



Determination of amount of alpha carbon atom lied in gap between protein segments using Perl programming

Wachiraporn Wanichnopparat, Chulalongkorn University, Bangkok Thailand

19th July 2011 – 8th September 2011

Abstract

Nowadays, protein structural could be obtained from a consequence of many approaches such as X-ray crystallography and nuclear magnetic resonance (NMR). However, there were some obstacles from limitation of methods and sample preparation. Electron density map that obtained as a result of them became a main source for building protein structure. In this study, we aimed to determine number of alpha carbon atom that lied in distance between two segments of protein by Perl language programming. To get a coordinate of all atoms in protein, PDB data file of Glucose -6-phosphate dehydrogenase were downloaded and collected. Then, distances between 2-10 atoms were computed and counted their frequencies. The graphs of frequency were plotted and normalized to calculate a probability. The results demonstrated the distance and probability of number of atoms found during distances which was similar between two chains of G6PD. Thus, it should be concluded that, in this case, coordinate of atoms in electron density map should be applied to find out alpha carbon atoms which located during each distance in protein molecule. However, there were many factors that should be considered when the model was constructed.

Content

• Introduction	3
• Theory	3-4
i) Perl	
ii) Electron density map	
• Objective	4
• Method	5
i) Data collection	
ii) Distance measurement	
iii) Frequency analysis	
• Results	6-8
i) Glucose-6-phosphate dehydrogenase	
• Discussion	9
• References	10
• Acknowledgements	10

- **Introduction**

There are several techniques to determine protein structure such as X-ray Crystallography and Nuclear Magnetic Resonance (NMR). Although, these approaches are famous in structural biology, some data are also unproved due to many factors such as phase and resolution. To evaluate structure of protein, three dimensional structures, secondary, tertiary and quaternary conformation, always are relevant. These conformations depend strongly on many forces between atoms. Moreover, there are some connections between their shape and some activities in cells. So, protein that loses some data from the experiments cannot build a molecular model precisely. Electron density map of protein is a solution that used to solve the problem through computational and bioinformatics approaches. In this study, some proteins were used as a subject to evaluate their structure using data collection in a database. The distance of next alpha carbon of missing residues were interested and found out. Furthermore, all steps above were manipulated using Perl programming. The results should be expected that the position of unknown residues were discovered and secondary structure of proteins were predicted.

- **Theory**

Perl

Perl, Practical Extension and Report Language, was developed by Larry Wall in 1987. This programming language was adapted its features from several other languages such as C, Shell scripting and AWK. Although Perl could be applied with text files, it could be worked with graphic modeling as well. Because of it was not complicated to manipulate, it became famous in Bioinformatics field. Many researches applied this programming language to operate their works. For example, PSPP software which predicted structure of protein used Perl language programming in processing(1). In this study, Perl was used to manage and collect some data contained in PDB files.

Electron density map

Almost all data of protein structures is from x-ray crystallography and nuclear magnetic resonance (NMR). To construct a model of protein, two things that important to prove were phase problem and electron density map. In the experiment, a value that could be known was intensity of wave, while a phase that changed during diffraction could not be measured. So, the lost phase, called phase problem, which contained important information would be solved for getting some good results. To solve phase problem, there were many methods used such as isomorphous replacement and anomalous scattering etc.

Another one that was solved for building a model was an electron density map. It was obtained from interpretation of a diffraction pattern using Fourier transform. So, electron density map corresponded to coordinate of all atoms in unit cell of molecule. If the resolution and phase determination were good enough, the quality of density map of electron would be good to construct a model of protein molecule. However, there were several factors such as biological sample and environment that could be interfered during experiment. Sometimes, the electron density map was also provided in low resolution led to difficulty to build a main chain of protein molecule. Only some part of molecule could be constructed with a bad density map.

