ARTICLE
Genomics & Informatics
Vol. 9(2) 74-78, June 2011

DOI: 10.5808/GI.2011.9.2.74
pISSN 1598-866X eISSN 2234-0742

Analysis of the Globular Nature of Proteins

Sunghoon Jung[1,2] and Hyeon Seok Son[1,2,3*]

[1]Laboratory of Computational Biology & Bioinformatics, Institute of Health and Environment, Graduate School of Public Health, Seoul National University, Seoul 151-742, Korea, [2]Interdisciplinary Graduate Program in Bioinformatics, College of Natural Science, Seoul National University, Korea, [3]SNU Bioinformatics Institute, Seoul 151-742, Korea

Abstract

Numerous restraints and simplifications have been developed for methods that anticipate protein structure to reduce the colossal magnitude of possible conformational states. In this study, we investigated if globularity is a general characteristic of proteins and whether they can be applied as a valid constraint in protein structure simulations with approximated measurements (Gb-index). Unexpectedly, most of the proteins showed strong structural globularity (i.e., mode of approximately 76% similarity to the perfect globe) with only a few percent of proteins being outliers. Small proteins tended to be significantly non-globular (R^2=0.79) and the minimum Gb-index showed a logarithmic increase with the increase in protein size (R^2=0.62), strongly implying that the non-globular characteristics might be more acceptable for smaller proteins than larger ones. The strong perfect globe-like character and the relationship between small size and the loss of globular structure of a protein may imply that living organisms have mechanisms to aid folding into the globular structure to reduce irreversible aggregation. This also implies the possible mechanisms of diseases caused by protein aggregation, including some forms of trinucleotide repeat expansion-mediated diseases.

Keywords: protein aggregation disease, protein globularity, protein structure analysis

Introduction

Protein structure is considered to be the most important primary information in molecular biology, especially in pharmaceutical studies. The difficulty, however, of deriving structural information from proteins using protein crystals or protein solutions leads to the development of protein structure prediction methods based on amino acid sequences and other already revealed structures. The most critical problem of the protein structure anticipation method is the colossal magnitude of possible conformational states. Numerous restraints and simplifications have been developed to be appropriately applied to reduce the search space while minimizing false structures.

Recently, Palù *et al*. (2004) incorporated globularity as their protein folding simulation criterion using Constraint Logic Programming. Globularity, expressed by the radius of gyration, was used to improve the packing and accuracy of NMR structures in previous research (Kuszewski *et al*., 1999), and the validity of the globular restraint for NMR protein structure determination was examined by Huang and Powers (2001). Globularity was also successfully used to assess the quality of models submitted to the Critical Assessment of Techniques for Protein Structure Prediction center (CASP; Constantini *et al*., 2007). Although protein globularity is assumed to be a valid criterion in many studies, to our knowledge, an analysis of the globularity of proteins investigating a whole database of protein structures had not been performed.

Here, we investigated if globularity is a general character of most proteins and whether it can be applied as a valid constraint in protein structure simulations. We used virtually all the protein structures in the Research Collaboratory for Structural Bioinformatics' Protein Data Bank (RCSB's PDB) database. We removed redundant entries and divided the proteins into subcategories to enable a more detailed analysis.

Chaperones are known to protect the aggregation of misfolded proteins by binding and aiding the recycling of the folding process, especially in the endoplasmic reticulum (ER) during protein synthesis. The delineation of correct- and misfolded states by chaperones suggests a conundrum because many more proteins exist than chaperones and related molecules. Complete recognition of a correctly folded structure by a structural protein-protein binding site interaction is almost impossible because one protein might possess numerous structural characters. Accordingly, a possibility exists that the globular character might be the checkpoint of the correct folding in biological organisms. This assumption is supported by the aggregation of misfolded proteins because non-globular proteins tend to bind more tightly

*Corresponding author: E-mail hss2003@snu.ac.kr
Tel +82-2-880-2746, Fax +82-2-762-2888
Accepted 23 May 2011

