Dadimodoweb user's guide

Roudenko Olga, Thureau Aurelien, and Perez Javier

Soleil Synchrotron
liste-exp-dadimodo@synchrotron-soleil.fr
https://dadimodo.synchrotron-soleil.fr

1 Introduction

Dadimodo [6] 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 structure of a multi-domain protein complex against SAXS experimental data, the degrees of freedom being the backbone dihedral angles of the protein flexible fragments. We use an "all-atoms" structure representation with energy control for every newly generated model so as to prevent steric clashes and converge to a physically feasible structure.

The present version takes its roots from [4,1]. The program described in [1] had configuration requirements that were hard to understand by external users. Hence, it was only used at Soleil for data analysis in collaborative studies. To overcome this drawback we designed an algorithm which automatically deduces, from user defined rigid parts (section 2.1), the transformations that are applicable to the structure at hand so as to allow search space exploration.

In addition, the introduction of an adaptive behavior of the rotation limit for backbone dihedral angles, inspired from state-of-the-art Evolutionary Algorithms, provides the updated Dadimodo with a more efficient search mechanism.

In terms of implementation, the calculation is sped up as a result of parallel execution on the cluster, enabled by the DEAP library [2].

Finally, the user can now choose between PepsiSAXS [3] and Crysol [7] for the evaluation of the molecule model against SAXS data.

Currently, Dadimodo is available to users via a web page hosted by Soleil Synchrotron. The user is thus provided with a distant access to the Soleil HPC resource where the computation is performed. The use of this service is open to all researchers from the corresponding scientific community whether their SAXS data were produced at Soleil synchrotron or at any other SAXS platform. For IT security reasons, every user is kindly asked to create an account before submitting his data.

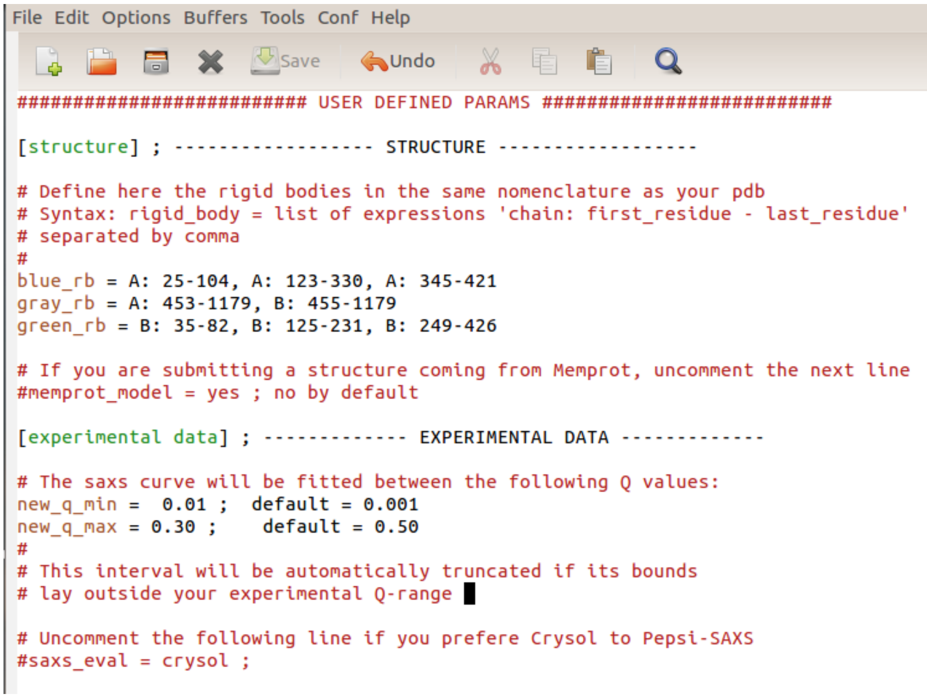
2 Input files

Three input files are required: the PDB file corresponding to a complete (all atoms) structure of the studied complex, the multi-column ascii file containing the experimental SAXS curve and configuration file where the rigid bodies are defined as well as Crysol, PepsiSAXS and some other parameters which are presented in more detail in section 2.1.

Let us take the structure of *Mycobacterium tuberculosis* DNA gyrase [5] as an example for the illustrative purposes. This complex is composed of two protein chains (A and B), 1179 residues in each.

