Application of reduced space modeling and sparse experimental restraints to structure determination of proteins and protein assemblies

research project realized by Dominik Gront*

Part I

Outgoing phase

1 Summary of the project objectives

The main objectives for this stage of the project were to:

- learn the research methodology as practiced by the outgoing laboratory, in particular the Rosetta modeling approach
- bridge Rosetta package with BioShell, a modeling suite that had been developed by the fellow
- develop a novel protein structure modeling protocol, based on a combination of Rosetta with other, coarse-grained methods, that would be able to utilize a very wide range of sparse and inaccurate experimental data

2 Main results

Rosetta software, which has been continuously developed in the outgoing host laboratory for over fifteen years, builds protein models from short (3 and 9 amino acid long) fragments extracted from structures already known (i.e. already studied experimentally). The fragments play the central role in Rosetta modeling methodology. Within the first few months of the project it became clear that tinkering with the fragment picking module would be the best way to achieve most of the project's objectives. This however appeared to be a very difficult task. In fact, the then-existing fragment picking module was the oldest part of Rosetta still in use and seemed to be the weakest part of the Rosetta package. All the other algorithms had been already updated or replaced with novel versions. Development of the new method would ensure the highest gain in terms of scientific quality of all the aspects related to the project. Therefore the new fragment picking method has been designed, implemented and tested.

Thanks to the very careful object-oriented design, the program is not only a stand alone executable, but also a very versatile software library that can be easily extended to accommodate future applications. Scientific results as well as extensive benchmark of the new method have been already published[1]. The developed software is publicly available as a part of the Rosetta† distribution. Moreover, a new version of BioShell package‡ has been released during this stage of the project. The published version allows package users conveniently process Rosetta input data and computed results. It had played an important role in efficiently conducting large scale computations that had been necessary for the project.

The new fragment picker offers a lot of new functionality. It accomodates various kind of experimental data already at the stage of fragments selection, even before the main simulations starts. In fact, the implementation of the experimental data was one of the most challenging parts of the project. This feature is a very important improvement in respect to the field of biomolecular modeling with experimental data. Approaches used so far utilize experimental data as a part of their scoring system, e.g. as restraints. The modeling process - usually a Monte Carlo or a Molecular Dynamic simulation - is pushed towards the correct solution because incorrect ones were penalized. The penalty function is realized as a scoring term that measures the (dis)agreement between the current conformation and the experimental results

* dgront@chem.uw.edu.pl † http://www.rosettacommons.org/ ‡ http://www.bioshell.pl

provided as input data to the method. Usually such an approach yields more accurate results than simulations without the experimental data. However, in majority of cases the penalty function hampers the sampling process which makes the simulation less efficient. On the contrary, the approach developed in the course of this project removes any structural fragment that do not comply to the data before the simulation. This greatly reduces the search space that has to be sampled, eliminates errors and improves modeling accuracy.

Currently, the software can utilize the following types of experimental data:

- backbone chemical shifts (from solution or solid state NMR)
- NOE restraints (from solution NMR)
- disulfide bond locations (known from protein mass spectrometry or from NMR experiments)

The work is still continued and additional types of experimental data will be added. Current development focuses on Solid State NMR measurements and Residual Dipolar Couplings.

The second part of the project was to implement global long range experimental data into the Rosetta modeling protocols. Data of this kind depends on the overall geometry of the whole molecule rather than on the position of particular atoms. Because Rosetta protocols utilize only short fragments (now from 3 to 15 amino acid long) - much shorter than the modeled protein, such global data cannot be used for fragment selection. This information therefore may only be used during simulation as a part of a scoring function.

A new scoring term which employs Small Angle Xray Scattering (SAXS) profiles to assesses a protein conformation has been designed and implemented in Rosetta. In order to finalize this objective it was necessary to:

- choose and implement in Rosetta the most suitable algorithm for computing SAXS spectra. After preliminary research, Debye equation has been chosen for this purpose due to its simplicity and possible integration with other energy-related procedures already present in the software
- derive pseudoatomic form factors for Rosetta reduced representation. Form factors, which describe scattering of Xray radiation of any known atom type are the necessary parameters for the Debye equation. Since the reduced representation of Rosetta involves artificial pseudoatoms (each of them substitutes one amino acid residue side chain), a custom form factor function had to be derived for each of them.
- optimize the numeric procedure for Debye calculations to achieve the necessary computational efficiency. A novel algorithm has been designed for this purpose.

Currently the SAXS modeling protocol is extensively tested.

3 Implications of the project

The main achievement of the project was to develop a novel fragment picking process driven by experimental data. The software is currently used in more than twenty research laboratories all over the world.

Part II

Integration phase

4 Summary of the project objectives

The main objectives for this stage of the project were to:

- conduct large-scale simulations with available experimental data that would show the applicability of the research methodology developed by the fellow
- exchange the research experience and the new knowledge learned during the outgoing phase

5 Main results

The fellow introduced the developed software to the computational research groups affiliated at the return host institution. For instance, he conducted a tutorial session to teach PhD student on using the software. Thanks to his initiative, the University of Warsaw became a member of the Rosetta Commons - a group that gathers all the research laboratories that work on Rosetta software development. As a consequence, the institution gained access to the full version of the software. It also creates a framework for the future collaborations with other groups that belongs to the Rosetta Commons.

During this period several new modules have been implemented in the Rosetta package. As a result of extensive benchmarks conducted by the fellow, the module for fragment picking has been updated and fine tuned. Moreover, the fellow has conducted extensive simulations on a benchmark set of proteins that utilizes SAXS experimental restraints. The results show that in general the assumed approach may be very useful in the determination of quaternary structure of protein complexes.

During the integration phase, BioShell software package has been greatly updated and a new version released. The work on package has been started in 2005 as a part of the fellow's PhD projects. The software is still actively developed and used by several research groups in Poland. The collaboration with the outgoing host resulted in numerous addition to the package. The new applications include:

- automated compilation and testing server, which greatly facilitates the effort of several programmers that work on the project; the employed solutions are inspired to the methods used in Rosetta project
- fold recognition (threading) application, based on one-dimensional threading algorithm i.e. on a sequence alignment of sequence profiles combined with structure profiles

Finally, the Rosetta methodology has been applied by the fellow in his current research projects. Most notably, the fellow uses the software to *in-silico* predict the effects of single point mutations in enzymes.

6 Implications of the project

The main achievement of this phase was to propagate the Rosetta research methodology at the University of Warsaw, in particular at the Faculty of Chemistry. The results obtained by the fellow were also promoted by his participation in numerous conferences, either as oral presentations as well as by posters.

Part III

Summary and perspectives

The fragment picking algorithm developed by the fellow became the indispensable part of Rosetta methodology. It greatly facilitates introduction of various kinds of experimental data. Many of them have been utilized in the course of the project. Others are still tested in ongoing research. The Rosetta methodology has been introduced to a few laboratories working at the host's institutions and is routinely used in their studies. The experience gained by the applicant will certainly have very positive influence on his future research career.

References

[1] Gront D, Kulp DW, Vernon RM, Strauss CEM and Baker D, *"Generalized fragment picking in Rosetta: design, protocols and applications"*, PLoS ONE, 2011; 6(8): e23294.