BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net **Hypothesis**

Volume 6(4)

# Neurocognitive derivation of protein surface property from protein aggregate parameters

**Hrishikesh Mishra, Tapobrata Lahiri***

Division of Applied Science and Indo-Russian Center for Biotechnology, Indian Institute of Information Technology, Allahabad, India; Tapobrata Lahiri - Email: tlahiri@iiita.ac.in; Phone: 91-532-2922242; Fax: 91-532-2430006; *Corresponding author

Received February 13, 2011; Accepted February 17, 2011; Published May 07, 2011

**Abstract:**
Current work targeted to predicate parametric relationship between aggregate and individual property of a protein. In this approach, we considered individual property of a protein as its Surface Roughness Index (SRI) which was shown to have potential to classify SCOP protein families. The bulk property was however considered as Intensity Level based Multi-fractal Dimension (ILMFD) of ordinary microscopic images of heat denatured protein aggregates which was known to have potential to serve as protein marker. The protocol used multiple ILMFD inputs obtained for a protein to produce a set of mapped outputs as possible SRI candidates. The outputs were further clustered and largest cluster centre after normalization was found to be a close approximation of expected SRI that was calculated from known PDB structure. The outcome showed that faster derivation of individual protein's surface property might be possible using its bulk form, heat denatured aggregates.

## Background:

Protein aggregation has been considered as an unwanted and unproductive phenomenon in biological applications involving proteins **[1]**. It can be defined as a process by which a homogeneous protein solution separates into two phases comprising aggregate phase having significant intermolecular interactions and the other one having dilute supernatant of isolated protein **[2]**. Like crystallization, aggregation is governed by details of structure of protein and chemical and physical conditions of environment i.e., protein solution **[3]**. Aggregates may be formed by various mechanisms and may be classified in several manners into soluble/insoluble, native/denatured, covalent/noncovalent and reversible/irreversible etc **[4]**. Aggregation is generally accompanied by conformational change of protein, which can be induced by thermal, enzymatic or chemical perturbations affecting the native folded structure of protein **[5]**.

Several studies have been done which show specificity of aggregates. Study done by Bohr et al., indicates that given some conditions, aggregates are strongly affected by shape of each protein molecule **[6]**. As per another study protein monomer surface characteristics profoundly affect the structure and morphology of protein aggregates. Physicochemical nature of protein surfaces including distribution of hydrophobic and hydrophilic sites on the monomer surface controls the organization of aggregates **[7]**. Some evidences regarding specificity of aggregates, show that a minor change in amino acid sequence of protein can prevent or increase aggregation of protein. In a study done on viral coat proteins, King et al., found that mutant viral coat protein having a single amino acid change, folded at low temperatures normally, but at higher temperatures it self-assembled into aggregates. This aggregation at high temperatures was not found in normal protein **[8]**. Another study done by David Brems et al., on bovine growth hormone, showed that mutation prevented its aggregation but did not affect its folding **[9]**. These studies gave rise to concept that aggregation may also be preprogrammed into amino acid sequence just like folding and aggregates should not be considered as just a nonspecific mess **[8]**. We started our work considering that, as protein folding and aggregation both are linked to amino acid sequence, pattern of protein folding and aggregation should be considerably specific to concerned protein. In our previous studies, we showed that multifractal property of heat denatured aggregates 'Intensity level based Multi-Fractal Dimension' (ILMFD) of different proteins differ from each other and can be used to discriminate them **[10, 11]**. In this work the scope of ILMFD has been extended by including four different fractals viz., perimeter fractal dimension giving $\text{ILMFD}_P$, perimeter-area relationship giving $\text{ILMFD}_{PAR}$, area fractal dimension giving $\text{ILMFD}_A$, and, perimeter-area fractal dimension defining $\text{ILMFD}_{PA}$.