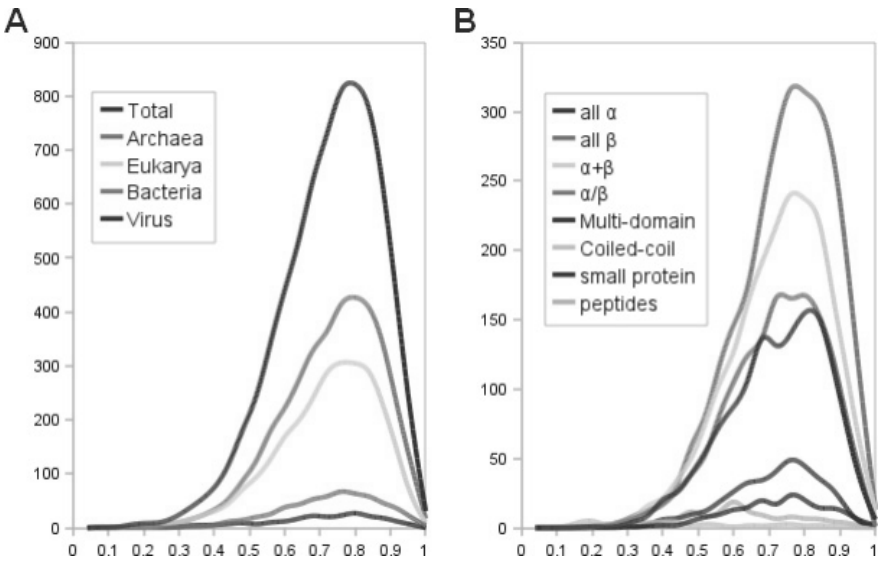

Fig. 1. Distribution of Gb-indices among proteins of different types and different sources of organisms. (A) Distribution of Gb-indices of proteins from four different types of organisms (archaeal, bacterial, eukaryotic and viral proteins). All four proteins from different organisms showed similar distributions. Details of the distribution of the values are listed in Table 1. (B) Distribution of Gb-indices of eight different types of proteins according to the SCOP classification. Peaks were between 0.7 and 0.8, except the peak of peptides (0.6) and coiled-coil (0.43) proteins.

with one another due to the larger surface area provided by the loss of globularity.

Unexpectedly, most of the proteins showed strong structural globularity (i.e., mode of approximately 76% similarity to the perfect globe) with only a small proportion of the proteins being outliers. This strong perfect globe-like character implies partial validity to the postulation that living organisms have mechanisms to aid folding into globular structures to reduce irreversible aggregation. It also implies the possible mechanisms of protein aggregation diseases including some forms of trinucleotide repeat expansion-mediated diseases.

Methods

Data sets

PDB files were collected based on the Structural Classification of Proteins (SCOP). We used all-alpha, all-beta, alpha/beta, alpha+beta, multidomain proteins, and other minor proteins including peptides, small proteins, and coiled-coil proteins. We excluded membrane proteins and peptides because of their topological difference and lipid membrane surrounding environment, which has different characteristics from that of soluble proteins. We also excluded fragmented and nucleic acid-containing structures, but included ligand-bound proteins. We removed structures that have 90% or more sequence identity to others to reduce redundancy. Redundancy also arose from proteins with multiple folds belonging to different sources of organisms or different SCOP classes. PDB entries with redundant source organisms were removed but structures with two or more different SCOP classes were not excluded to allow the investigation of as many protein structures as possible.

In total, 7,131 PDB structures were analyzed with 1,365 all-alpha chain, 1,503 all-beta chain, 2,690 alpha/beta chain, 2,067 alpha+beta chain, and 182 multidomain chain containing proteins and 547 other proteins. Programs to sort proteins according to their structural classification, source of organism, oligomeric states, and to filter out redundant and fragmented structures were all written in JAVA language.

Globularity measurement

We defined new simple geometric quantities to represent globularity other than the radius of gyration because the radius of gyration might misinterpret internal cavities. Our globularity index (Gb-index) was defined as the ratio of the length of the longest displacement of any two atoms of the protein to the average of the longest lengths of two displacements that are orthogonal to each other and to the longest displacement. This approximated measure was chosen because cubic proteins are assumed to be extremely rare in real cases. The orthogonal criterion was surveyed within a 2° span from a perfect orthogonal angle. A range of 2° was successful for all the cases tested. We calculated the mean, standard deviation (S.D.), median, and the minimum and the maximum values of these indices of globularity. All the necessary programs were written in JAVA.

