Modeling Protein Structures and their Complexes with Sparse Experimental Data

Dominik Gront, Maciej Błaszczyk, Jacek Wabik, and Andrzej Kolinski

University of Warsaw, Faculty of Chemistry, Pasteura 1, 02-093 Warsaw, Poland
E-mail: {dgront, mblaszczyk, jwabik, kolinski}@chem.uw.edu.pl

BioShell project has been started in 2005 as a set of stand-alone programs aimed on simplification of typical bioinformatics tasks. Since then it has evolved to become a fully featured scripting language for biomolecular modeling and structural bioinformatics. Most recently, the development of the package is focused on incorporating various types of experimental data into protocols for structure prediction of proteins and their complexes. In this work we present an application of Small Angle Xray Scattering (SAXS) profiles to the determination of mutual domain orientation in multi-domain proteins. Preliminary results suggest that the scattering data can be successfully used in studies of large macromolecular assemblies providing that the structure of the individual interacting partners are known.

1 Introduction

One of the most important challenges in modern structural biology is the characterization of multi-domain macromolecular complexes that govern a major part of important cellular functions. These large biomachines are very difficult targets for standard experimental methods. Due to their size and flexibility investigation with X-ray crystallography or NMR spectroscopy is not always easily accomplishable. Multidisciplinary methods are amongst the most promising approaches. Structures of separate protein domains, that have been previously determined with Xray or NMR methods may be properly combined into whole complexes with the help of Electron Microscopy (EM) or Small Angle Xray Scattering (SAXS) data. SAXS profile is a function of all atomic coordinates for a given system and provides only an averaged description of macromolecular size and shape encoded in a very synthetic way. Informational entropy analysis suggests that a scattering profile may be used to determine only several independent degrees of freedom. The data however is considered to be satisfactory for the unique definition of geometry of a macromolecular assembly.

SAXS data has been already incorporated into many modeling platforms, e.g. NIH-Xplor[1] or ATSAS[2]. Here we describe its successful combination with the BioShell modeling platform. BioShell[3] package has been originally created for structural bioinformatics[4]. The suite of programs was growing and new functionalities have been emerging. After several years of development the package is capable to deliver all the necessary modeling routines[5]. BioShell routines may be called from any programming language that operates within Java Virtual Machine (JVM), most notably from Java implementations of Python (jython) and Rubby (jRuby) as well as from Java itself. This makes it a very versatile platform that may be quickly applied to various modeling projects.

2 Materials & Methods

In this contribution SAXS data has been used in determination of three-dimensional (3D) structure of two-domain proteins, provided that the high-resolution structures of both domains are known. Conformation of the linker that connects domains has been subjected to conformational sampling while the structure of each of the domains has not been altered. SAXS data has been applied as the only scoring term to guide the walk in the conformational space towards the correct geometry.

2.1 Conformational Sampling

CartesianProteinSystem module of the BioShell package has been used for conformational sampling. Due to its very general and careful design, a wide range of mover objects are available for introducing conformational changes. In the course of this work, RandomDihedralMover has been used to modify Φ,Ψ dihedral angles at randomly selected position in the linker region. The mover object proposes new Φ,Ψ coordinates according to an empirical probability distribution that has been obtained from loop conformations extracted from a non-redundant set of high-resolution protein structures. A trial conformation has been accepted or rejected according to the Metropolis criterion with SAXS-based χ^2 statistics. Simulated annealing Monte Carlo protocol was used to generate short trajectories, starting from a random conformation of the linker. The structure of the lowest energy (i.e. of the best χ^2 fit) has been reported from each trajectory.

2.2 SAXS-based Model Assessment

In this work we followed the frequently used approach to simulate SAXS intensity $I(q)$ at any arbitrary scattering vector length q with the Debye formula:

$$I(q) = \sum_{i=1}^{N_{at}} \sum_{j=1}^{N_{at}} f_i(q) f_j(q) \frac{sin(qd_{ij})}{qd_{ij}} \tag{1}$$

where d_{ij} is the distance between i-th and j-th atom of the molecule. Atomic form factors $f(q)$ were properly corrected to reflect the effect of displaced solvent. Following the work by Fraser et al.[6], dummy solvent atoms were placed at all atomic positions within the macromolecule with the form factors computed from on average electronic density of bulk water. The Debye formula was employed in computations of a theoretical SAXS spectrum for each trial protein conformation, i.e. at every Monte Carlo move. The quality of the fit between the experimental scattering data and those predicted from the models is described by the χ^2 statistics over the set of N_q values:

$$\chi^2 = \frac{1}{N_q - 1} \sum_{k=1}^{N_q} \Big[\frac{I_{\text{reference}}(q) - I_{\text{model}}(q)}{\sigma(q)} \Big] \tag{2}$$

PDB id	N_{res}	linker residues from	to
1a62A	125	42	47
1d09B	153	94	100
1mgtA	169	49	61
1ammA	174	79	88
1nkrA	195	96	102
1knyA	253	123	127
1ctuA	294	170	188
1j8mF	295	83	95
1jpnA	296	86	98
1ca1A	370	244	256
1bagA	425	344	349
1eovA	487	128	137

Table 1. Benchmark set used in this study. Each of these proteins comprises two domains. Conformational sampling has been restricted to the inter-domain linker region defined in the table.

