The IIOAB Journal

IIOAB ISSN: 0976-3104

RESEARCH: BIOINFORMATICS

MULTIFRACTAL ANALYSIS OF PROTEIN AGGREGATES TO DERIVE PROTEIN-SPECIFIC SIGNATURE

Hrishikesh Mishra, Tapobrata Lahiri*

Division of applied sciences and Indo-Russian center for biotechnology, Indian Institute of Information Technology, Allahabad–211012, INDIA

Received on: 3rd-Aug-2010; **Revised on:** 13th-Oct-2010; **Accepted on:** 19th-Oct-2010; **Published on:** 15th-Nov-2010.

***Corresponding author:** Email: tlahiri@iiita.ac.in Tel: +91-5322922229; Fax: +91-5322430006

ABSTRACT

Deriving a property of a protein that is unique to it has well known significance since the study on ab initio model based derivation of protein structure where uniqueness of protein sequence is taken as the source of specificity of protein structure. In this direction, Heat denatured protein aggregates (HDPA) of proteins were studied with an objective to derive some multi-fractal markers specific to constituent protein that may be further utilized to extract information of the seed protein. Since Ordinary microscopic images of aggregates were analyzed to extract Intensity Level-based Multifractal Dimension (ILMFD) features. ILMFD features include four different features, perimeter fractal dimension (ILMFD$_P$), perimeter-area relationship (ILMFD$_{PAR}$), Area fractal dimension (ILMFD$_A$) and Perimeter-area fractal dimension (ILMFDP$_A$) that were calculated using fractal computations considering perimeter, and area of aggregate images. Feed forward backpropagation network was used to classify the proteins using different ILMFD parameters. It was found that ILMFD features could discriminate the proteins used in our study, that points to their potential to serve as unique property or marker of a protein. Further to validate the results, the outputs from ANN were clustered, and the outputs clustered in the largest cluster were found to significantly improve the result in class decision given by ANN.

.

Keywords: ***Heat denatured protein aggregate; Multifractal dimension; Protein-marker; clustering***

[I] INTRODUCTION

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]. generally accompanied by conformational change of protein, which can be induced by thermal, enzymatic or chemical perturbations affecting the native folded structure of protein [3]. But recently several studies have pointed towards specificity of aggregates to their seed proteins.

Bohr et al (1997) [4]. through their experiment on native protein aggregate by electronic, atomic force and ordinary microscope, have shown that the structure of aggregates of proteins are strongly influenced by shape of constituent individual protein molecules. Also, from the study done by Taubes we find that protein aggregates are not as nonspecific as earlier believed [5].

In a simulation study on protein aggregation, Patro and Przybycien showed that variation in monomer surface property significantly affects the structure of kinetically irreversible protein aggregates [6]. Another study indicated that aggregation properties are affected by structural changes in proteins. Change in protein structure is found to significantly affect the topology and energetics of contacts within aggregates and thermodynamic drive towards aggregation [7].

Recently it has been found that aggregation is a generic property of polypeptide chains and aggregation propensity differs with difference in structure and environment [8]. Some studies have shown 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 [5]. 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 indicate that aggregation may be preprogrammed into amino acid sequence just like folding and aggregates should not be considered as just a nonspecific mess.

Taking cue from these studies, we have been searching towards a suitable aggregate feature that can be used as a marker showing specificity of aggregates to individual proteins. The background behind this effort is the expectation that such unique aggregate based feature should be map-able to individual protein property, especially structural property. Like the structural information we draw from protein crystal which is eventually an ordered assembly of protein and mostly scarce, we want to draw structural information of protein from its aggregate which is apparently not so ordered but available for almost all the proteins. To accomplish this goal, in this work we have extracted a scale and rotation independent feature Intensity Level-based Multifractal Dimension (ILMFD), based on mass fractal dimension of aggregates to study the rough pattern of heat denatured aggregates [10, 11]. As ILMFD is basically an aggregate-image based feature, it is quite likely that it captures the rough shape and texture of 3D-aggregates in its 2D-projected form. In current work, we have extracted three more features, to be included in ILMFD feature set and utilized them in a novel neuro-clustering classifier. This new approach has shown promise to significantly increase the specificity of ILMFD feature set to individual protein.

