

**Two factor correlation studies on
X-Ray Solution Scattering of
Biological Macromolecules**

**Josh Skodack
Michigan State University REU
Summer 2004
Advisor: William Wedemeyer**

Two factor correlation studies on X-Ray Solution Scattering of Biological Macromolecules

Josh Skodack

ABSTRACT

Traditional methods of determining biological macromolecule structure, such as NMR and X-Ray Crystallography have serious limitations that limit their use. It is believed that recent advances have made x-ray solution scattering applicable in a wide range of conditions for a larger number of macromolecules. I used a new statistical model to interpret x-ray solution scattering patterns of a small protein molecule. It is hoped that by measuring the correlation between the scattering patterns of different sites will allow for more accurate, precise and easier structure determination.

INTRODUCTION

Since the beginning of the post-genomic era, there has been a need to determine the structural properties of biological macromolecules. Traditional methods of determining the structure includes high-resolution techniques in the angstrom range, such as X-Ray crystallography and Nuclear Magnetic Resonance (NMR). However, both of these techniques have their limitations. X-Ray crystallography requires high quality protein crystals while NMR requires small soluble proteins. However, X-Ray Solution scattering is applicable for a wide range of proteins in a number of conditions. In X-ray

solution scattering, a target protein scatters a coherent beam of x-rays. Diffracted x-ray intensity is recorded as a function of the scattering angle (Bailey-Kellog, et al., 2003).

Currently, the true resolution of solution scattering is unknown. I am pursuing a study of the resolution limits of this technique. To determine the resolution limits a series of mutants of Protein L are prepared and we compare the conformational changes from a single change in the amino acid code. The Protein L mutants I examine in this study are Protein L K23C and Protein L K42C (O'neil, et. al.). The two mutants contain a mutation at the 23rd and the 42nd residue respectively.

Protein L is a small 7 kDa protein consisting of 62 residues. The structure of protein L as determined by NMR consists of a four-stranded β -sheet packed against a single α -helix. The order of the packing of Wild Type Protein L is $\beta\beta\alpha\beta\beta$ (figure 1). The two β -turns are diametrically opposed and this allows for ideal experimental comparison between the Protein L variants (Gu, et al., 1997).

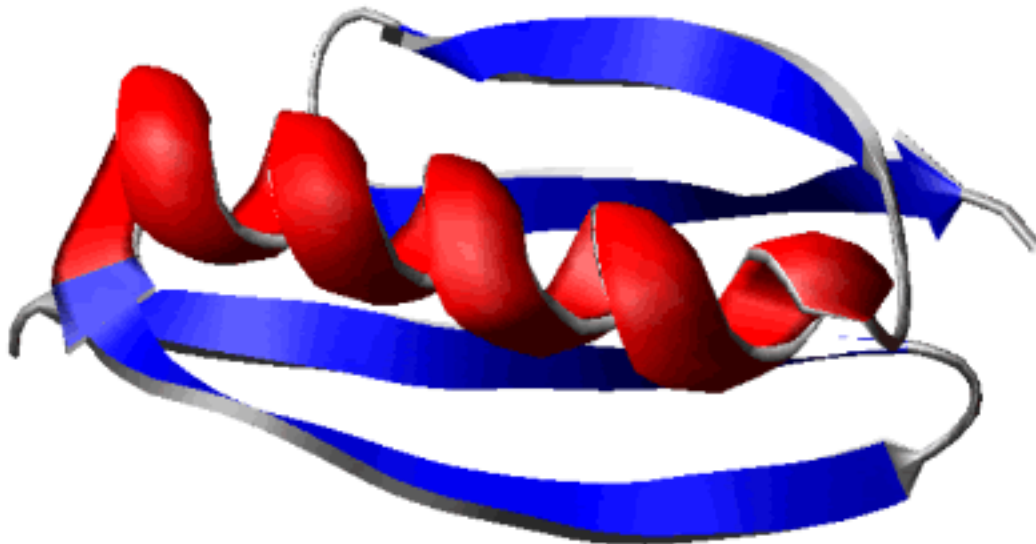


Figure 1: Wild Type Protein L with bbabb structure. (PDB)

Materials and Methods

The Protein L DNA variants, K23C and K42C were obtained from the Michigan State University Genomics Center. Mutants were designed using the Quick Change site directed mutantgenesis kit (Stratagene) using the suggested protocol and primer design program and the pET15-b plasmid as the template DNA. Mutant DNA was amplified through PCR.

