Evolutionary refinement of the 3D structure of multi-domain protein complexes from Small Angle X-ray Scattering data

Olga Rudenko
Synchrotron Soleil
Paris area, France
roudenko@synchrotron-soleil.fr

Aurelien Thureau
Synchrotron Soleil
Paris area, France
thureau@synchrotron-soleil.fr

Javier Perez
Synchrotron Soleil
Paris area, France
perez@synchrotron-soleil.fr

ABSTRACT

DADIMODO is an Evolutionary Algorithm based software program for solving an inverse problem arising from biological Small Angle X-ray Scattering (SAXS) data analysis. The problem consists in refining the three-dimensional model of a multi-domain protein complex against SAXS experimental data. We use an "all-atoms" structure representation allowing for an energy control for every newly generated model so as to prevent steric clashes and converge to a physically feasible structure. Sophisticated variation operators had to be specifically tailored for this application and constitute our optimization algorithm's particularity. Atomic structure manipulations being computationally expensive, we report the critical implementation on a computer cluster for parallelized feasibility check and evaluation stage. DADIMODO has been successfully used in many biostructural data analysis cases and was recently made available to the corresponding research community as a web service.

CCS CONCEPTS

• **Applied computing** → *Computational biology*; *Molecular structural biology*; • **Theory of computation** → *Evolutionary algorithms*;

KEYWORDS

protein structure, biological small angle X-ray scattering (SAXS), experimental data analysis, evolutionary algorithm

ACM Reference Format:
Olga Rudenko, Aurelien Thureau, and Javier Perez. 2019. Evolutionary refinement of the 3D structure of multi-domain protein complexes from Small Angle X-ray Scattering data. In *Genetic and Evolutionary Computation Conference Companion (GECCO '19 Companion), July 13–17, 2019, Prague, Czech Republic.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3319619.3322002

1 INTRODUCTION

Proteins are large biomolecules consisting of long chains of amino acid residues. They are involved in important processes such as metabolic reactions, DNA replication *etc.* Protein three-dimensional structure determination gives access to important information regarding their biological functions. Small Angle X-ray Scattering (SAXS) is a widely used experimental technique in structural molecular biology allowing to study protein complexes in solution.

DADIMODO, a software program for biomolecular structure refinement from SAXS, takes its roots from [2, 5]. It is based on "all-atoms" structure representation which is computationally more costly than the often used coarse-grain representation but has the advantage of preserving the physical feasibility of solutions. A very important improvement since [2] is the automated configuration of the variation operators based on a graph representation of the structure (with rigid domains as nodes and flexible fragments as edges). Previously tedious input configuration thus became accessible to external users, making DADIMODO evolve from an "expert only" tool to a service open to the corresponding scientific community. Additionally, an adaptive mechanism for the previously static mutation radius was introduced to make the search process more robust. The latest DADIMODO version is based on a distributed EA implementation using the DEAP library [3] and recently became available as a web service [6].

2 STRUCTURAL INFORMATION FROM SAXS

Protein chains are commonly organized in "rigid" (energetically stable) domains connected by relatively flexible chain fragments (Fig. 1). By illuminating samples with X-rays and studying the scattered radiation, SAXS is used specifically to guess the 3D structure of the flexible parts while rigid domains structure is typically determined using biocrystallography. Biological SAXS provides information about the size and shape of molecules with a resolution notably lower than the atomic level. The existing programs for SAXS data fitting tend to reflect this fact by proposing variants of a coarse-grain model representation, which may sometimes lead to physically unacceptable structures. The complete all-atoms structure representation adopted in our case needs longer calculations but allows the integration of an energy constraint preventing from potential unphysical solutions.

2.1 Evolutionary SAXS data analysis

In the present work, the degrees of freedom of a 3D structure model adjusted from the SAXS data are the dihedral angles (ϕ, ψ) of all the residues composing the flexible peptide chains fragments (typically several tens).

2.1.1 Variation operators. No crossover operator is used. To introduce mutations, we distinguish four types of flexible fragments: linkers, terminal parts, double linkers and loops (Fig. 1). Linkers

GECCO '19 Companion, July 13–17, 2019, Prague, Czech Republic

ACM ISBN 978-1-4503-6748-6/19/07…$15.00
https://doi.org/10.1145/3319619.3322002

and terminal parts are modified in the same way: they are folded by adding some $(\Delta\phi, \Delta\psi)$ to the dihedral angles of one randomly chosen residue. When dealing with a double linker, it is physically impossible to modify one of its parts without applying a correlated modification to the other. A rotation around an axis passing by a randomly chosen C_α atom on each of its parts is applied (Fig. 1 right). Clearly, such transformation can also be applied to loops.

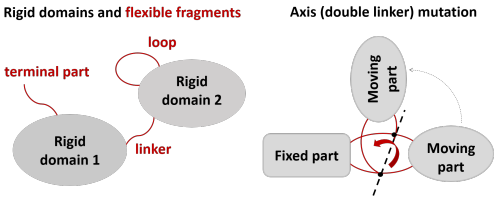

Figure 1: *DADIMODO*: **search space topology**

2.1.2 Energy constraint. Molecular mechanics assume the steric energy of a molecule to arise from a few, specific interactions such as stretching/compressing of bonds, torsional effects, attractions/repulsions of atoms, and electrostatic interactions. The MMTK software [4] is used to calculate the new 3D coordinates of all atoms moved by rotation. Since the axis mutation tends to "break" the peptide chains in unnatural ways, MMTK also "restores stereochemistry" by modifying atom positions in neighboring residues. Such local adjustments are part of the mutation operator. After each rotation, the steric energy terms are calculated and each of them is required not to exceed twice (heuristic choice [2]) the corresponding energy term of the initial structure. New structures are obtained either by one randomly chosen (ϕ, ψ) linker mutation or by a random axis rotation. We do not mutate several residues at once since even after a single rotation, a feasible child is not guaranteed. If the mutation yields an energetically unacceptable structure three times, the parent structure is duplicated since no feasibilty fixing mechanism is available.