Results and Discussion

The distribution of the degree of globularity was analyzed according to the source of organisms (Fig. 1A) and SCOP classification (Fig. 1B). Except coiled-coil

Table 1. Gb-index of Different Types of Proteins

Type	Mean(s.d.)	*Median*	Mode	Min	Max	Number
SCOP classes						
all-α	0.70 (0.14)	*0.72*	0.84	0.19	0.99	1,365
all-β	0.70 (0.14)	*0.71*	0.80	0.14	0.97	1,503
$\alpha + \beta$	0.73 (0.14)	*0.72*	0.76	0.12	0.99	2,690
α / β	0.71 (0.13)	*0.74*	0.76	0.15	0.99	2,067
multidomain	0.70 (0.13)	*0.72*	0.76	0.28	0.96	182
coiled-coil	0.42 (0.22)	*0.40*	0.20	0.08	0.86	49
peptides	0.59 (0.16)	*0.58*	0.60	0.24	0.93	129
small proteins	0.69 (0.13)	*0.71*	0.76	0.21	0.97	369
Source of organisms						
Archea	0.71 (0.13)	*0.73*	0.72	0.28	0.98	568
Eukarya	0.70 (0.14)	*0.72*	0.72	0.12	0.99	3,672
Bacteria	0.71 (0.14)	*0.73*	0.76	0.15	0.99	2,657
Virus	0.67 (0.18)	*0.70*	0.76	0.08	0.95	234
Total	0.71 (0.14)	*0.72*	0.76	0.08	0.99	7,131

proteins and peptides, all kinds of proteins of SCOP classifications including all-alpha, all-beta, alpha/beta, alpha+beta, multidomain, and small proteins showed mean Gb-indices from 0.69 to 0.73 and modes from 0.76 to 0.84. Their median values ranged from 0.71 to 0.74. The mean and mode of all proteins was 0.71 (S.D. 0.14) and 0.76, respectively, with a median of 0.72. The mean Gb-index of peptides was 0.59 (S.D. 0.16) and the median was 0.58. Coiled-coil proteins showed the lowest Gb-index with a mean of 0.42 (S.D. 0.22) and a median of 0.40. Modes of the Gb-indices of peptides and coiled-coil proteins were 0.6 and 0.2, respectively. Details of the values of each type of protein are shown in Table 1.

No significant difference of globularity was observed between proteins from different organisms. Gb-indices showed similar average (0.70-0.71) and median (0.72-0.73) values among proteins from different organisms, except viral proteins, which showed a slightly lower average Gb-index of 0.67 (S.D. 0.18) and median of 0.70. The modes of the Gb-indices of the proteins from different organisms ranged from 0.72 to 0.76 with the mode of 0.76 for the whole protein, 0.72 for archaeal and eukaryotic proteins, and 0.76 for bacterial and viral proteins.

Viral coiled-coil proteins gave a minimum Gb-index of 0.08 (PDB ID: 1pjf). The mean Gb-index of viral proteins was slightly lower (mean of 0.67) compared to proteins from other sources (mean of 0.71). Though this deviation may have been due to the small sample size (234 entries) compared to other proteins (total of 7,131), the lower globular character might have originated from the abundance of structural capsid proteins.

Ninety-five percent of the proteins from any biological source had Gb-indices higher than 0.453, and 97% of the proteins had Gb-indices higher than 0.413. This result strongly indicated that almost every protein is globular, partly validating previous attempts that used the globularity criterion in anticipating protein structures (Palù *et al.*, 2004). However, non-globular, linear proteins were observed, as represented by low Gb-indices, which implies that the folding criterion based on globularity might not be suitable for all cases. Rather than the uniform mathematical formula for the radius of gyration according to the length of the polypeptide chain (Skolnick *et al.*, 1997), sequence- and other character-based homology search for expected globularity might be more suitable for proteins with varying degrees of globularity.