ARP/wARP, a program for build macromolecular model, always used electron density map to form structure of molecule via hybrid model. So, density map would condense to a set of free atoms. After refinement, it was separated into two groups, gain chemical density atoms and free atoms. Hybrid model would integrate two groups of information. Finally, atomic position in hybrid model would be a subject to construct a model of protein later. Furthermore, main chain of protein could be extracted using hybrid model(2).

For low resolution of electron density map, there were many methods used to identify some domain of protein molecule. The electron density map was separated into segment following by pattern recognition. At last, the model of some domains would be compared if it matched superimpose with structure in PDB database. Thus, for using electron density map to build the model, it was not necessary to known protein conformation and target domain(3).

- **Objective**

Because of data obtained from experiments often lost some parts of molecule. Sometimes, main chain of protein structure could not build completely. Thus, this study aimed to find out amounts of alpha carbon atom located in each distance during gap of protein segments using Perl programming.

- **Method**

Data Collection

The PDB files of interested proteins were downloaded in <http://www.pdb.org>(4). In this case, 2 proteins were been subjects. Firstly, Glucose-6-phosphate dehydrogenase (G6PD) that prepared from *Leuconostoc mesenteroides* was an enzyme in pentose phosphate pathway. This protein contained 2 chains of polypeptides with 485 amino acid residues in both chains. In addition, G6PD data were collected from x-ray crystallography at resolution 2.0 angstrom(5).

Distance Measurement

To find an alpha carbon position of next amino acid residue of protein, a coordinate of alpha carbon of each amino acid residues were collected from PDB files. A distance between an alpha carbon atom and an adjacent one was calculated according to this equation.

$$\text{Distance (Angstrom)} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

In addition, a distance between 3 atoms of alpha carbon was calculated using the equation above following by 4 atoms until 10 atoms of alpha carbon. All steps were done by programming via Perl.

Frequency Analysis

The distances between 2-10 atoms of alpha carbon were measured and collected as described previously. After that, frequencies of each distance, which ranges were difference in 1.0, 0.5 and 0.1 angstrom, were count and drew a bar graph respectively. To refine the data, distances between atoms in every distance were normalized into 1 and plotted a line graph to identify an area under curve.

- **Results**

Glucose-6-phosphate dehydrogenase

After coordinated of alpha carbon of amino acid residues were collected via Perl programming, average distances between each pair of alpha carbon atoms were calculated. The results were displayed in a table below (Table1). It was noticed that the average distances were not different in both chains of G6PD. Furthermore, frequency of distances in each range of pairs were counted and demonstrated by a bar graph. The consequence of G6PD chain A was compared with the consequence of G6PD chain B (Figure1). It could be suggested that trends of every couples always be in the same direction. In addition, the highest scores were established in the vicinity of distance. For instance, the frequency of distance between each pair of 5 alpha carbon atoms that located in 5.5 – 6.0 and 6 – 6.5 angstrom was the highest frequency in G6PD chain A and G6PD chain B respectively (Figure1c).

The frequencies accumulated from previous study were normalized to 1 in every distance from minimum to maximum values. After that, line graphs were plotted using normalized data and adjusted to an area curve (Figure2). This area curve also represented a probability of number of alpha carbon atoms which could be found during a period of each distance. For example, at distance between 7-7.5 angstrom, it could be found 7 atoms for 5.7% (figure2a). Although, the average distances were similar, bar graphs and area curves of G6PD chain A was not alike to G6PD chain B at some points. Hence, it should be proposed that the structure of G6PD chain A and chain B molecule were similar with quite different in their size. Moreover, the number of alpha carbon atom of amino acid residue could be predicted in each distance using the data from distance between pairs of atoms and density of alpha carbon atom in the range of distance.

Table1 Distance between each pair of atoms of G6PD from *Leuconostoc mesenteroides*.