Taking cue from the published works on nature of aggregates showing specificity to their seed proteins, in this work we have put our effort to investigate whether multifractal property, i.e., ILMFD features derived from microscopic images of heat denatured protein aggregates (HDPA), are linked to the surface geometrical property of its seed protein represented by surface roughness index (SRI). We utilized recurrent backpropagation network to generate mapping function to derive SRI as a function of ILMFD with subsequent clustering of its outputs. Finally the largest cluster center of the outputs was utilized to derive SRI. Accuracies of methods utilizing different ILMFD features were compared. The result of our approach indicated that the estimated function can be utilized as an experiment support system to derive information about surface property of proteins for which no folded structure is available yet.

## Methodology:

### Preparation of Heat Denatured Protein Aggregates:

Proteins viz., albumin, cytochrome c, ferritin hemoglobin, insulin and lysozyme were obtained from Sigma Aldrich (USA). Protein solutions were prepared in milipore water at concentration of 25 mg/cc and put in hot water bath having temperature 100°C for 15 minutes to obtain Heat Denatured Protein Aggregates (HDPAs).

### Microscopic Visualization of Aggregates and Image Collection:

Suspension of HDPAs kept at hemocytometer slides (Model: Neubauer Chamber, Marienfeld, Germany) and covered with thin microscopic glass cover slip, was visualized at 400X magnification using phase contrast microscope (Leica Model DML-B2). Digital images of aggregates were captured using a camera (Canon PowerShot S50) at optical zoom 2X. Thus cumulative optical zoom of the microscope and camera was 800X. 50 images of different HDPAs were captured for each protein.

### Image Processing:

Each aggregate image converted to grey scale and was resized from original size of 2592x1944 pixels was resized to 1/3rd of the original size to ease further processing. Portion of image having aggregate was segmented out from each image making the background pixel intensity zero. Each segmented image having intensity range from 0 to 255, was splitted into 10 binary images on the basis of fixed intensity-ranges by applying the rule described by Singh *et al.* (2005) **[12]**.

### Intensity level based Multi-Fractal Dimension for Aggregate Images:

Four types of ILMFD features were calculated for each of the aggregate image using box counting method **[13, 14]**. Areas (A), perimeter (P) of aggregates were measured at different box sizes (S) of pixel unit for 10 binary images representing each aggregate image. Area was measured as the number of boxes required to cover the aggregate part of image. Similarly, perimeter of aggregate was measured as the number of boxes making the periphery of the aggregate part of segmented aggregate images. Perimeter fractal dimension was measured as the slope of the linear regression plot (LRP) between log (P) and log(S), at different scales (box size), for 10 binary images derived from segmented aggregate image. Perimeter-area relationship was calculated as slope of LRP between log(P) and log(A) at different scales. Area fractal dimension was calculated as the slope of LRP between log(A) and log(S), considering different scales i.e., box sizes. Similarly perimeter-area fractal dimension was calculated as slope of LRP between two variables derived from P, A, and S as x= log(P/S), and y=(log(A))/2 - log(S). Thus each aggregate image was represented by 10 fractal dimensions, cumulatively referred to as ILMFD **(see Supplementary material)**. As we had measured four different types of fractal dimensions for each binary image, four different types of ILMFD parameters were obtained as $ILMFD_A$, $ILMFD_P$, $ILMFD_{PA}$, and $ILMFD_{PAR}$ from area fractal dimension, perimeter fractal dimension, perimeter area fractal dimension and perimeter area relationship respectively.

### Computing SRI for proteins:

Surface property of folded native proteins was represented by surface roughness index (SRI) which was computed following published protocol **[15]**. PDB files representing native folded structure of proteins were procured from Protein data bank (PDB). PDB coordinates were changed into orientation invariant coordinate system (ICS) to represent molecules in uniform manner. First the centre of gravity (CG) was calculated by taking the mean of coordinates of the atoms. Calculated CG was taken as the origin O of ICS. The point on molecule surface that was placed at maximum Euclidian distance from O was taken as Z point and the line connecting O and Z was taken as Z axis. For fixing X and Y planes, a slice of thickness ΔZ, chosen by trial and error, was considered. The point at maximum distance in the slice was fixed as T point and the line joining O and T was considered as X axis. The Y axis was deduced as the axis perpendicular to both the X and Z axis. After calculating x, y, and z axes of ICS, the Cartesian coordinates of proteins were transformed to invariant coordinates. Finally surface of each protein was divided into eight octants and standard deviations of distances of $C_\alpha$ atoms of surface residues in each octant from ICS origin were calculated. This set of eight standard deviations constituted SRI for concerned protein as shown in **Table 1(see Supplementary material)**.