To prepare each separate Protein L mutant, Protein L was cloned into pET 15b display vector (Gu et al., 1995) between the DPN I restriction enzyme sites. At these two sites the enzyme DPN I cuts the both the plasmid vector and the Protein L DNA. The Protein L DNA is inserted into the plasmid vector. The mutant DNA-plasmid complexes transform into competent BL21 E. coli cells. To insure only e-coli cells with the plasmid complex were allowed to reproduce; the cells were screened for resistance to the Carbenicillin antibiotic.

To isolate Protein L, a starter culture of cells with the protein L variant and plasmid complex was grown in LB-Broth and 1xCarbenicillin. A one-liter solution of LB-1xCarbencillin-1 M glucose was inoculated with the starter culture and incubated for four hours. Protein L production was induced in the culture following incubation by addition of 1mM IPTG to the culture. The culture was allowed to incubate for four hours and then cells were recovered by centrifugation. The recovered cells are lysed by addition of 1xbinding buffer, 1x PMSF and 1x benzamidine. Cell debris, DNA, RNA is removed from the solution by decanting the supernant. Protein L is further purified by column chromatography. Protein L comes off the column, on average, between 40 mM and

50mM Imidazole Buffer. Gel electrophoresis of all column fractions is run on 10% SDS-Page gel to determine relative protein concentration.

To insure a measurable signal from solution scattering, the protein L K23C and protein L K42C mutants were labeled by tagging the cystein residue with mercury at their respective mutant residue sites. Concentrating the protein solutions to a final concentration of 25 mg/ml in 50 mM, Imidazole Buffer reduces the Background noise from solvent particles.

Solution scattering was performed using a rotating anode tube with Cu K_{α} nonpolarized radiation at an average wavelength of 1.542 Å. The scattered X-Ray radiation was collected using the Rigaku R-AXIS II system. William Wedemeyer writes software used for solution scattering. The diffracted X-Ray radiation was collected for each mutant as a function of intensity (I) versus scattering angle (s). Background noise from solvent particles and from the surrounding environment is further eliminated by subtracting the Intensity versus scattering angle of 50 mM Imidazole buffer solution from the Intensity measurement of each mutant.

DISCUSSION

The basic theory relies on assuming a single reflection of the incoming X-ray plane wave with an x-ray scattering particle such as an electron. Measuring differences in reflected radiation Intensity vs. scattering angle and comparing it to a mutation of the same protein with a tag on a nearby residue can estimate the distance between the two residues. Large-scale integration of this technique will allow much more accurate and precise model. An incoming x-ray plane wave $f(\mathbf{r}) = A_0 e^{i\mathbf{k} \cdot \mathbf{r}}$, with initial amplitude A_0 , wave vector $\mathbf{k} = \mathbf{k}_0 = 2\pi/\lambda$ is incident on a protein and the atoms within the protein become

sources of spherical waves. In this study it was assumed that only elastic scattering occurs, so the modulus of the scattered wave $k_1=k_0$. The amplitude of the scattered wave is described by the scattering length f . The scattering length with X-rays interacting with electrons is $f=N_0r_0$, where N_0 is the number of electrons and $r_0 = 2.82e-13 \text{ cm}$ is the Thomson radius. The scattering is described in the Fourier Transform from the real space of \mathbf{r} to the reciprocal space of the scattering vectors $\mathbf{q} = \mathbf{k}_1 - \mathbf{k}_0$ (Svergun).

The scattered ray is described by the function of the momentum transform, $f(\mathbf{r}) = A_0\rho_e g(\mathbf{r})$ where ρ_e is the electron density and $g(\mathbf{r})$ is the probability for scattering. In the reciprocal space (figure 3) \mathbf{q} is the difference between scattered and the incoming ray wave number, $\mathbf{q} = \mathbf{k}_1 - \mathbf{k}_0$. The total amplitude for scattering in reciprocal space will be $G(\mathbf{q}) = A_0\rho_e \int (d\mathbf{r} * g(\mathbf{r}) * e^{-i(\mathbf{k}-\mathbf{k}_0)\mathbf{r}}$, where $g(\mathbf{r})$ is the difference in the scattering probability from the protein $\rho(\mathbf{r})$ and the solvent $\rho_s(\mathbf{r})$.

The physical distance between two scattering events can be derived from the correlation between the two scattering events (Haag). In statistical mechanics the correlation between two events is the expected product of the two data values, $X(\mathbf{r})$ and $Y(\mathbf{r})$, is $Z'(\mathbf{q}) = X'(\mathbf{q})Y'(\mathbf{q})$. If $Z'(\mathbf{q}) = X'(\mathbf{q})Y'(\mathbf{q})$ in the phase domain, then the Fourier transform of the correlation between the two scattering events in the spatial domain is $z(\mathbf{r}) = \int (d\mathbf{r}' x(\mathbf{r}') y(\mathbf{r}-\mathbf{r}'))$. Similarly, if either $X(\mathbf{r})$ or $Y(\mathbf{r})$ are even functions, then $z(\mathbf{r}) = \int (d\mathbf{r}' x(\mathbf{r}') y(\mathbf{r}+\mathbf{r}'))$.

