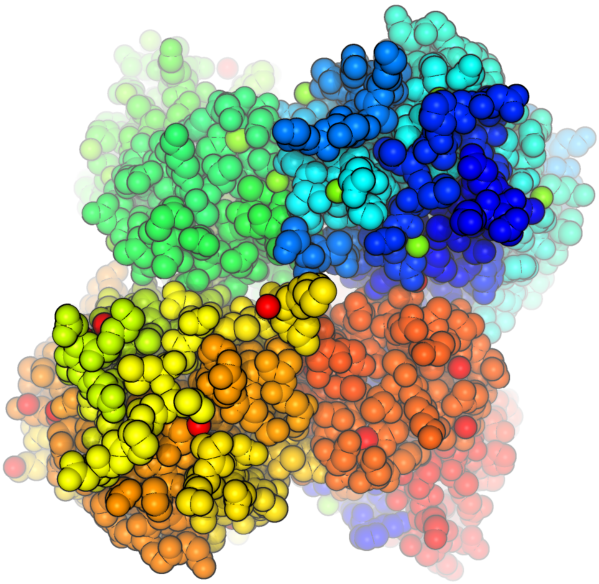
Geometrically Stable Substructures (GeoStaS)

ver. 1.1

Manual & tutorial

Julia Romanowska, Krzysztof Nowiński, Joanna Trylska



June 9, 2012

Chapter 1

Introduction

This chapter describes the basic information on where the program is available, what is needed to be able to use it and how to run the program.

1.1 How to obtain the program

The GeoStaS program is a freely available open source software under GNU General Public License[1], which means there is no charge for using it and the underlying Java code can be accessed and freely modified, provided that the information of the source and authors is passed along.

From the web page bionano.icm.edu.pl/Software/GeoStaS or bitbucket.org/jrom/geostas one can download the following packages:

- tar ball with Java classes source code;
- HTML version of the documentation of the classes (*JavaDoc*);
- compiled program, with all the libraries, for any operating system;
- README and LICENSE texts;
- exemplary trajectory (useful with the tutorial at the end of the manual!);
- this manual in PDF.

1.2 Requirements

1.2.1 Software

Java Runtime Environment (version 1.6 or higher) must be installed in the operating system, for the program to run correctly. The program has been tested on Kubuntu 10.04, Kubuntu 11.04, Fedora 14, Fedora 15, Windows 7, Windows Vista, Windows XP and MacOS. The following Java environments were tested:

- OpenJDK Runtime Environment (IcedTea6 1.9.8) (6b20-1.9.8-0ubuntu1 10.04.1), (IcedTea6 1.10.02) (6b22-1.10.02-0ubuntu1 11.04.1), (IcedTea6 1.11pre) (6b23 pre11-0ubuntu1.11.10.2), (IcedTea6 1.10.3) (fedora-59.1.10.3.fc15-x86_64), and (IcedTea6 1.10.6) (fedora-63.1.10.6.fc15-x86_64) on Linux;
- Java(TM) SE Runtime Environment build 1.6.0_26-b03 and build 1.6.0_20-b02 on Windows;

[1] http://www.gnu.org/licenses/licenses.html#GPL

- Java(TM) SE Runtime Environment build 1.6.0_29-b11-402-10M3527 on MacOS.

1.2.2 Hardware

Depending on the size of the trajectories you want to analyze, the required amount of RAM may vary. The larger, the better. ;-)

1.3 Running the program

1.3.1 On Linux:

Open your favorite terminal and type:

```
java -jar program_name
```

When loading large trajectories, the `-Xmx` option is quite useful, which sets the maximum RAM volume that Java machine will use. For example, if you want to load a 2 GB trajectory, you could allow Java to use 4 GB of RAM with the command:

```
java -Xmx4g -jar program_name
```

This option is useful especially if you notice a program stopping with the output saying "Java out of heap space". More information about running Java on Linux can be found on the Sun web pages[2].