### Mapping of ILMFD to SRI using Neurocognitive Protocol:

We used a protocol employing a recurrent backpropagation neural network (RBPN) supported by a two tier clustering of decisions obtained through RBPN to predict SRI for 4 ILMFD types. Each of the four ILMFD features was used separately for prediction of SRI of constituent proteins, up to first tier clustering of decisions. Decisions obtained from different ILMFD features through first tier of clustering, were combined (details given in latter section) to get a combined mapping efficiency of all the four ILMFD features. The clustering method was however adopted from the work of Wallis and Bülthoff **[16]**, who attributed human cognitive success to its capability to combine decisions obtained from temporal inputs. In our work we have converted the idea of temporal inputs into input ensemble or multiple inputs, assuming inputs followed Ergodic hypothesis **[17]**.

The ILMFD data obtained from six proteins, comprised of 50 images for each protein. ILMFD data for 35 images of each protein was used as training input. Data for remaining 15 images for each protein was used as test set. SRI values of these proteins were used as target output for training. Training ILMFD data was normalized by subtracting their respective column 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 initial mapping ILMFD to SRI, we used Elman network, which is a recurrent backpropagation network (RBPN) 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. 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. Trained Elman network was simulated with test data normalized using mean derived from training input data. Outputs of network for test data were reverted from 0-1 range to original range by multiplying with corresponding index-maximum. To improve the results of RBPN, SRI values mapped for each protein using different ILMFD features were clustered separately to evaluate the general tendency of the mapping decisions. For this purpose hierarchical clustering was applied and hierarchy was chosen as 4 after certain trials. Center of the largest cluster was considered as the output decision of first tier of clustering. Then largest clusters obtained for decisions from each of the 4 ILMFD features were combined and again clustered using hierarchy level as 2. Finally the predicted SRI was calculated as mean and median of decisions clustered in the larger cluster.

### Computing Efficiency of Mapping of ILMFD to SRI:

First, output of RBPN was converted to mapped SRI (*SRI map*$_i$) for any i-th ILMFD data by multiplying it with $SRI_{max}$. Mapping deviation (i.e., the error in mapping) for i-th data of each protein was calculated as given in **Supplementary material**.