In scattering experiments, the amplitude $G(\mathbf{q})$ cannot be directly measured. Only the intensity of scattered photons, $H(\mathbf{q}) = G(\mathbf{q})G(\mathbf{q})^*$, in a general direction \mathbf{s} can be directly measured.

In the examination of the scattering events between two proteins differing only in the position of a mercury tag, let $H(\mathbf{q})$ be defined as the intensity of the scattering events of $G(\mathbf{q})$ of mutant K23C and $G'(\mathbf{q})$ of mutant K42C with a time separation of $t = \infty$. The autocorrelation between the two events of scattering is then $h(\mathbf{r}) = \int (d\mathbf{r}' g(\mathbf{r}') g(\mathbf{r}' + \mathbf{r}))$. Unfortunately, this model will result in a “smeared” pattern of peaks because it will give a peak whenever the radius of the scattering Ewald sphere is equal to $r_{\text{atom}, k}$. A correction factor is added to give sharp distinct peaks in the autocorrelation function.

If we model the scattering probability $g(\mathbf{r})$ as a delta point particle that goes to infinity as $\mathbf{r} = \mathbf{r}_{\text{atom}}$ and $g(\mathbf{r}) = \sum \delta(\mathbf{r} - \mathbf{r}_{\text{atom}, k})$, then this will give distinct peaks in $h(\mathbf{r})$ whenever $\gamma = \mathbf{r}_{\text{atom}, k} - \mathbf{r}_{\text{atom}, l}$.

RESULTS

Initially, Protein L samples with a mutation at the 23rd and 42nd residues were prepared for investigation. The samples prepared did not contain a mercury tag at the 23rd and 42nd residues.

The first samples tested were Protein L K23A and Protein L K42A. The two samples resulted in null result. We believe that no result was achieved because of a low electron density of the target 23rd and 42nd residues of Protein L. The resultant scattering pattern showed no change between Protein L K23 and Protein L K42. Peaks from background water molecules further distorted the image. It was decided that protein L samples tagged with mercury would increase the probability of scattering at the 23rd and 42nd residues.

Currently, both protein samples of mercury tagged Protein L K23C and Protein L K42C are nearing completion. Protein L K23C and Protein L K42C have been purified

through column electrophoresis and will have the 23rd and 42nd residues tagged shortly with mercury. Following tagging, the protein solutions will be concentrated to a concentration of 25 mg/ml of protein in solution.

Preliminary data will be taken on the R-AXIS to determine early resolution limits and overall sample quality. If sharp distinct peaks are seen in the autocorrelation function from the two intensity patterns of protein L K23C and K42C, then further data collection will be taken using the advance photon source at Argonne National Laboratory later this month.

ACKNOWLEDGEMENTS

This work is supported by discussions with William Wedemeyer, Russell LeClain, and Terry Ball. A NSF Research Experience for Undergraduate fellowship provided financial support.

REFERENCES

Kuhlman B., O'Neill J. W., Kim D. E., Zhang K. Y. J., Baker D. (2002). Accurate Computer-based Design of a New Backbone Conformation in the Second Turn of Protein L. *J. Mol. Biol.* **315**, 471-477

O'Neil J. W., Kim D. E., Baker D., Zhang K. Y. J. (2001) Structures of the B1 domain of protein L from *Peptostreptococcus magnus* with a tyrosine to tryptophan substitution.

Acta. Cryst. D. **57**, 480-487

Pillardy J., Czaplewski C., Liwo A., Wedemeyer W., Lee J., et al. (2001) Development of Physics-Based Energy Functions that Predict Medium-Resolution Structures for Proteins

of the α , β , and α/β Structural Classes. *J. Phys. Chem. B.* **105**, 7299-7311

Svergun D. L., Petoukhov M. V., Koch M. H. J. (2001) Determination of Domain

Structure of Proteins from X-Ray Solution Scattering. *Biophysical Journal.* **80**, 2946-

2953

Svergun D. L., Koch M. H. J. (2003) Small-angle scattering studies of biological

macromolecules in solution. *Rep. Prog. Phys.* **66**, 1735–1782

Drenth J. (1994) *Principles of Protein X-Ray Crystallography*, Springer-Verlag, New York.

Gu, H., Yi Q., Bray S. T. , Riddle D. S., Shiau A. K., Baker D. (1995) A phage display system for studying the sequence determinants of protein folding. *Protein Sci.* **4**, 1108-

1117.