Relationship between B-factor and average shortest path in the protein structure

Kamal Mirzaie Badrabadi*[,1], Mehdi Mirzaie[2,3]

[1] Department of Computer Engineering, Maybod Branch, Islamic Azad University, Maybod, Iran
[2] Department of Computational Biology, Faculty of High Technologies, Tarbiat Modares University, Tehran, Iran
[3] School of Biological Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

*Corresponding Author: email address: k.mirzaie@maybodiau.ac.ir (K. Mirzaie Badrabadi)

ABSTRACT

Protein structural flexibility is important for catalysis, binding, protein design, and allostery. Some simple methods have recently been introduced to compute protein flexibility directly from the protein structure without any mechanical models. For example the atomic mean square displacement (or B-factor) is related to the number of neighboring atoms. The protein structure can be modeled as a graph where nodes represent atoms and edges can be defined by Delaunay tessellation procedure with weight equal to d^2 where d is the Euclidean distance between pair of atoms. In this study, we show that the average of shortest path for each atom in this graph is related to the B-factor.

Keywords: Protein structure; Flexibility; B-factor; Delaunay graph; Shortest path.

INTRODUCTION

Protein flexibility is related to its function. Structural flexibility is essential for its activity but, on the other hand, structural stability requires rigidity. Flexible regions are found in catalytic sites, binding sites, antigenic regions, and allosteric hinge sites. Proteins with similar functions have similar excess of flexibility in their optimum reaction conditions. The core of protein is relatively tightly packed and surface residues are the most flexible [1].

The protein flexibility has also been studied by normal mode analysis (NMA) [2]. The NMA calculates the 2nd derivative matrix (also called the Hessian matrix) of the total potential function of the energy-optimized structure, and then calculates the normal mode eigenvectors and eigenvalues of the matrix. The Elastic Network Model (ENM) or Gaussian Network Model (GNM) [3] is a coarse-grained version of NMA. The ENM or GNM is based on a simpler mechanical model that all neighboring residents that are within a certain cut-off distance are connected to each other by a uniform harmonic potential function and a Hessian matrix is calculated based on that simple force field, and the modes of motion are calculated through the diagonalization of this matrix.

Recently, simple methods have been introduced to compute protein flexibility directly from the protein structure without any mechanical models, for example the square of the atomic distance from the center of mass of the protein [4], the number of surrounding atoms [5], and average shortest path lengths in residue interaction network [6]. Unlike NMA, this method does not require matrix operations like matrix diagonalization and therefore can handle quite large proteins.

The B-factors of protein crystal structures reflect the fluctuation of atoms about their average positions and provide important information about protein dynamics. Atilgan et al. in 2004 [6] showed that the average of the shortest path lengths of a residue to all other residues in a protein is related to B-factor. In this study, we show that the average shortest path lengths in a weighted graph constructed by Delaunay procedure is also related to atomic fluctuations. There are generally two types of representations for protein structures as a network. In the first presentation, two atoms are considered in contact, if they are separated by a distance of less than τ (in this study 6Å), we call this *threshold model*. In the second presentation, the Delaunay tessellation procedure is used to determine the nearest

neighbors of each atom. In this method, the nearest neighboring points in a three-dimensional space can be determined by Voronoi tessellation method, which partitions the space into convex polytopes called Voronoi polyhedra. For a given set of points, the Voronoi polyhedral is the region of space around each point that all points of this region are closer to this point than any other. A group of four points, whose Voronoi polyhedra meet at one vertex, forms another basic topological object called the Delaunay tessellation simplex. Two points are defined to be the nearest neighbors, if they are two vertices of an edge in a simplex and are separated by a distance of less than 6Å. In fact, two atoms are defined to be in contact, if they are two vertices of an edge in a simplex.

METHODS

In order to obtain the nearest neighbors in a protein structure, the amino acids in a protein chain are represented by all heavy atoms and then the Delaunay tessellation of the resulting point set is computed using Qhull [7]. Two atoms are defined to be in contact, if they are two vertices of an edge in a simplex. Then, two atoms are in contact if they do not shield from contact by other atoms. In our model, each contact extracted by Delaunay procedure is an edge with weight equal to d^2 where d is the Euclidean distance between pair of atoms. We show that the average path length distance of each atom to other atoms is related to B-factor, as defined bellow.

Consider a weighted graph G with the set of vertices V. Let $d(u,v)$, where $u, v \epsilon V$ denote the shortest distance between u and v. Assume that $d(u,v) = 0$ if $v$cannot be reached from u. Then, the average path length for vertex u is (1):

$$\frac{1}{n(n-1)} \sum_{v \in V-\{u\}} d(u,v)$$

, where n is the number of vertices in G

To measure the performance of our methods, we calculated the correlation coefficient between the average of shortest path and B-factor as given by (2)

$$r = \frac{\sum_{i=1}^{M}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{[\sum_{i=1}^{M}(x_i - \overline{x})^2][\sum_{i=1}^{M}(y_i - \overline{y})^2]}}$$

Where x_i and y_i are the experimental and predicted values of B-factor of the i-th atom, and x and y are their corresponding sample means. Here we consider every chain in the data set as a single dynamical module and compute its r profile.

