameter (or a set of parameters) that served as the output of the preceding node as well as input to the next node. Therefore in course of traversing from the starting node to the end node via many other nodes of the graph, chance of accumulation of error was very high due to additive effect. Therefore to make our layered model robust against the possible erroneous or noisy data, we introduced the concept of clustering of decisions, which was actually done by clustering of the mapped outputs (considered as output-decisions) of the recurrent network. While table 3 gives the histogram of ME indicating general trends of large amount of data grouped around smallest ME value, similar findings were found as result of clustering of decisions as shown in table 4 to 7. Table 8 further shows the extent of improvement of efficiency in mapping by decision-clustering in comparison to simple average technique.

Significance of clustering of decisions

As a solution we introduced the concept of clustering of many decisions obtained from multiple test data. While the largest cluster was considered and tested to obtain general tendency of decision, other smaller clusters were discarded as noises. Interestingly, the theoretical background of this protocol was also matching with the findings of Wallis and Bülthoff (2001)[25] on human cognition process involved in object recognition. Wallis and Bülthoff described that human recognizes an object correctly from its temporal description. While, technical translation of this concept immediately gives the idea of utilization of video data of an object for its recognition, we extracted the meaning of video as “multiple still images” or in general, “multiple data”. In our case, multiple input data were fed into our predictive model to obtain general tendency of the decision. Result obtained in Table 2 helped in strengthening our notion in initializing the decision-clustering protocol. Finally, table 3 to 7 showed the benefit of this protocol to enhance the mapping efficiency of SRI even for the test protein insulin.

Specificity of mapping

Our main target was to find the value of SRI for a protein for which no already evaluated structure was available. Although SRI is a surface-roughness profile of a protein and thus may not be the best discriminatory and unique property of a protein, in our pilot study, we tested the efficacy of our layered model in specifically mapping SRI values of the chosen proteins, some of which are having similar SRI values (example, Cytochrome c and Ferritin). For testing, whether the mapped output resulted after decision clustering for a particular protein is spurious, we calculated ME taking expected SRI of other proteins and found that it is giving least mapping error with the expected SRI of the same protein as shown in the table 9. The result confirms that our methodology can be used to find the SRI of a protein for which its already evaluated structure is not available.

Significance of the layered protocol adopted in this study

As mentioned in the introduction-section, surface active incidents like aggregation are specific to individual protein. Direct evaluation of important surface properties of protein, e.g., active site from any experimental method may be very difficult if not impossible for most of the cases. Therefore, our hypothesis was that a layered methodology might be a better choice to find protein functional site from simple experiments where

i) the first layer was designed with simple experiment (or assembly of such experiments in general) to generate output parameter (or a set of output parameters), say, OP_1, that could be considered as effective and important to predict complex biomolecular properties (e.g., protein surface property, SRI, in our case). As an example, in our study, a simple experiment based on i) heat denaturation of protein forming its aggregate and finally ii) visualization, acquisition and analysis of ordinary microscopic image of such aggregates, gave rise to the generation of parameter ILMFD.

ii) the second layer was designed with experiment or computational predictive model or a set of combination of many simple experiments and computational models, that would systematically incorporate the output of first layer to generate output of the second layer, say, OP2. At this layer we have utilized recurrent network as the predictive model to generate the mapped predicted surface parameter, SRI^{pred}.

iii) formation of methodological layers may be continued following similar principle, although in our study, we used two layers only.

We have already shown that further incorporation of SRI^{pred} generated in the last (i.e., second) layer of our methodology may be useful to identify more specific properties of protein, like, protein-SCOP-family and active sites[16]. In that case we should consider the methodology adopted in this work as the addition of another layer to give the desired result.

We may summarize findings of our study in the following manner. Current protocol of finding information about individual properties of protein (e.g., surface properties giving functional or active site) requires pre evaluated 3D structure of protein molecule. As the prerequisite of evaluated protein structure can not be fulfilled utilizing currently available methods like NMR and crystallography for most of the proteins of known sequences, there was a requirement of an alternative approach which could bypass the evaluation of protein structure for derivation of its individual information. As aggregates of proteins were shown as specific to individual proteins by many researchers, we considered that there might be a scope of finding an approach to utilize the specificity of aggregates to find individual information about protein. We started our exercise with the fact that, protein aggregation being a surface mediated phenomenon, it can be used to derive surface related information of its smallest subunit i.e., individual protein. Result of our work showed that intensity level based multifractal dimension of microscopic images of protein aggregates were specific to their surface property, surface roughness index. The novelty of this approach is to give a layer wise semi-empirical protocol to find solution of a complex biological problem where each layer was designed as a combination of simple experiments and computational models and, output of a preceding layer can be incorporated systematically to its next layer to yield the final solution. The success of this work eventually invoked the possibility of designing a very fast layer based divide and conquer strategy to address the complexity associated to the problem of derivation of protein structural property.