[II] MATERIALS AND METHODS

2.1. Formation of protein aggregates

The proteins used were Albumin, Cytochrome c, Ferritin, Hemoglobin, Insulin and Lysozyme. Proteins were purchased from Sigma Aldrich Inc. (USA). Water used in the experiments was purified by Millipore Water System (Model: Millipore, USA). Each Protein was dissolved in Millipore water at concentration of 25 mg/cc and kept at 100°C for 15 minutes to obtain its Heat Denatured Protein Aggregates (HDPAs).

2.2. Procuring microscopic images of aggregates

Suspension having homogeneously distributed HDPAs was spread over Hemocytometer slides (Model: Neubauer Chamber, Marienfeld, Germany) and visualized under phase contrast mode of compound light microscope (Leica Model DML-B2) at 400× magnification. Images of aggregates were captured using a digital camera (Canon PowerShot S50) attached with the microscope, at 2× optical zoom, resulting in to total magnification factor of 800×. For each protein 50 images of HDPAs at different fields of views were captured to create an aggregate image dataset.

2.3. Preprocessing and intensity plane slicing of images

Each aggregate image was converted to grey scale and resized to 1/3rd of the original size 2592x1944 pixels, to reduce computational complexity. Background of each aggregate image was made black using Adobe Photoshop 7.0 to nullify the effect of background on ILMFD parameters calculated form images. Each image, was split into 10 binary images on the basis of fixed intensity-ranges by applying the rule that in a binary image representing an intensity range, only the pixels having intensity values falling in that intensity range , would be kept as 1, while all other pixels would be assigned a value of zero. Computationally, intensity interval between maximum and minimum intensity of a particular image was divided into 10 smaller and equal intervals or ranges.

2.4. Deriving ILMFD features from aggregate images

Area (A), perimeter (P) of aggregates for 10 binary images representing each aggregate image were calculated at different scales of measurement (S) using box counting method. Four types of ILMFD features were derived using Area, and perimeter calculated at different scales of measurement i.e., box size [12,13]. Area (A) was calculated as number of boxes covering the aggregate in the image. Similarly, perimeter (P) was measured as the number of boxes making the periphery of the aggregate in the image. Perimeter fractal dimension was calculated as the slope of the linear regression plot between log(P) and log(S). Perimeter-area relationship was calculated as slope of linear regression plot between log(P) and log(A) at different box sizes (S). Area fractal dimension was calculated as the slope of linear regression plot between measured log(A) and log(S). Similarly perimeter-area fractal dimension was calculated as linear regression plot between two variables x, and y where x= log(P/S), and y=(log(A))/2 - log(S) [14]..Thus each aggregate image was represented by 10 fractal dimensions (one for each binary image), cumulatively referred to as ILMFD where, D_i is fractal dimension of one intensity level:

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

Thus we derived four different types of ILMFD parameters 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.

2.5. Classification by ILMFD parameters using artificial neural network

Each of the four ILMFD features was used separately for classification of images into different classes based on their constituent protein. The classification decisions from different ILMFD features were obtained using feed forward backpropagation networks where normalized values of ILMFD features I_F was used as input vector. ILMFD data set obtained from 300 images of all proteins, was divided in to training and test sets, by randomly picking data for 210 images as training set and for remaining 90 images as test set. Such five training and test sets were chosen randomly for training and testing the neural network based classifier. Each training and test ILMFD data was normalized by subtracting their column mean calculated from respective training ILMFD data.