3 Results

A benchmark set of twelve proteins (summarized in Tab. 1) have been utilized to test the protocol. These targets have been chosen to cover the typical range of polypeptide chain lengths, at the same time providing representatives for all the major protein architectures. For each of these proteins 50 000 structural models have been calculated. Results have been summarized in the Fig. 1 where each dot represents a single model. The plotted range of the coordinate root mean square deviation (crmsd) between a model and the native

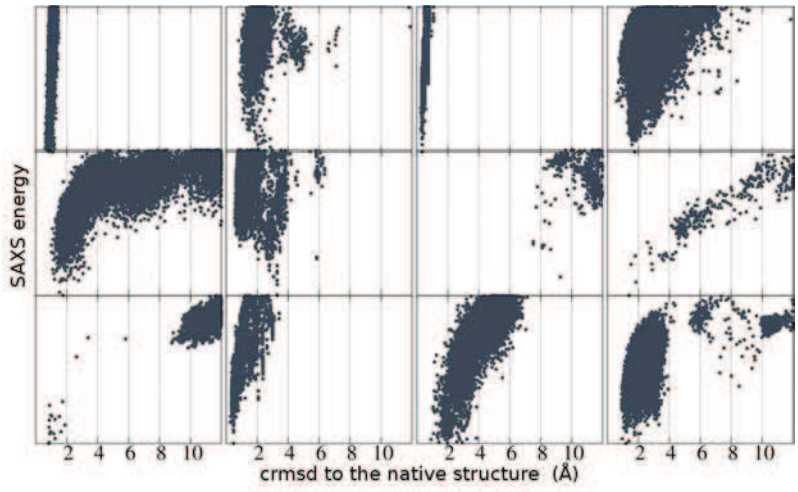

Figure 1. Scatterplots that show modeling results for the twelve proteins in the benchmark set. SAXS energy (Y axis, arbitrary units) is plotted as a function of crsmd (from 0.0 to 12.0 Å in each of the scatter plots).

structure (X axis) has been set to [0.0 Å,12.0 Å] although for some of the cases much larger values have been observed. Since SAXS scattering intensity $I(q)$ grows quadratically with the number of atoms in the scattering system, χ^2 statistics also varies greatly with the size of a target protein. Therefore the SAXS energy values on the Y axis are shown in arbitrary units, scaled separately for each box of the multipart plot. In all but one of the test cases (`1d09B`), SAXS energy decreases as the conformational sampling is approaching the native conformation according to a funnel-like dependence. Additionally, for two other cases (`1eovA` and `1j8mF`) conformational sampling turned out to be inefficient in the proximity of the native structure. Domain packing in the two other test cases: `1a62A` and `1mgtA` is relatively tight which explains the fact that the sampling process yielded mostly very good structures. Some conformations that otherwise could have low SAXS energy were excluded due to steric clashes.

4 Conclusions

From the perspective of typical biomolecular modeling methods, the test systems presented in this contribution should be considered as quite large. The problem has been made computationally tractable by freezing all but a small fraction of the degrees of freedom. Only from 12 to 36 main chain dihedral angles have been subjected to conformational sampling. This however granted enough flexibility to the protein chain to sample the whole space of mutual orientations between the domains. In general, SAXS - based score has been able to pinpoint the correct conformation. It should be also mentioned that even very small deviation in domain orientation (e.g. a tiny rotation of one of the domains in respect to the other one) may result in relatively large crmsd value.

Acknowledgments

Support from Marie Curie fellowship (FP7-people-IOF) for DG is acknowledged.

References

1. A. Grishaev, and J. Wu, and J. Trewhella, and A. Bax *Refinement of Multidomain Protein Structures by Combination of Solution Small-Angle X-ray Scattering and NMR Data* J. Am. Chem. Soc **127**, 16621-16628, 2005.
2. M. Petoukhov, and P. Konarev, A. Kikhney, and D. Svergun *ATSAS 2.1 - towards automated and web-supported small-angle scattering data analysis*. J. Appl. Cryst. **40**, 223-228, 2007.
3. `http://www.bioshell.pl`
4. D. Gront, and A. Kolinski *BioShell - a package of tools for structural biology computations*, Bioinformatics **22**, 621-622, 2006.
5. D. Gront, and A. Kolinski *Utility library for structural bioinformatics*, Bioinformatics **24**, 584-585, 2008.
6. R. Fraser,and T. Macrae,and E. Suzuki *TITLE GOES HERE* J. Appl. Crystallogr **11**, 693-694, 1978.