Acknowledgements

We are thankful to MHRD, Govt. of India for approval of grant (No.F.26-14/2003.T.S.V.Dated 14.01.2004) for this work.

Author contributions

Hrishikesh Mishra is responsible for the experiment on protein aggregates, its microscopic study, application of recurrent network and overall application of other computational protocols. Tapobrata Lahiri is responsible for the overall idea and design of experiments and computational protocols and analysis of results.

Figure legends

Figure 1: Architecture of recurrent neural network used

Tables

Table 1: SRI values for proteins in our study

Proteins	Octant							
	1	2	3	4	5	6	7	8
Albumin	15.1	16.075	22.405	19.448	18.138	9.480	15.043	7.996
Cytochrome c	5.379	5.830	4.702	2.545	5.112	4.191	3.401	3.699
Ferritin	4.123	5.229	4.991	3.079	6.327	3.139	3.655	3.227
Hemoglobin	5.951	8.818	8.593	10.985	9.285	4.963	8.706	8.363

Table 2: Difference of SRI values between all possible pairs of proteins

Albumin	-			
Cytochrome c	0.950	-		
Ferritin	0.975	0.183	-	
Hemoglobin	0.602	0.623	0.628	-
	Albumin	Cytochrome c	Ferritin	Hemoglobin

Table 3: Grouping of outputs at each interval of frequency histogram generated by data of protein albumin

Bin Serial Number	Bin center	Number of outputs clustered
1	0.023	34
2	0.044	8
3	0.064	2
4	0.085	2
5	0.105	1
6	0.126	2
7	0.146	0
8	0.167	0
9	0.187	0
10	0.208	1

Table 4: Result of k-means clustering for albumin

Cluster Serial number	Number of vectors clustered	Deviations of cluster-centers from original SRI values
1	3	0.150
2	5	0.077
3	11	0.036
4	31	0.015

Table 5: Result of k-means clustering for cytochrome c

Cluster Serial number	Number of vectors clustered	Deviations of cluster-centers from original SRI values
1	2	0.678
2	18	0.078
3	25	0.069
4	5	0.296

Table 6: Result of k-means clustering for ferritin

Cluster Serial number	Number of vectors clustered	Deviations of cluster-centers from original SRI values
1	15	0.142
2	18	0.029
3	12	0.215
4	5	0.459

Table 7: Result of k-means clustering for hemoglobin

Cluster Serial number	Number of vectors clustered	Deviations of cluster-centers from original SRI values
1	10	0.207
2	15	0.062
3	21	0.049
4	4	0.161

Table 8: Mapping error (ME) and mapping efficiency for proteins in our study calculated from “simple average” and “largest cluster centre” of the mapped output

Protein	Average		Calculated from largest cluster centre	
	ME %	MEF%	ME %	MEF%
Albumin	3.472	96.528	1.477	98.523
Cytochrome c	14.587	85.413	6.922	93.077
Ferritin	16.355	83.645	2.868	97.132
Hemoglobin	10.072	89.928	4.942	95.058

Table 9: ME values calculated for proteins considering calculated SRI of other proteins(including the proteins of study) as expected SRI.

Expected SRI taken from **ME calculated for**	Albumin (original)	Cytochrome c (original)	Ferritin (original)	Hemoglobin (original)
Albumin	0.015	0.952	0.976	0.600
Cytochrome c	0.924	0.069	0.179	0.582
Ferritin	0.966	0.183	0.029	0.614
Hemoglobin	0.571	0.638	0.646	0.049

Methods

Experimental

Proteins and other materials used in study

The proteins of some well-known pathophysiological significance have been selected for our study because the broader objective of our work is to develop a tool for faster derivation of their functional sites. For our study, we used five proteins, albumin, haemoglobin, ferritin, and cytochrome c. The proteins were obtained from Sigma Aldrich (USA) and were of analytical grade. Milipore water was used as a medium to dissolve the proteins to make solution of proteins.

Getting aggregates of proteins

Each protein was suspended in Millipore water at concentration of 50 mg/cc and was kept in hot water bath at 100°C for 15 minutes to get its Heat Denatured Protein Aggregates (HDPAs).