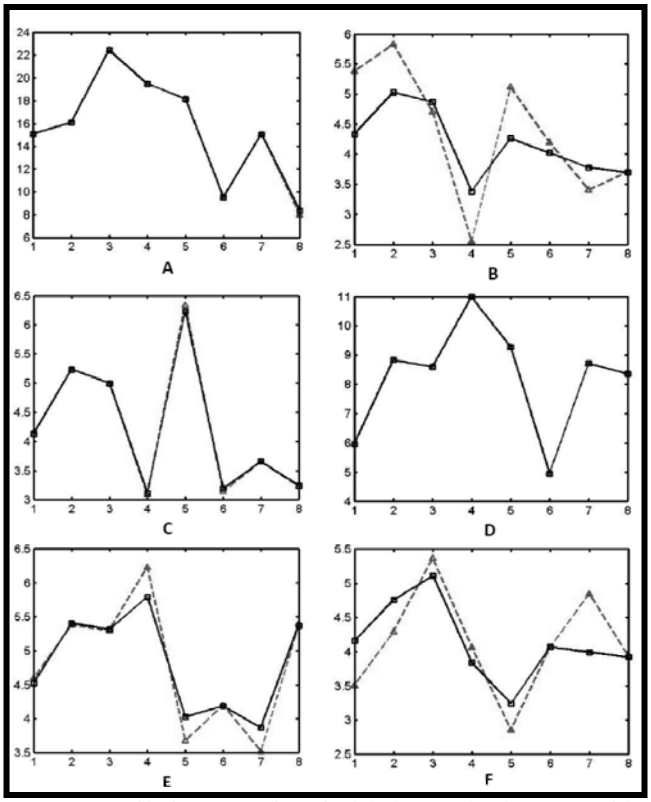

**Figure 1:** Graphical representation of original SRI (triangles connected by dotted lines) and predicted SRI (squares connected by solid lines) values obtained from cluster median after application of second tier of clustering. Plots A-F represent proteins albumin, cytochrome c, ferritin, hemoglobin, insulin, and lysozyme respectively.

**Discussion:**

**Mapping Efficiency of RBPN and clustering of decisions protocol:**

Initially RBPN was utilized for mapping ILMFD into SRI. Its average mapping efficiency for training set was found to be 95.93%, 88.57%, 90.18%, and 98.08% for $ILMFD_P$, $ILMFD_A$, $ILMFD_{PA}$, and $ILMFD_{PAR}$ respectively. Table 1 shows a comparative profile of mapping efficiencies by RBPN and first tier of clustering for test set **(Table 1)**. **Table 2** shows mapping efficiency by application of second tier of clustering. The result of application of second tier of clustering shows improvement in mapping efficiency as compared to that obtained without clustering on test set **(Table 2 see Supplementary material)**. For further validation, target SRI and mapped SRI values obtained at second tier of clustering were compared graphically **(Figure 1)**.

**Potential applicability of ILMFD based protocol:**

Current methods for deriving proteins' surface proper-ties predominantly rely on its evaluated structure. Structure evaluation by X-Ray crystallography, NMR methods and all other methods including comparative model based methods has typical shortcoming of nearly 10% applicability among the proteins with known sequences. Therefore for rest of the proteins having no known and reliable structural model, it still remained a difficult task if not impossible to derive information about its structural properties. In this situation, we propose an Experiment Support System (ESS) to derive crucial information from simple but universally applicable experiments which were not apparent from the firsthand outputs of these experiments. We made the target as a protein surface geometric property, SRI and, first hand output, ILMFD of experiment on microscopic image of HDPA as input-feed to the ESS template **[10, 11, 15]**. Since importance of rough part of a protein surface to attach small molecules was already known, we have chosen SRI as our target parameter **[18]**. This is further justified because it was already reported that SRI represented scale and rotation independent profile of protein surface roughness and was capable to predict SCOP family of protein **[15]**.

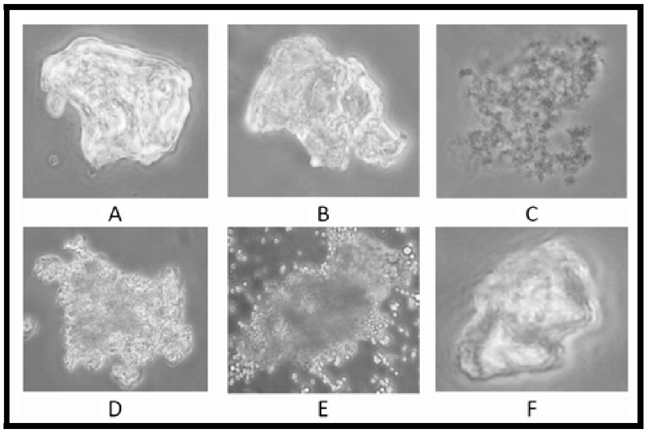

**Figure 2:** Representative HDPA images of A) albumin, B) cytochrome c, C) ferritin, D) hemoglobin, E) insulin, and F) Lysozyme