Data Set

We selected our date set from PDB-REPRDB 972 protein chains with a length larger than 60 residues [8]. All structures are solved by X- ray crystallography with resolution higher than 2.0 A˚ and R-factors less than 0.2. All chains are of pair-wise sequence identity less than 25%.

Figure 1 shows the computed and the X-ray B-factor profiles of the two proteins 1FUPA and 1KQFB perform well in our method. The correlation coefficients between the calculated and experimental B-factors are 0.90 and 0.85, respectively.

Figure 2 illustrates the distributions of the correlation coefficients between the X-ray and the computed B-values using our method. The p-values for 962 of the proteins (99% of proteins in dataset) are less than 0.01. The mean correlation coefficient is 0.53 and 63% of the proteins in the data set have correlation coefficient > 0.5.

Atilgan et al. in 2004 [6] showed that the average of the shortest path lengths of a residue to all other residues in a protein is related to B-factor. But Delaunay graph of protein is considerably sparser than threshold graph. In fact, in our data set, on average, the proteins in threshold graph have three times more neighbors than those of Delaunay graph.

Also it is noticeable that the average of shortest path for each atom in Delaunay model is equal to threshold model (data are not shown). Therefore, in the proposed method, we are able to calculate the shortest path between atoms in a significantly shorter time without loss of any information.

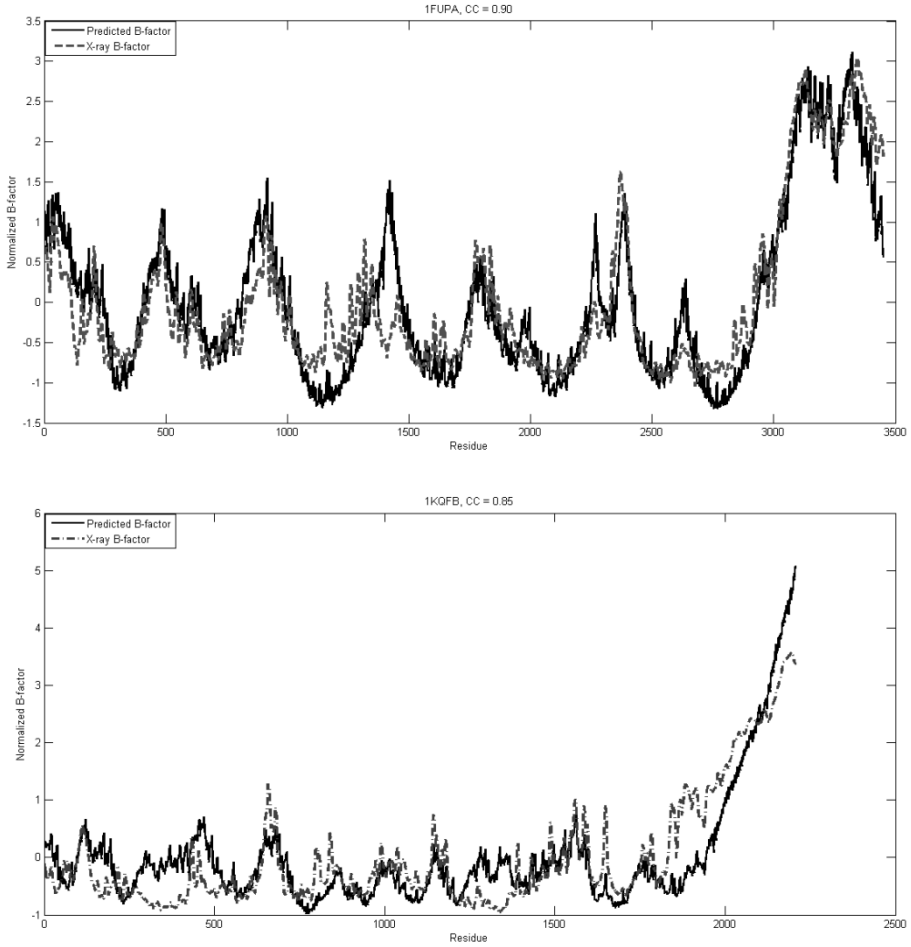

Figure 1. The computed (solid line) and X-ray (dotted line) B-factors of 1FUPA and 1KQFB structures