A few percent of non-globular proteins existed although most of the proteins were globular. We investigated the possible relationship with the size of the protein and the tendency to lose the globular structure. We drew a graph of the mean, minimum, and maximum Gb-indices of proteins along with the number of atoms in the proteins (Fig. 2). In all cases, the means were always approximately 0.7 and the maximum Gb-indices were always just below 1.0. However, the minimum globularity of proteins showed logarithmic growth with the square correlation coefficient (R^2) of 0.62 to the regression line, indicating that smaller-sized proteins were more likely to deviate from globular structures.

We also analyzed the relationship of the numbers of proteins with globularity lower than 0.453, i.e., the lowest 5% of non-globular proteins with the size of the proteins (Fig. 2). The findings showed that the smaller the protein, the more non-globular in structure, with the square correlation coefficient (R^2) of 0.79 for the observed data and the power regression line. This and the

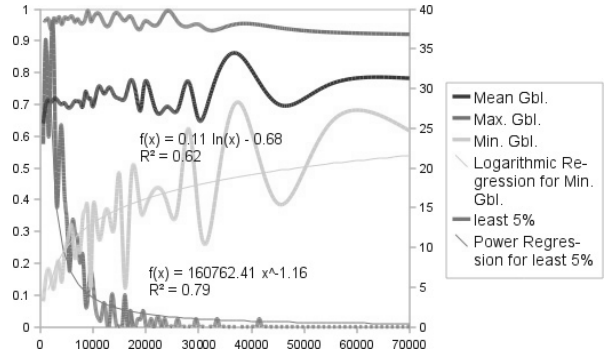

Fig. 2. Change in protein globularity with protein size. The mean, maximum, and minimum Gb-indices were plotted against the atom numbers of proteins. As the atom number increased, the minimum globularity measure showed a logarithmic increase, indicating the square correlation coefficient of 0.79 with the regression lines. The maximum and mean values, however, stayed rather constant along the whole range of protein size. The relationship between protein size and the minimal globularity index indicates that non-globular structure might be more permissible in smaller proteins. The number of proteins with a globularity index lower than 0.453 (the lowest 5% of non-globular proteins) was also plotted against protein size. The square correlation coefficient of 0.62 was shown with the decreasing power regression line, possibly indicating again that the small size permits less globular structures of proteins.

Fig. 3. Structure of the amyloid-beta peptide. The fibrillar structure of the amyloid-beta peptide is shown by a ball-and-stick model and backbone helix ribbon. The Gb-index was 0.2094, which belongs to the lowest 0.24% of non-globular proteins. This non-globular structure of the molecule might aid in irreversible aggregation. Image was prepared with Sirius visualization system.

logarithmic increase in the minimum Gb-index with increasing protein size strongly implied that non-globular characteristics might be more acceptable for smaller proteins than larger ones.

Calnexin (Bergeron *et al.*, 1994) and calreticulin (Michalak *et al.*, 1999) are chaperones that are known to retain inappropriately folded proteins in the ER. The delineation of correct and incorrect folded proteins is known to be the function of another ER enzyme, glucosyl transferase (Ellgaard *et al.*, 2001). The interplay of these three key enzymes retains incompletely folded proteins in the ER. A significant portion of proteins in the ER are misfolded and translocated back into the cytosol and degraded. (Plemper and Wolf, 1999). How the myriad of numbers of misfolded proteins is recognized individually, and why misfolded proteins are likely to aggregate with each other to make plaques or crystals in side the ER, remains unclear.

565 eucaryotic secretory proteins, which originates from the rough ER and proceeds to Golgi apparatus, showed mean Gb-index of 0.70 (s.d. 0.14) and minimum and maximum Gb-indices of 0.12 and 0.96 each, partly indicating that globularity might be useful in preventing the irreversible aggregation. 31 proteins (i.e., 5.49% of the 565 proteins) showed Gb-index lower than 0.453, which is the threshold value of the least 5% of non-globular proteins. The degree of globularity of these proteins was as strong as, or might have been even stronger than non-secretory proteins considering the smaller sizes with mean atom number of 2686 (s.d. 3285) than the size of the total proteins investigated with mean atom number of 4808 (s.d. 5834) and the correlation of the loss of the globularity with protein's small size.