Visualization of aggregates through phase contrast Microscopy and Creation of Aggregate Image Dataset

Homogeneously distributed HDPAs were visualized at 400× magnification under phase contrast mode of Leica Microscope Model DML-B2. We used hemocytometer slides (Model: Neubauer Chamber, Marienfeld, Germany) which had sample mounting area of depth 0.1 mm so that the aggregates were less likely to be deformed. Slides were covered with a thin microscopic glass cover slip and put under the microscope to collect digital images of aggregates using a camera (Canon PowerShot S50) attached with the microscope. Optical zoom of camera was adjusted to 2×. Thus total optical zoom combining the microscope and the camera was 800×. 50 images of HDPAs at different fields of views were taken for each protein. Thus a dataset Comprising 200 images of protein aggregates was created and kept for further analysis.

Computational

Preprocessing of aggregate images

Each image of original size in pixel 2592x1944 was converted to grey scale and resized to 1/3rd of the original size. Aggregate part was segmented out from each resized gray scale image using Adobe Photoshop 7.0 keeping the background pixel intensity as zero.

Dividing the images into intensity plane slices

Each segmented image of aggregates was divided into ten equally spaced intensity levels considering the intensity range of the image from zero to maximum intensity (255) following the protocol given by Singh et al[26]. Each preprocessed image, **I** was sliced into 10 binary images from fixed intensity-intervals by the following protocol:

$\boldsymbol{B_i} = 1 \text{ for } \boldsymbol{s_i} \leq I < \boldsymbol{s_{i+1}}, \text{ and}$

$B_i = 0, \text{ otherwise} \qquad \text{for } \mathbf{i} = 1 \text{ to } 11$

where

$s_i = 0$ for $i = 1$, and

$s_i = (m/10)(i-1) + 1$, for $i > 1$ where m is the maximum intensity of the image, I.

and,

B_1 is referred as binary image form of 1st bit plane,

B_2 is referred as image of 2nd bit plane,

and so on up to 10th bit plane.

Calculation of ILMFD for aggregate images

Fractal dimensions were measured for each of the 10 binary images using box counting method[27,28]. Thus each aggregate image was represented by a set of 10 fractal dimensions, D whereas, each fractal dimension corresponds to one intensity level:

$$D = \{D_i\}_{i=0}^{10}$$

Calculation of SRI for proteins

SRI was calculated following the protocol given by Varadwaj et al[12]. First PDB files corresponding to individual proteins were downloaded from Protein data bank (PDB). Next, all PDB file coordinates were converted to orientation invariant coordinate system (ICS) to put all proteins in similar coordinate space. Surface of each protein was divided into eight octants and surface roughness index was measured as set of eight standard deviations of distances of residues in each octant calculated from ICS-origin.

Application of Recurrent neural network for predicting SRI through mapping ILMFD to SRI

The ILMFD data obtained from four proteins, comprised of 50 images for each protein (albumin, cytochrome c, ferritin and hemoglobin). ILMFD and SRI values of these proteins were used as input and target output respectively for mapping. ILMFD data was normalized by subtracting their mean from them.

For the purpose of mapping the output data (i.e., SRI data having a set of 8 surface roughness indices) the target was scaled to the range 0 to 1 by dividing each of them by their corresponding index-maximum. 8 such maxima thus constitute the SRI_{max}.

For mapping ILMFD to SRI, we used Elman network, which is a backpropagation network having a feedback connection from the output of hidden layer to its input with delay of one time step. The network architecture used in our work comprised of three layers viz., input, hidden and output layer comprising 10, 12 and 8 neurons respectively (Fig. 1). Hidden layer was the recurrent layer. Transfer functions in the hidden layer and output layer were tan sigmoid and log sigmoid respectively. Mean square error was used as a performance function.

Assessment of average efficiency of mapping ILMFD into SRI for a particular protein

We calculated the efficiency of mapping of ILMFD to SRI for a particular protein, p for each j-th data of this protein, by using the following steps:

Step 1) first we calculated the predicted SRI for j-th data of this protein, as

$\mathrm{SRI}_j^{\mathrm{pred}} = \mathrm{output}_j \times \mathrm{SRI}_{\max}$

Step 2) next we calculated the mapping error (ME) for j-th data of protein, p by taking mean of the squared deviation of $\mathrm{SRI}_j^{\mathrm{pred}}$ from its corresponding target SRI_j as:

$$ME_j^p = \sqrt{\frac{\displaystyle\sum_{i=1}^{8} \frac{\left(net\_o_{ji} - t\arg\_o_{ji}\right)^2}{(net\_o_{ji}{}^2 + t\arg\_o_{ji}{}^2)/2}}{8}}$$

Where net_o_{ji} is i-th element of the rescaled output, $\mathrm{SRI}_j^{\mathrm{pred}}$ and $targ\_o_{ji}$ is the i-th index of the corresponding actual SRI_j.