Range (between atoms)	Average Distance of each protein (Angstrom)	
	G6PD chain A	G6PD chain B
2 atoms	3.7986	3.8037
3 atoms	6.0158	6.0085
4 atoms	7.4744	7.4605
5 atoms	9.1115	9.0954
6 atoms	11.0300	11.0055
7 atoms	12.6236	12.6022
8 atoms	13.9617	13.9459
9 atoms	15.2967	15.2804
10 atoms	16.5640	16.5458
11 atoms	17.6845	17.6703

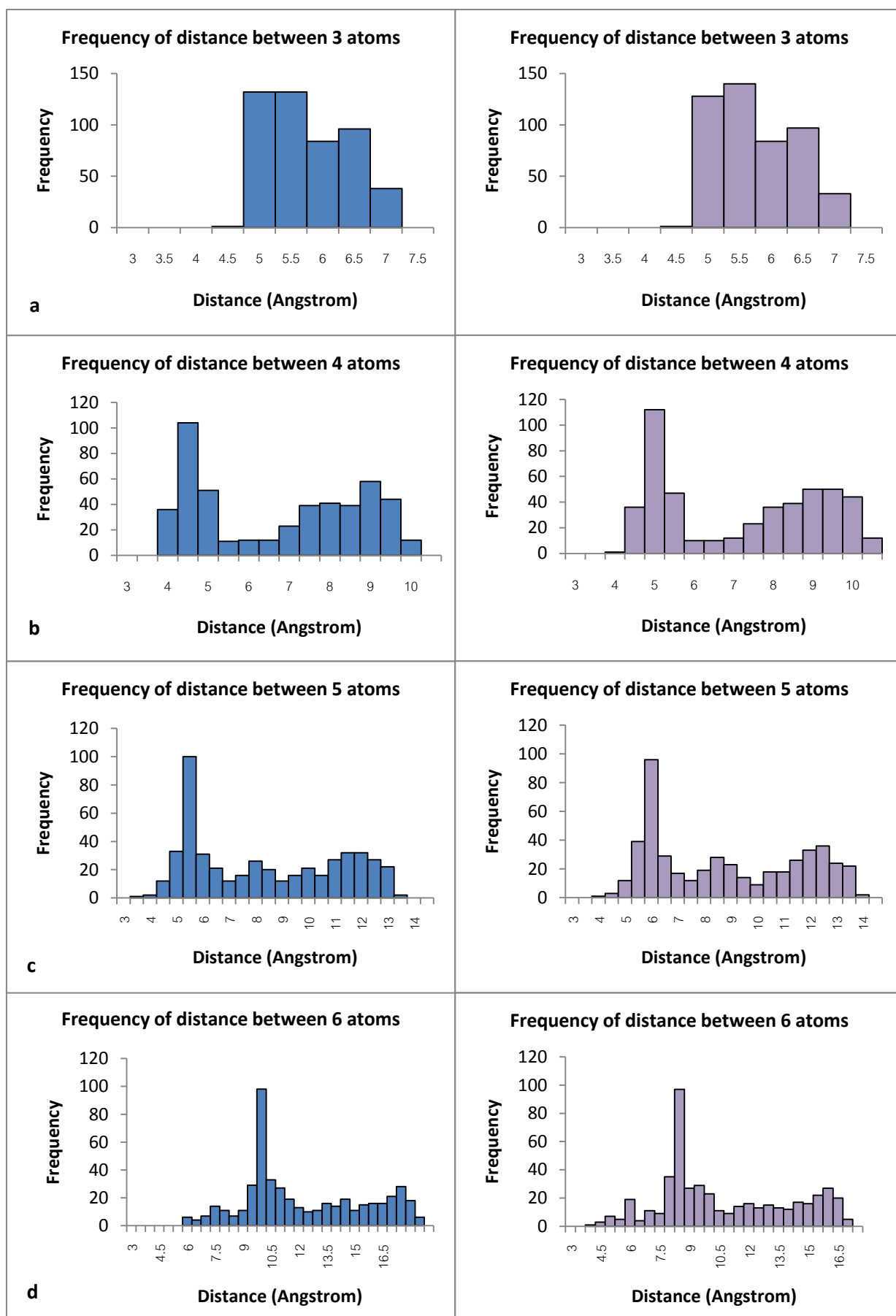


Figure1 Bar graph represented frequency of distance between atoms at 0.5 angstrom difference. The left site (blue) displayed chain A, while the right site (purple) displayed chain B.