Same network architecture was used for all the four ILMFD features. It consisted of one hidden layer apart from input and output layers. While the input layer comprised of 10 neurons, hidden layer consisted of 8 nodes. Output layer had six neurons to represent six classes of our interest. Tan sigmoid transfer function was used in hidden and output layers. Mean square error was used as performance function. Trained networks were simulated with normalized test ILMFD feature data for validation.

ISSN: 0976-3104

2.6. Classification by ILMFD parameters using neuro-clustering classifier

Classification decisions obtained for test sets of each protein were clustered using k means clustering. Value of k was kept as 2, considering possibility of two types of decisions i.e., correct or incorrect. Centre of the decisions grouped in larger cluster were matched with correct decisions of test dataset to validate the decision tendency of trained networks.

[III] RESULTS AND DISCUSSION

3.1. Potential of ILMFD features to classify and recognize individual proteins

The neural networks trained with different ILMFD features, were simulated for their respective test sets. To remove the possibility of any bias, training and testing of neural networks, was done using five different randomly chosen training and test sets from whole data. Results for neural networks giving maximum efficiency of protein classification on test sets as well as average efficiency of all networks for each ILMFD feature are shown in table 1. Maximum efficiency of protein classification on test set was found for features ILMFD$_P$ and ILMFD$_{PAR}$. Similarly network-average of efficiencies of five networks for classification of proteins using these same features were found to be the maximum among the features selected in our study **[Table-1]**.

Protein-wise sensitivity and specificity of classification using neural networks giving maximum efficiency on test set are given in **Table-2**. In **Table-3**, the efficiency in classifying a protein using decision clustering model of neural network outputs (we referred as neuro-clustering) has been shown.

ILMFD Feature	Training Set		Test Set	
	Efficiency of network giving Maximum Efficiency on test set	Average Efficiency of five networks	Efficiency of network giving Maximum Efficiency on test set	Average Efficiency of five networks
ILMFD$_A$	93.81	94.57 ± 0.87	74.44	70.44 ± 3.20
ILMFD$_P$	98.57	94 ± 6.96	80	76.88 ± 2.65
ILMFD$_{PA}$	96.67	94.95 ± 1.37	67.78	63.11 ± 3.08
ILMFD$_{PAR}$	94.76	94.85 ± 2.90	80	75.78 ± 2.53

Table: 1. Results of classification in percentage for proteins using different ILMFD features of HDPAs

Class	ILMFD$_A$		ILMFD$_P$		ILMFD$_{PA}$		ILMFD$_{PAR}$	
	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec
Albumin	61.54	76.62	57.14	84.21	63.64	68.35	90	78.75
Cytochrome c	64.71	76.71	80	80	37.5	74.32	63.16	84.51
Ferritin	73.68	74.65	100	75.34	78.57	65.79	100	76
Hemoglobin	83.33	72.22	94.12	76.71	81.25	64.86	93.33	77.33
Insulin	100	70.13	92.86	77.63	83.33	63.89	82.35	79.45
Lysozyme	60	76.25	46.15	85.71	60	69.33	57.14	84.21

Table: 2. Results of neural network based classification in percentage for each protein based on four different ILMFD features: ‘Sens’ and ‘Spec’ represent sensitivity and specificity respectively

	ILMFD$_A$	ILMFD$_P$	ILMFD$_{PA}$	ILMFD$_{PAR}$
Efficiency	100	100	67	100

Table: 3. Results of percentage efficiency for neuro-clustering based classification for each protein based on four different ILMFD features

Efficiency of ILMFD parameters to classify proteins is indicative to their potential to recognize and discriminate each of the proteins and thus to serve as markers for each proteins. For this reason we present efficiency-profile of different ILMFD features in **Tables-1**, **-2** and **-3**.

3.2. Selection of protein aggregates as study material

The idea behind this work originated from various studies on protein aggregates, indicating the specificity of aggregate properties to their constituent proteins. On the other hand limited applicability of experimental methods like x-ray crystallography, NMR, prediction methods like homology modeling and threading for protein structure determination etc. predicated the need for search of novel methods for determination of protein structure and structure based features. Easy availability of protein aggregates through simpler experimental set up as compared to protein crystals, encouraged us to investigate the possibility of deriving some protein specific aggregate features, which would be protein specific and could be further used to map some functionally important structure features like protein functional sites.