Naturally for all 50 ILMFD data of the protein, p the mean mapping error can be calculated as:

$\mathrm{ME}^p = <\{\mathrm{ME}_j{}^p\}_{j=1}{}^{50}>$

Finally to assess the mapping efficiency (MEF) of prediction of SRI for a particular p-th protein we used the formula as:

$\mathrm{MEF}^p = (1 - \mathrm{ME}^p)\times 100$

Overall efficiency of the network was however calculated as the average of the above efficiency taken over the data of all 4 proteins,

$\mathrm{MEF} = \langle \mathrm{MEF}^{p} \rangle$

Assessment of efficiency of mapping ILMFD into SRI for a particular protein by clustering of decision protocol

For a particular protein, in order to screen out the general tendency of mapping decision from that generated through noisy input feature, we have adopted two methods:

1) calculation of statistical mode of decisions using 10 interval frequency histogram of mapping errors ranging from least to maximum mapping error and finally choosing the decision for which the mapping error is closest to the mid point of the highest-frequency-interval. This exercise has been done in order to visualize whether there is any major tendency of decisions as well as to identify presence of "decisions generated through noisy input feature" in other bins.
2) Clustering of decisions (i.e., neural network outputs) methodology has been adapted from method 1, to get the general tendency of decision for data where primarily we should not take the help of expected decision and thus measurement of mapping error is not possible. In this direction, we have applied k-means clustering fixing cluster number, k as 4 after certain trials.

References

1. Pettit, F.K., Bare, E., Tsai, A. & Bowie, J.U. HotPatch: a statistical approach to finding biologically relevant features on protein surfaces. *J. Mol. Biol.* **369**, 863-879 (2007).
2. Fernández, C. & Wider, G. TROSY in NMR studies of the structure and function of large biological macromolecules. *Curr. Opin. Struct. Biol.* **13**, 570-580 (2003).
3. Smyth, M.S. & Martin, J.H. x Ray crystallography. *Mol. Pathol*. **53**, 8-14 (2000).
4. Qu, X., Swanson, R., Day, R. & Tsai J. A guide to template based structure prediction. *Curr. Protein Pept. Sci.* **10**, 270-285 (2009).

5. Singh, S. & Lahiri, T. Study on comparative model strategies for batter prediction of protein structure. Thesis in M.Tech. (Information Technology & specialization in Bioinformatics) accepted at Indian Institute of Information Technology, Allahabad, India, 43 (2009).
6. Baker, M.L., Ju, T. & Chiu, W. Identification of secondary structure elements in intermediate-resolution density maps. *Structure* **15**, 7-19 (2007).
7. Zhou, Z.H. Towards atomic resolution structural determination by single-particle cryo-electron microscopy. *Curr. Opin. Struct. Biol.* **18**, 218-228 (2008).
8. Shibata-Seki, T., Masai, J., Ogawa, Y., Sato, K. & Yanagawa, H. Application of atomic force microscopy to protein anatomy: Imaging of supramolecular structures of self-assemblies formed from synthetic peptides. *Appl. Phys. A* **66**, S625–S629 (1998).
9. Giocondi, M.C., Milhiet, P.E., Lesniewska. E., Le, G.C. Atomic force microscopy: from cellular imaging to molecular manipulation. *Med. Sci.(Paris)* **19**, 92-99 (2003).
10. Topf, M. & Sali, A. Combining electron microscopy and comparative protein structure modeling. *Curr. Opin. Struct. Biol.* **15**, 578-585 (2005).
11. Mistry, J., Bateman, A. & Finn, R.D. Predicting active site residue annotations in the Pfam database. BMC Bioinformatics. 8, 298-311 (2007).
12. Varadwaj, P.K., Lahiri, T. & Tsodikov, O. Surface roughness index a novel approach to compare protein surfaces*. Proc. ICISIP 2005.* 474-478 (2005).
13. Tateno, M., Yamasaki, K., Amano, N., Kakinuma, J., Koike, H., Allen M.D. & Suzuki M. DNA recognition by beta-sheets. *Biopolymers* **44**, 335-359 (1997).
14. Via, A., Ferre, F., Brannetti, B. & Helmer-Citterich, M. Protein surface similarities: a survey of methods to describe and compare protein surfaces. *Cell. Mol. Life Sci.* **57**, 1970-1977 (2000).
15. Pettit, F.K. & Bowie, J.U. Protein surface roughness and small molecular binding sites. *J.Mol.Biol.* **285**, 1377-1382 (1999).
16. Singha, S., Lahiri, T., Dasgupta, A.K. & Chakrabarti, P. Structural classification of proteins using surface roughness index. *Online J. Bioinformatics* **7**, 74-84 (2006).
17. Lahiri, T., Mishra, H., Sarkar, S. & Misra, K. Surface characterization of proteins using Multifractal property of heat-denatured aggregates. *Bioinformation* **2**, 379-383 (2008).
18. Lahiri, T., Mishra, H., Kumar, U. & Misra, K. Derivation of a protein-marker from heat-denatured protein-aggregate. *Online J. Bioinformatics* **10**, 29-39 (2009).
19. Zhang, B., Rychlewski, L., Pawłowski, K., Fetrow, J.S., Skolnick, J. & Godzik, A. From fold predictions to function predictions: Automation of functional site conservation analysis for functional genome predictions. *Protein Sci.* **8**, 1104-1115 (1999).
20. Slama, P., Filippis, I. & Lappe, M. Detection of protein catalytic residues at high precision using local network properties. *BMC Bioinformatics* **9**, 517-529 (2008).
21. Kinoshita, K. & Nakamura, H. Identification of the ligand binding sites on the molecular surface of proteins. Protein Sci. **14**, 711-718 (2005).