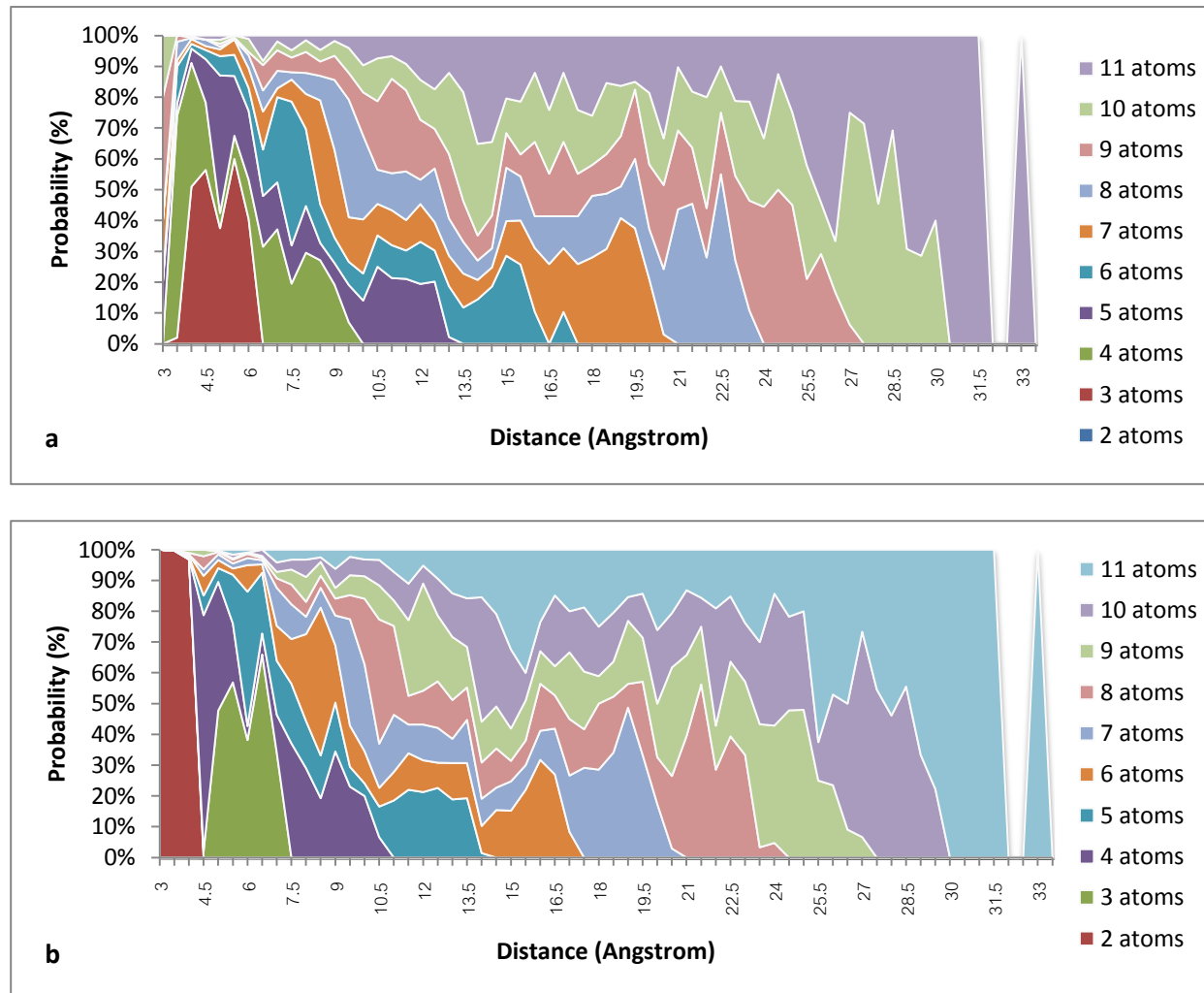


Figure2 Area curve demonstrated a probability which could be found alpha carbon atoms in distance period. An upper (a) and lower (b) ones were belonged to chain A and chain B of G6PD respectively.