3.3. Suitability of aggregate data representation

An aggregate image represents the natural three dimensional texture of aggregate in two dimensional forms [15, 16]. Each image was sliced into different intensity planes using gray level intensity based method into binary images where each intensity level was supposed to grossly capture the three dimensional depth of the aggregate. At each intensity level four types of fractal dimensions were calculated. The whole set of fractal dimensions calculated from all the intensity planes constitute the multi-fractal features for a particular aggregate image. As we had considered area, and perimeter measurements at different scales of measurement, it is quite likely that information on geometrical rough-pattern of aggregate surface and perimeter was suitably represented through these multi-fractal dimension features. ILMFD feature set tries to capture the roughness pattern of aggregates at surface and peripheral parts, which may be specific to the aggregates of particular protein. Thus ILMFD features have potential to be used as a protein specific aggregate feature.

3.4. Robustness of ILMFD Features

The probable reason behind the high efficiency obtained through ILMFD features may be suitable representation of possibly unique patterns of aggregate surface and perimeter using ILMFD features. ILMFD features capture aggregate surface and perimeter patterns hidden in various intensity-depths. Moreover, the number of intensity levels was fixed after several trials, to get sufficient and equitable representation of intensity depths. Further studies may be done to find an optimum number of intensity levels to represent various intensity depths in aggregate images more reasonably.

3.5. Applicability of ILMFD features

The work dealt with development of a new approach for deriving structural information of protein by using light-microscopic images of protein aggregates as input data that is subsequently processed and mined for this purpose. The objective of ILMFD based classification of protein was to find a set of features which will serve as unique structural or functional signature. As aggregation is generally driven by interaction of its constituent proteins it is interesting to see whether this interaction has specificity to the structure or function of its ingredient i.e., individual protein. The proteins chosen by us have diverse function and patho-physiological behavior. Therefore a specific pattern of these proteins was expected from their aggregates. The capability of ILMFD features towards this direction to discriminate the proteins selected for this study is quite encouraging. For enhancing the discriminatory (i.e., classifying) power a novel neuro-clustering approach was adopted in which the overall efficiency of classification was found to increase significantly. Moreover, our approach utilizes a very simple protocol based on computation of data obtained from heat-denaturation of protein and ordinary microscopy. Therefore it is worth investigation to see whether this protocol may be utilized as a tool to identify proteins on the basis of their structural or functional families without taking the help of their PDB structures.

3.6. Applicability of neuro-clustering classifier

Concept of neuro-clustering classifier was introduced taking cue from the decision making process of human brain. Notion of group decision was applied using multiple instead of a single test data. In most cases larger cluster of decisions was found to represent the general tendency of decisions, while the smaller cluster was found to represent noise. Centre of larger cluster was found to be a close approximation of correct class decision in majority of cases leading to significant improvement in classification efficiency for all the ILMFD features except $\mathrm{ILMFD}_{\mathrm{PA}}$ **[Table-3]**. This kind of approach may find its applications in various other classification problems dealing with biological data.

[V] CONCLUSION

Promising results obtained from this study show the specificity of aggregate properties to constituent proteins. In the context of limited applicability of conventional complex methods for protein structure determination like x-ray crystallography and NMR, aggregation based methods hold potential to serve as starting point for development of novel alternative methods for derivation of protein structural features. Further exploration is required in this direction to develop methods to utilize these aggregate features to derive functionally important structural feature of protein like functional sites which lie on surface.

Moreover the concept of neuro-clustering introduced in this work proved to be a very useful classifier to handle complexity in input test dataset which points possible scope of its applications in other biological classification problems also.