The globular structure of proteins might help prevent irreversible and pathological aggregation because it has the minimum surface area of a specified volume. The size of the interacting surface area is widely known to strongly correlate with the binding strength. Two perfect globes will have virtually no contacting area because of the convex shapes of both. However, two rodlike proteins would line up side by side with the strong interaction through the long contactable area. The relationship between low protein size and the propensity to lose globular structures might be explained by the smaller contactable surface area of smaller proteins than larger proteins with the same globularity.

In conclusion, We investigated the degree of structural globularity of proteins with an approximate measurement. The results strongly indicated that virtually every protein (95%) was significantly globe-shaped with Gb-indices larger than 0.453. Some oddities were found mainly among small proteins. The small size of the protein and the tendency to have significantly non-globular structure showed a rather high correlation (R^2 = 0.79).

The minimum Gb-index showed a logarithmic increase along the increase of protein size (R^2 = 0.62).

The suggestion that globularity might function to prevent aggregation is somewhat intriguing considering the interest in protein aggregation associated with neurodegenerative diseases. Pathogenic aggregations of proteins may have been caused by the loss of globularity of normal proteins or by the overproduction of the proteins with low globularity. Fig. 3 displays the least globular structure (model 7) of the amyloid-beta peptide (PDB ID: 1BA4), which is known to aggregate and make plaques in neurons in Alzheimer's disease. The Gb-index of this structure was 0.2094, which is in the smallest 0.24% of the 7,131 proteins examined. Although confirming that the less globular structure has a primary effect on pathological aggregation is not sufficient, one can still aid in the irreversible aggregation. This supposition might also be applied to some form of trinucleotide repeat expansion diseases, which also show pathological plaques; i.e., inserted polyamino acids might induce the loss of globularity in normal proteins.

Our Gb-index has shortcomings in the delineation of some polygonal structures from globular proteins. Although most of the polygons would be implausible for real protein structures, cubic structures and other structures with strongly planar or concave surfaces are still possible. The analysis of the curvature of the surface of a protein would result in a more accurate inference of the degree of globularity for these exceptional cases. Also, the search span of 2° for searching the orthogonal displacement to the longest displacement among all possible atom pairs may be too small and might cause the Gb-index to decrease. Our finding regarding the distribution of globularity of known protein structures can be used to suggest proper constraints for protein folding algorithms. This analysis of the simple physical character of protein structures might be helpful in anticipating protein structures.

Acknowledgements

We acknowledge the contribution of all those who have made their invaluable data publicly available. This work was partly supported by the BK 21 project in 2011.

References

Bergeron, J.J., Brenner, M.B., Thomas, D.Y., and Williams, D.B. (1994). Calnexin: a Membrane-bound Chaperone of the Endoplasmic Reticulum. *Trends Biochem. Sci.* 19, 124-128.

Constantini, S., Macchiano, A.M., and Colonna, G. (2007). Evaluation of the Structural Quality of Modeled Proteins by Using Globularity Criteria. *BMC Struct. Biol.* 7, 9.

Ellgaard, L. and Helenius, A. (2001) ER quality control: towards an understanding the molecular level. *Curr. Opin. Cell. Biol.* 13, 431-437.

Huang, X., and Powers, R. (2001). Validity of Using the Radius of Gyration as a Restraint in NMR Protein Structure Determination. *J. Am. Chem. Soc.* 123, 3834-3835.

Kuszewski, J., Gronenborn, A.M., and Clore, G.M. (1999). Improving the Packing and Accuracy of NMR Structures with a Pseudopotential for the Radius of Gyration. *J. Am. Chem. Soc.* 121, 2337-2338.

Michalak, M., Corbett, E. F., Mesaeli, N., Nakamura, K., and Opas, M. (1999). Calreticulin: one protein, one gene, many functions. *Biochem. J.* 344(Pt2), 281-292.

Palù, A.D., Dovier, A., and Fogolari, F. (2004). Constant Logic Programming Approach to Protein Structure Prediction. *BMC Bioinfomatics* 5, 186.

Plemper, R.K., and Wolf, D.H. (1999'). Retrograde Protein Translocation: ERADication of secretory proteins in health and disease. *Trends Biochem. Sci.* 24, 266-270.

Skolnick, J., Kolinski, A., and Ortiz, A.R. (1997). MONSSTER: A Method for Folding Globular Proteins with a Small Number of Distance Restraints. *J. Mol. Biol.* 265, 217-241.