2.1 Configuration file

The Dadimodoweb configuration file contains the following sections: `Structure, Experimental data, Pepsi-SAXS options, Crysol options` and `Optimization parameters`. The three last sections are optional since they do not necessarily need to vary from one case to another and the default parameter values will apply if the user does not have any particular preferences. Parameters related to the studied structure and experimental data are proper to the study at hand and are expected to be defined by the user.



```
########################## USER DEFINED PARAMS ##########################

[structure] ; ------------------ STRUCTURE ------------------

# Define here the rigid bodies in the same nomenclature as your pdb
# Syntax: rigid_body = list of expressions 'chain: first_residue - last_residue'
# separated by comma
#
blue_rb = A: 25-104, A: 123-330, A: 345-421
gray_rb = A: 453-1179, B: 455-1179
green_rb = B: 35-82, B: 125-231, B: 249-426

# If you are submitting a structure coming from Memprot, uncomment the next line
#memprot_model = yes ; no by default

[experimental data] ; ------------- EXPERIMENTAL DATA -------------

# The saxs curve will be fitted between the following Q values:
new_q_min =  0.01 ;  default = 0.001
new_q_max = 0.30 ;    default = 0.50
#
# This interval will be automatically truncated if its bounds
# lay outside your experimental Q-range

# Uncomment the following line if you prefere Crysol to Pepsi-SAXS
#saxs_eval = crysol ;
```

Fig. 1: Dadimodoweb config file

Rigid bodies definition The only mandatory part for the user to fill in the configuration file is the rigid bodies definition. Figure 2 illustrates the way Mtb gyrase structure will be split in rigid and flexible parts following the definition from the configuration file from figure 1. However, the default values for the experimental curve bounds are taken rather large and will not often apply well. The user would better take care of it by himself.

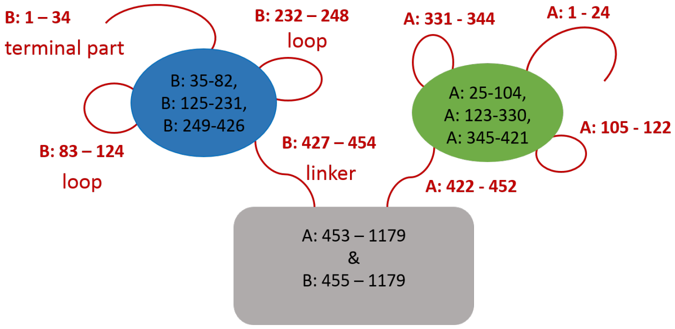


Fig. 2: Mtb gyrase scheme: rigid bodies and flexible parts. Numbers represent residues in peptide chains A and B

Crysol or PepsiSAXS Let us notice, that PepsiSAXS is a default SAXS simulator in Dadimodoweb. The user can uncomment the *saxs_eval* parameter line to switch to Crysol if needed. It is not necessary to remove PepsiSAXS parameters section when using Crysol or vice versa; useless parameters will be ignored. To have more information about the PepsiSAXS or Crysol parameters, we refer to the corresponding documentation.

Optimization

(ϕ, ψ) *rotation amplitude* The degrees of freedom of a 3D structure model adjustable from the SAXS data are all the dihedral angles (ϕ, ψ) of the residues composing the flexible parts of the structure (red cords in figure 2). There are typically several dozens such residues which gives an idea of the problem dimension.

We consider four types of flexible fragments: linkers, terminal parts, double linkers and loops. Except for the double linkers that happen quite rarely, all these types are present in our Mtb gyrase structure schema (Fig. 2).

Linkers 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 (Fig. 3 - left). When dealing with a double linker, it is physically impossible to modify one of its linkers without applying a correlated modification to the

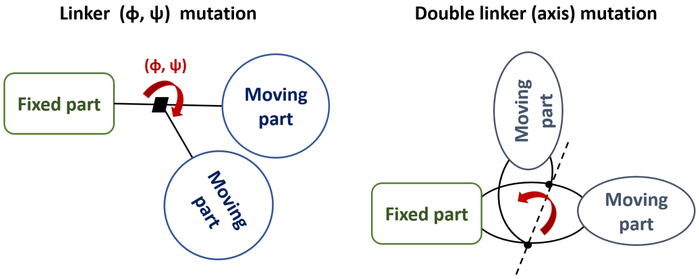


Fig. 3: *Dadimodo* search

other. It has been proposed in [1] to use a rotation around an axis passing by a randomly chosen C_α atom on each of its parts as illustrated in figure 3 - right. Let us notice that a loop can be seen as a double linker if we declare one residue in the middle of the loop as rigid body. The double linker rotation around axis can than be applied to loops.

The user parameter in the Optimization configuration section that governs the rotation amplitude is `mut_sigma`:

```
mut_sigma = 15
```

means that for generating a new structure, a random value of $(\Delta\phi, \Delta\psi)$ is picked following the normal distribution with $\sigma = 15°$. The user can decrease this value especially if the flexible parts are folded so as to clash easily except for small rotations.

Number of iterations Let us notice that the stopping criterion in Dadimodo is the maximum number of iterations which by default is fixed to 200. From our experience, it is amply sufficient. However, in some cases 200 iterations are too many and they pointlessly take hours of computation time. To adjust this value to the user's particular case, it may be useful to plot the evolution of the best fitness (fit quality) over iterations (see section 3). If after 100 iterations there is no significant improvement for any of the 5 runs, the total number of iterations may be reduced for the next submission.