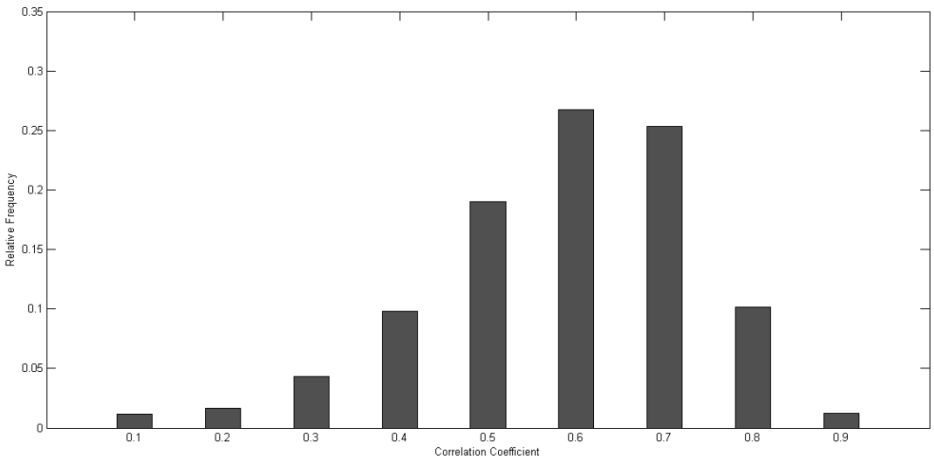

Figure 2. The distribution of correlation coefficients between the X-ray and the computed B-factor

In CN model [9], the average correlation coefficient is 0.51 and there are 54% of the proteins in the data set with a correlation coefficient > 0.5. Therefore, our method performs better than the CN model. However, in GNM model the average correlation coefficient is 0.56 and there are 69% of the proteins in the data set with a correlation coefficient > 0.5.

Structural flexibility is essential for the activity of protein but, on the other hand, structural stability requires rigidity. Flexible regions are found in catalytic sites, binding sites, antigenic regions, and allosteric hinge sites. Hence, by suggesting parameters related to B-factor and therefore protein flexibility, we can obtain structural information about these items. For example, there are observations [10, 11] that the catalytic residues usually have smaller B-factors than others. According to our model, this observation is equivalent to the observation that the catalytic residues usually have small average of the shortest path and this means that these residues are located near the centroid of the enzymes. This is consistent with studies in this area [12, 13].

ACKNOWLEDGMENTS

The authors would like to thank Maybod Branch of Islamic Azad University, for their funding this project.

REFERENCES

1.Vihinen, M., E. Torkkila, and P. Riikonen, Accuracy of protein flexibility predictions. Proteins: Structure, Function, and Bioinformatics, 1994. 19(2): p. 141-149.

2.Levitt, M., C. Sander, and P.S. Stern, Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. Journal of Molecular Biology, 1985. 181(3): p. 423-447.

3.Kidera, A. and N. Go, Refinement of protein dynamic structure: normal mode refinement.Proceedings of the National Academy of Sciences, 1990. 87(10): p. 3718-3722.

4.Shih, C.h., et al., SHORT COMMUNICATION A Simple Way to Compute Protein Dynamics Without a Mechanical Model. 2007. 38(December 2006): p. 34-38.

5.Lin, C.-P., et al., Deriving protein dynamical properties from weighted protein contact number. Proteins: Structure, Function, and Bioinformatics, 2008. 72(3): p. 929-935.

6.Atilgan, A.R., P. Akan, and C. Baysal, Small-World Communication of Residues and Significance for Protein Dynamics. 2004. 86(January): p. 85-91.

7.Barber, C.B., D.P. Dobkin, and H. Huhdanpaa, The quickhull algorithm for convex hulls, in ACM Transactions on Mathematical Software (TOMS). 1996. p. 469-483.

8.Noguchi, T. and Y. Akiyama, PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. Nucleic Acids Research, 2003. 31(1): p. 492-493.

9.Halle, B., Flexibility and packing in proteins. Proceedings of the National Academy of Sciences, 2002. 99(3): p. 1274-1279.

10.Yang, L.-W. and I. Bahar, Coupling between Catalytic Site and Collective Dynamics: A Requirement for Mechanochemical Activity of Enzymes. Structure, 2005. 13(6): p. 893-904.

11.Yuan, Z., J. Zhao, and Z.-X. Wang, Flexibility analysis of enzyme active sites by crystallographic temperature factors. Protein Engineering, 2003. 16(2): p. 109-114.

12.del Sol, A., et al., Residue centrality, functionally important residues, and active site shape: Analysis of enzyme and non-enzyme families. Protein Science, 2006. 15(9): p. 2120-2128.

13.Avraham Ben-Shimona and M. Eisenstein, Looking at Enzymes from the Inside out: The Proximity of Catalytic Residues to the Molecular Centroid can be used for Detection of Active Sites and Enzyme–Ligand Interfaces. journal of molecular biology, 2005. 351(2): p. 309–326.