**Aggregates as starting point of protocol:**

The question of using aggregates as the starting point of our ESS can be answered considering, first, one of the factors governing protein aggregation is protein surface and secondly, like ordered assembly of proteins (crystals), disordered and irregular aggregate patterns are also worth investigation by the commonly available tools, e.g., fractal dimensional analysis. For this reason, we have considered scale and rotation independent intensity level multifractal dimension of aggregate images obtained from standard and same experimental condition. As we have utilized hemocytometer slides under transmission light microscopy, the images formed could be considered as 2 dimensional projections of the aggregates as shown in **Figure 2**. Our basic assumption was that there was a statistically comprehensible homogenous spread of the aggregate mass within its microenvironment and therefore its intensity distribution would reflect the same within its image. Moreover the raw intensity profile could be non-specific in nature because of presence of various sizes of aggregates. To solve this problem, we have considered the scale dependence of aggregate mass represented by its corresponding intensities at its different levels by calculating fractal dimension at each level of intensity.

**Handling of noisy data with clustering of decisions:**

There was presence of noisy data due to possible accumulation of errors at each level of the overall experimental steps however simple it is, and its manifestation to form and give raw images of aggregates. Moreover, since both of the parameters ILMFD and SRI were extracted from irregular or rough objects, success in getting a robust approximation of their derived values was difficult to obtain. This further pointed out presence of ambiguous or confusing data especially for ILMFD and also showed difficulty involved in designing a function for one to one mapping from ILMFD to SRI. To rule out the contribution of these noisy data in the overall mapping process we introduced the concept of "use of multiple test data" to predict general tendency of the mapped values as obtained through the ESS-outputs based on RBPN. In this context, we assumed that there should be fewer occurrence of noisy data and therefore if we cluster the mapped outputs obtained from multiple test data, the smaller clusters should contain those noisy data. On the contrary, applying same logic we may expect that the largest cluster should hold the information of the general tendency of the larger section of the data which were assumed as predominantly set of correct data only. Although this argument points existence of only two such clusters but after several trials we found best result of mapping for 4 clusters. Probable explanation of more than 2 clusters might be due to presence of different amount of noise within the set of data, whereas largest cluster showed the data with least noise and the remaining smaller 3 clusters represented data of higher noise levels.

**Conclusion:**

In this work, we have shown that a simple experiment on heat denatured protein aggregation can be made more effective in drawing useful information on protein surface roughness with the help of Experiment Support System based on the computational strategy proposed by us. In this direction, our study shows the specificity of protein aggregate feature, intensity level based multi fractal property of heat denatured protein aggregates, to accurately predict Surface Roughness Index of the single seed protein without using its already evaluated structure. It also indicates that structural information of a protein may be conserved in its aggregates.

**Acknowledgement:**

We are thankful to MHRD, Govt. of India and ICMR, Govt. of India for approval of grants (No.F.26-14/2003.T.S.V.Dated 14.01.2004 and No. 52/8/2005-BMS, dated 4/2/10) for this work.

**References:**

**[1]** Okanojo M *et al. J Biosci Bioeng.* 2005 **100**: 556 [PMID: 16384796]
**[2]** Pappu RV *et al. Arch Biochem Biophys.* 2008 **469**: 132 [PMID: 17931593]
**[3]** Pullara F *et al. Biophys J.* 2007 **93**: 3271 [PMID: 17660322]
**[4]** Cromwell ME *et al. AAPS J.* 2006 **8**: E572 [PMID: 17025275]
**[5]** Weijers M *et al. Protein Sci.* 2003 **12**: 2693 [PMID: 14627731]
**[6]** Bohr H *et al. Z Phys D.* 1997 **40**: 513
**[7]** Patro SY & Przybycien TM. *Biophys J.* 1996 **70**: 2888 [PMID: 8744327]
**[8]** Taubes G. *Science* 1996 **271**: 1493 [PMID: 8599100]
**[9]** Brems DN *et al. Proc Natl Acad Sci USA.* 1988 **85**: 3367 [PMID: 3130626]
**[10]** Lahiri T *et al. Bioinformation* 2008 **2**: 379 [PMID: 18795110]
**[11]** Lahiri T *et al. Online J Bioinformatics.* 2009 **10**: 29
**[12]** Singh R *et al. Lect Notes Comput Sc.* 2005 **3832**: 713
**[13]** Kawaguchi E & Taniguchi R. *IEEE Trans Syst Man Cybern.* 1989 **19**: 1321
**[14]** Zmeškal O *et al. HarFA e-journal.* 2001 **1**: 3
**[15]** Singha S *et al. Online J Bioinformatics.* 2006 **7**: 74
**[16]** Wallis GM & Bülthoff HH. Proc Natl Acad Sci USA. 2001 **98**: 4800 [PMID: 11287633]
**[17]** Birkhoff GD. *Am Math Mon.* 1942 **49**: 222
**[18]** Pettit FK & Bowie JU. *J Mol Biol.* 1999 **285**: 1377 [PMID: 9917382]