To modify this value, the following line has to be added to the Optimization section:

```
max_nb_gens = 100
```

Syntax The user is not expected to write his configuration file from scratch. It is suggested to download the example available on the submission page and modify rigid bodies definition and adjust other parameters if needed. Some simple syntactic expectations are cited in the comment in the beginning of this file.

2.2 Molecule structure: PDB file

The residues in the input pdb file may be numbered separately in each peptide chain (like in our example Fig. 2) as well as transversely. The important thing is this numeration to correspond to the rigid bodies definition in the configuration file.

Add missing atoms/residues with Modeler

Ligands Ligands may only belong to a rigid part of the structure. Unlike peptides, which are identified via their number in a peptide chain, a ligand in rigid bodies definition is described via residue name and number.
.... example RB with a ligand ....

From Memprot output to Dadimodo input The PDB structure resulting from Memprot contains a particular membrane description (HETATM lines). To be correctly interpreted by Dadimodo dependencies (Crysol, PepsiSAXS and MMTK), before submitting to Dadimodoweb, this file has to be modified. The suitable modification can be done using the following command line:

```
sed -e '/HETATM/s/LYS....../LYS M9998/g'
    -e '/HETATM/s/LEU....../LEU M9999/g'
    memprot_output.pdb > dadimodo_input.pdb
```

It makes the heteroatoms LYS and LEU belong to the M chain and have residue numbers 9998 and 9999 respectively to make sure the chain name and residue numbers are uniques.
Then, Dadimodo needs to be told it is dealing with a structure coming from Memprot. To do so, the line

```
memprot_model = yes
```

in the Structure configuration section (Cf. Fig. 1) has to be uncommented. At last, the heteroatoms LYS and LEU take part of the rigid bodies definition under nicknames YYY and UUU respectively to prevent MMTK from interpreting them as standard peptides. Here is an example of definition of a rigid body containing the membrane part:

```
rb_with_membrane = A: 31-325, B: 31-325, C: 31-325, D: 31-325,
 F: 31-325,  YYY: 9998-9998, UUU: 9999-9999
```

2.3 Curve to fit: SAXS experimental data

Three column file is expected where the columns are q values, corresponding intensities and sigma values. A header made of comment lines starting with # is accepted.

3 Dadimodoweb output

Once a computation is completed, a zip file is sent to the user by email. This archive is decompressed to `dadimodo_results` folder, which contains three subfolders: `input_files, log` and `calculation_results`.

Since Dadimodo relies on a stochastic optimization process (Evolutionary Algorithm), for each submission several runs are performed, each resulting in a different solution. Currently, five runs per submission are reasonably affordable for the Soleil cluster. In principle, all runs are expected to converge to similar structures. If it is clearly not the case, and especially if the SAXS fitting quality of the resulting structure vary significantly from one run to another, we would suggest to let us know via `liste-exp-dadimodo@synchrotron-soleil.fr`.

Five resulting PDB files and corresponding simulated SAXS curves (`.fit` files) can be found in `calculation_results` folder. When Crysol is used for structures evaluation, five Crysol log files are also kept along.

In the `log` folder five ”run logs” are made available for troubleshooting purposes, they are indeed not very readable to the user. In this folder the user can also find, for each of five runs, an evolution log which is a multi-column file, the columns corresponding to the evolution of the best, average and worst fitness values over the iterations. These columns (especially the best fitness) can be plotted in particular to see if the significant improvement systematically stops long before the end of the iterations. In this case, the number of iterations could reasonably be reduced for the next submission, as explained in section 2.1.

References

1. Evrard, G., Mareuil, F., Bontems, F., Sizun, C., Perez, J.: 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**, 1264–1271 (2011)
2. Fortin, F.A., De Rainville, F.M., Gardner, M.A.G., Parizeau, M., Gagné, C.: Deap: Evolutionary algorithms made easy. J. Mach. Learn. Res. **13**(1), 2171–2175 (2012)
3. Grudinin, S., Garkavenko, M., Kazennov, A.: Pepsi-SAXS : an adaptive method for rapid and accurate computation of small-angle X-ray scattering profiles. Acta Crystallographica Section D: Biological Crystallography **D73**, 449 – 464 (2017), https://hal.inria.fr/hal-01516719
4. Mareuil, F., Sizun, C., Perez, J., Schoenauer, M., Lallemand, J., Bontems, F.: 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)**, 95–104 (2007)
5. Petrella, S., et al: Overall structures of *Mycobacterium tuberculosis* DNA gyrase reveal the role of a *Corynebacteriales GyrB* specific insert in ATPase activity. Structure (2019)
6. Rudenko, O., Thureau, A., Perez, J.: Evolutionary refinement of the 3d structure of multi-domain protein complexes from small angle x-ray scattering data. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion. pp. 401–402. Association for Computing Machinery, New York, NY, USA (2019), https://doi.org/10.1145/3319619.3322002

7. Svergun, D., Barberato, C., Koch, M.H.J.: *CRYSOL* – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. Journal of Applied Crystallography **28**(6), 768–773 (1995)