2.1.3 Adaptive mutation radius. Tuning the initial mutation radius is both important and difficult: it shall be large enough to overcome energetic barriers yet small enough to avoid high infeasibility rates. It is also highly specific to the problem instance and unintuitive for external users. An adaptation mechanism (absent in [2]) was introduced, and was observed to soften the consequences of inadequate choices of the maximum rotation angle. A variant of the 1/5 success rule is used since the state-of-the-art self-adaptation mechanisms that suppose modification of the step-size along with the individual at each evaluation are not applicable to our structures folded once at a time as explained in sect. 2.1.2.

2.1.4 Model evaluation. A *SAXS profile* (Fig.2) is a curve which represents circularly averaged scattering intensity as a function of a scattering vector q (normalized scattering angle). The *Crysol* program [7] calculates a simulated SAXS profile given the atomic structure of a biomolecule. The weighted sum of the squared differences between observed and calculated intensities (SAXS curves) is a model evaluation criterion for SAXS data analysis:

$$\chi^2 = \frac{1}{N-1} \sum \frac{(I_{obs}(q) - I_{calc}(q))^2}{\sigma^2(q)} \quad (1)$$

3 CASE STUDY

The study of Mycobacterium tuberculosis DNA gyrase [1] is one among several published biostructural data analysis cases for which DADIMODO has been successfully used. In figure 2(left) is presented an Mtb gyrase model resulting from a crystallographic study and completed using a molecular modeling tool. Typically, SAXS data analysis starts with one structure model. In order to provide our evolutionary algorithm with an initial population, we randomize around it with the aforementioned variation operators. A (10+20) ES scheme was used on one cluster node with 20 processors. 200 generations take for this (large) structure about 20 hours.

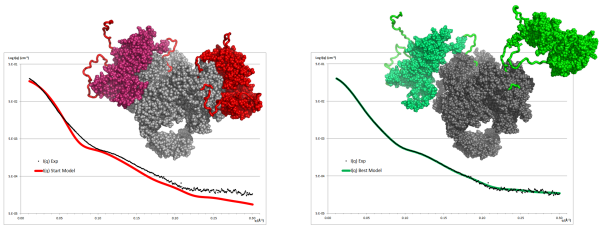

Figure 2: Structure model before ($\chi^2 = 45.24$) and after evolutionary SAXS refinement ($\chi^2 = 1.63$) with corresponding simulated vs. experimental SAXS curves

One of the best models produced by DADIMODO is represented in figure 2(right), indeed showing a very good agreement between the simulated and experimental SAXS curves. Taking SAXS averaging into account (sect. 2.1.4) and the fact that flexible protein fragments tend to oscillate in solution, the resulting model cannot be interpreted as an exact protein conformation but as the "most probable" conformational state of a protein molecule in solution. Structurally speaking, our result, that roughly characterizes the position of the smaller (green) domains respectively to the large (gray) one, contributes an important piece of information to a complex structural study suggesting a conformational state which might be targeted for drug discovery.

REFERENCES

[1] S. Petrella et al. 2019. Overall structures of *Mycobacterium tuberculosis* DNA gyrase reveal the role of a *Corynebacteriales GyrB* specific insert in ATPase activity. *Structure* (2019).
[2] G. Evrard, F. Mareuil, F. Bontems, C. Sizun, and J. Perez. 2011. DADIMODO: a program for refining the structure of multidomain proteins and complexes against small-angle scattering data and NMR-derived restraints. *Journal of Applied Crystallography* 44 (2011), 1264–1271.
[3] F.-A. Fortin, F.-M. De Rainville, M.-A. Gardner, M. Parizeau, and C. Gagné. 2012. DEAP: Evolutionary Algorithms Made Easy. *J. Mach. Learn. Res.* 13, 1 (2012), 2171–2175.
[4] Konrad Hinsen. 2000. The molecular modeling toolkit: A new approach to molecular simulations. *Journal of Computational Chemistry* 21, 2 (2000), 79–85.
[5] F. Mareuil, C. Sizun, J. Perez, M. Schoenauer, JY. Lallemand, and F. Bontems. 2007. A simple genetic algorithm for the optimization of multidomain protein homology models driven by NMR residual dipolar coupling and small angle X-ray scattering data. *Eur Biophys J* 37(1) (2007), 95–104.
[6] O. Roudenko, A. Thureau, and J. Perez. 2018. DADIMODO: Refining Atomic Models of Multi-Domain Proteins against SAXS Data. https://dadimodo.synchrotron-soleil.fr. (2018). [Online; accessed March-2018].
[7] D. Svergun, C. Barberato, and M. H. J. Koch. 1995. *CRYSOL* – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *Journal of Applied Crystallography* 28, 6 (1995), 768–773.