**Edited by P Kangueane**
**Citation: Mishra & Lahiri.** Bioinformation 6(4): 158-161 (2011)

## Supplementary material:

**Intensity level based Multi-Fractal Dimension:**

$$D = \{D_i\}_{i=1}^{10}$$

Where $D_i$ is fractal dimension of one intensity level.

**Computing Efficiency of Mapping of ILMFD to SRI:**

$$Map\_Err_i = \sqrt{\frac{\sum_{j=1}^{8} \frac{(SRI\_map_{ij} - SRI_{ij})^2}{(SRI\_map_{ij}{}^2 + SRI_{ij}{}^2)/2}}{8}} \quad (1)$$

Where *SRI_map*$_{ij}$ is j-th element of the mapped SRI, *SRI_map*$_i$ and *SRI*$_{ij}$ is the j-th index of the corresponding i-th data for $SRI_i$. Thus for 50 data of each protein mean mapping deviation was computed as arithmatic mean of all the 50 Map_Err values for the protein concerned. Finally efficiency of mapping for a protein p was calculated as:

$$Map\_eff_p = (1 - Map\_err_p) \times 100 \quad (2)$$

Overall efficiency of RBPN was calculated as arithmetic mean of the $Map\_eff_p$ for all six proteins. Mapping error and efficiency were calculated for first and second tier of clustering also as deviation of cluster center from SRI of concerned protein following calculation of Map_Err as described as above.

**Table 1:** Comparative profile of mapping efficiencies by RBPN and first tier of clustering for test set

| Protein | Efficiency (in %) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $ILMFD_P$ | | $ILMFD_A$ | | $ILMFD_{PA}$ | | $ILMFD_{PAR}$ | |
| | RBPN | First tier clustering | RBPN | First tier clustering | RBPN | First tier clustering | RBPN | First tier clustering |
| Albumin | 84.15 | 98.02 | 75.6 | 98.4 | 84.42 | 98.4 | 70.74 | 98.38 |
| Cytochrome c | 54.76 | 93.1 | 40.51 | 80.79 | 46.98 | 78.55 | 45.97 | 75.47 |
| Ferritin | 97.25 | 98.54 | 80.72 | 87.86 | 71.8 | 90.61 | 87.12 | 99.89 |
| Hemoglobin | 86.4 | 99.63 | 91.14 | 84.22 | 89.66 | 96.62 | 94.1 | 99.7 |
| Insulin | 90.67 | 98.58 | 70.57 | 84.22 | 84.41 | 90.02 | 81.67 | 94.68 |
| Lysozyme | 58.3 | 84.47 | 28.89 | 85.42 | 68.32 | 83.34 | 29.55 | 85.4 |
| Complete test set | 78.59 | 95.39 | 64.57 | 86.96 | 74.26 | 89.59 | 68.19 | 92.25 |

**Table 2:** Mapping efficiency by application of second tier of clustering

| Protein | Efficiency (%) | |
|---|---|---|
| | Based on Cluster Mean | Based on Cluster Median |
| Albumin | 98.40 | 98.41 |
| Cytochrome c | 82.41 | 84.56 |
| Ferritin | 95.69 | 99.14 |
| Hemoglobin | 96.74 | 99.88 |
| Insulin | 93.39 | 94.68 |
| Lysozyme | 85.26 | 89.00 |
| Complete test set | 91.98 | 94.28 |