22. Vermeer, A.W. & Norde, W. The thermal stability of immunoglobulin: unfolding and aggregation of a multi-domain protein. *Biophys. J.* **78**, 394-404 (2000).
23. Weijers, M., Barneveld, P.A., Stuart, M.A.C. & Visschers, R.W. Heat-induced denaturation and aggregation of ovalbumin at neutral pH described by irreversible first-order kinetics. *Protein Sci.* **12**, 2693-2703 (2003).
24. Elman, J.L. Finding Structure in Time. *Cogn. Sci.* **14**, 179-211 (1990).
25. Wallis, G.M. & Bülthoff, H.H. Effects of temporal association on recognition memory. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4800-4804 (2001).
26. Singh R., Samal, S. & Lahiri T., A Novel Strategy for Designing Efficient Multiple Classifier. *Lect. Notes Comput. Sc.* **3832**, 713 – 720 (2005).
27. Kawaguchi, E. & Taniguchi, R. The depth first picture-expression as an image thresholding strategy. *IEEE T. Syst. Man Cy.* **19**, 1321-1328 (1989).
28. Zmeškal, O., Veselý, M., Nežádal, M. & Buchníček, M. Fractal Analysis of Image Structures. *HarFA e-journal* 3-5 (2001).

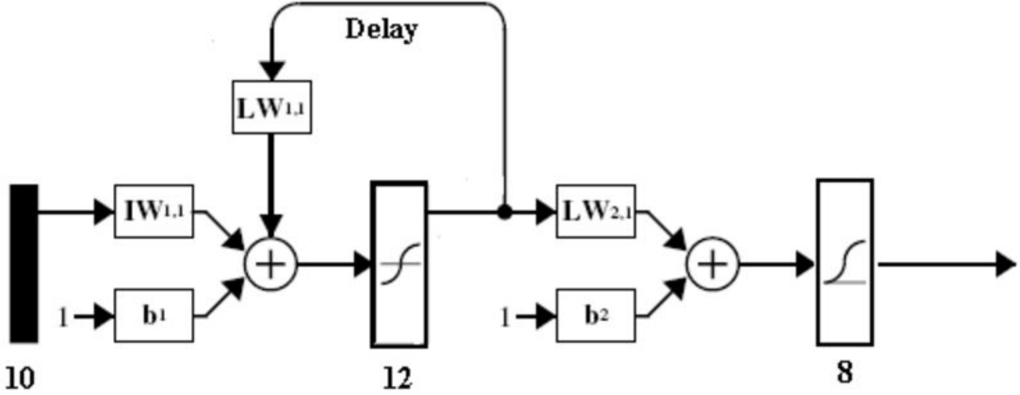

Figure 1 (Lahiri)