FINANCIAL DISCLOSURE

This work was financially supported by the Ministry of Human Resource Development, Govt. of India (grant no.F.26-14/2003.T.S.V.Dated 14.01.2004) and the Indian Council of Medical Research, Government of India (grant no. 52/8/2005-BMS dated 04.02.2010).

ACKNOWLEDGEMENT

We acknowledge helpful discussion with Prof. Harish Karnik, Department of Computer Science & Engineering, Indian Institute of Technology, Kanpur, India.

REFERENCES

[1] Okanojo M, Shiraki K, Kudou M, et al. [2005] Diamines prevent thermal aggregation and inactivation if lysozyme. *J Biosci Bioeng* 100: 56–561.

[2] Pappu RV, Wang X, Vitalis A, et al. [2008] A polymer physics perspective on driving forces and mechanisms for protein aggregation. *Arch Biochem Biophys* 469:132–141.

[3] Weijers M, Barneveld PA, Stuart MAC, et al. [2003] Heat-induced denaturation and aggregation of ovalbumin at neutral pH described by irreversible first-order kinetics. *Protein Sci* 12:2693–2703.

[4] Bohr H, Kühle A, Sørensen AH, et al. [1997] Hierarchical organization in aggregates of protein molecules. *Z Phys D* 40:513–515.

[5] Taubes G. [1996] Misfolding the Way to Disease. *Science* 271:1493–1495.

[6] Patro SY, Przybycien TM. [1994] Simulations of Kinetically Irreversible Protein Aggregate Structure. *Biophys J* 66:1274–1289.

[7] Pullara F, Emanuele A, Palma-Vittorelli MB, et al. [2007] Protein aggregation/crystallization and minor structural changes: universal versus specific aspects. *Biophys J* 93:3271–3278.

[8] Monsellier E, Ramazzotti M, Taddei N, et al. [2008] Aggregation propensity of the human proteome. *PLoS Comput Biol* 4:e1000199.

[9] Brems DN, Plaisted SM, Havel HA, et al. [1988] Stabilization of an associated folding intermediate of bovine growth hormone by site-directed mutagenesis. *Proc Natl Acad Sci USA* 85:3367–3371.

[10] Lahiri T, Mishra H, Sarkar S, et al. [2008] Surface characterization of proteins using Multifractal property of heat-denatured aggregates. *Bioinformation* 2:379–383.

[11] Lahiri T, Mishra H, Kumar U, et al. [2009] Derivation of a protein-marker from heat-denatured protein-aggregate. *Online J Bioinformatics* 10:29–39.

[12] Kawaguchi E, Taniguchi R. [1989] The depth first picture-expression as an image thresholding strategy. *IEEE T Syst Man Cyb* 19:1321–1328.

[13] Zmeškal O, Veselý M, Nežádal M, et al. [2001] Fractal Analysis of Image Structures. *HarFA e-journal* 1:3–5.

[14] Feder J. [1989] Fractals, Plenum Press, Newyork, USA.

[15] Tuceryan M, Jain AK. [1998] Texture Analysis, in: The Handbook of Pattern Recognition and Computer Vision, Chen CH, Pau LF, Wang, PSP., (eds) World Scientific Publishing Co., Hackensack, New Jersey, USA, p. 207–248

[16] Haralick M. [1979]. Statistical and Structural Approache to Texture, *Proc IEEE* 67:786–804.

ABOUT AUTHORS

***Mr. Hrishikesh Mishra** received M.Tech. degree in bioinformatics from Indian Institute of Information Technology, Allahabad, India. He is currently a Ph.D. scholar in bioinformatics at Indian Institute of Information Technology, Allahabad, India, under guidance of Dr. Tapobrata Lahiri. His research interests include artificial intelligence and computational biology. He has authored 6 publications.*

***Dr. Tapobrata Lahiri** is Associate Professor and Head of the Division of Applied Sciences and Indo-Russian center for Biotechnology at Indian institute of Information Technology, Allahabad, India. His research interests include artificial intelligence and biophysics. He has 36 publications in this credit.*