- **Discussion**

By using Perl programming, the electron density map which represented by coordinated of each atom from PDB files of protein could be sorted. After they were computed and analyzed, the distances between a couple and density of alpha carbon atoms in range could use to calculate how many residues are in between these two residues. For example, it could be found 3 atoms for 40.4% in chain A and 5 atoms for 43.6 % in chain B of G6PD at the distance between 6-6.5 angstrom. As same as to the distance between 16 – 16.5 angstrom, it could be found 9 atoms for 24.1% in chain A and 7 or 11 atoms for 14.8 % in chain B of G6PD(Figure2a and b). So, it should be suggested that the gaps between segments, which 6 angstrom apart, probably had 3 or 5 atoms located in the distance.

Although some residues were far away from their couple by their residue number, generally, it could be nearby each other compared with their distance. For instance, while the average distances between 6 alpha carbon atoms were 11.0300 and 11.0055 angstrom, the distance that could be found them frequently were less than their average about 9.5 – 10.0 and 8.5 – 9.0 angstrom in chain A and chain B of G6Pd respectively (Figure1D). The reason was due to a secondary structure which peptide chain was folded as helix shape or beta plate sheet. Thus, there were many factors that influences to the position of next alpha carbon atom such as hydrogen bond, hydrophobic and hydrophilic interaction between side chains of amino acids(6). It should be concluded that electron density map of protein could be used to find out the alpha carbon atom during distance between each pair alpha carbon atoms.

- **References**

1. Lee MS, Bondugula R, Desai V, Zavaljevski N, Yeh IC, Wallqvist A, et al. PSPP: a protein structure prediction pipeline for computing clusters. *PloS one*. 2009;4(7):e6254. Epub 2009/07/17.
2. Langer G, Cohen SX, Lamzin VS, Perrakis A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nature protocols*. 2008;3(7):1171-9. Epub 2008/07/05.
3. Heuser P, Langer GG, Lamzin VS. Interpretation of very low resolution X-ray electron-density maps using core objects. *Acta crystallographica Section D, Biological crystallography*. 2009;65(Pt 7):690-6. Epub 2009/07/01.
4. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Jr., Brice MD, Rodgers JR, et al. The Protein Data Bank: a computer-based archival file for macromolecular structures. *Journal of molecular biology*. 1977;112(3):535-42. Epub 1977/05/25.
5. Rowland P, Basak AK, Gover S, Levy HR, Adams MJ. The three-dimensional structure of glucose 6-phosphate dehydrogenase from *Leuconostoc mesenteroides* refined at 2.0 Å resolution. *Structure*. 1994;2(11):1073-87. Epub 1994/11/15.
6. Sweet RM, Eisenberg D. Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *Journal of molecular biology*. 1983;171(4):479-88. Epub 1983/12/25.

- **Acknowledgements**

I would like to express my special thanks and appreciation to Dr. Victor S. Lamzin, Dr. Philipp Heuser and Tim Wiegels for their supervising, valuable guidance and many kindly helps throughout the DESY summer student program. Thanks for Dr. Martin Tolkiehn and Todd Laurus for their guidance in exercise week period. Professor Helmut Dosch, Dr. Olaf Behnke, Dr. Andrea Schrader, Dr. Doris Eckstein, Dr. Rainer Gehrke and people who are behind the summer student program are also sincerely acknowledged for this invaluable experience. Furthermore, it is my great honor to be one of the Thailand representatives which passed the national selection of NSTDA and NSRC under the patronage of HRH Princess Maha Chakri Sirindhorn.