1.3.2 On Windows:

If you have a JRE installed properly, you should be able to run the program by double-clicking on the icon.

If you wish to adjust the parameters for java virtual machine, open the console by clicking on Start → "Run program", and typing "cmd" in the newly appeared window. Then, type a command (like in the Linux environment).

1.3.3 On Mac:

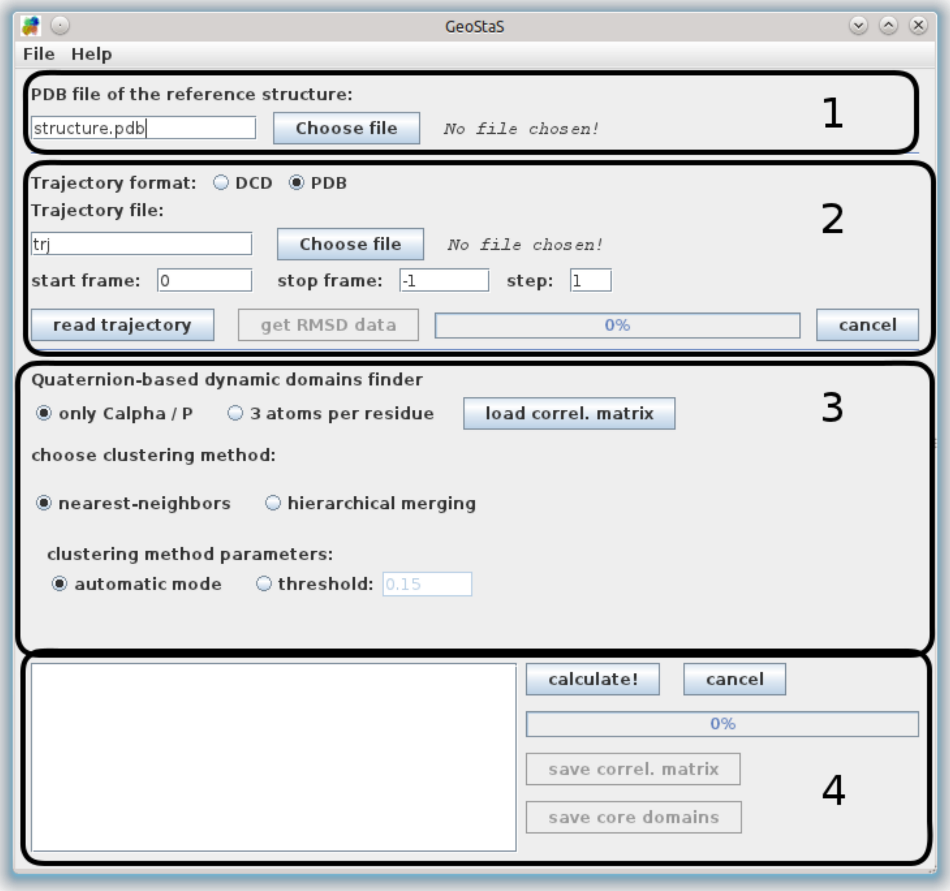
If you have a JRE installed properly, you should be able to run the program by double-clicking on the icon.

[2] http://download.oracle.com/javase/1.4.2/docs/tooldocs/linux/java.html

Chapter 2

Program description

2.1 Overview of the GUI



PDB file of the reference structure: structure.pdb

No file chosen!

Trajectory format: ○ DCD ● PDB

Trajectory file: trj

No file chosen!

start frame: 0 stop frame: -1 step: 1

0%

Quaternion-based dynamic domains finder

● only Calpha / P ○ 3 atoms per residue

choose clustering method:

● nearest-neighbors ○ hierarchical merging

clustering method parameters:

● automatic mode ○ threshold: 0.15

0%

Figure 2.1: The overview of the application window right after the start.

The following sections describe the details of graphical interface. The modules are marked in Fugure 2.1 with numbers:

1. reference structure loader;

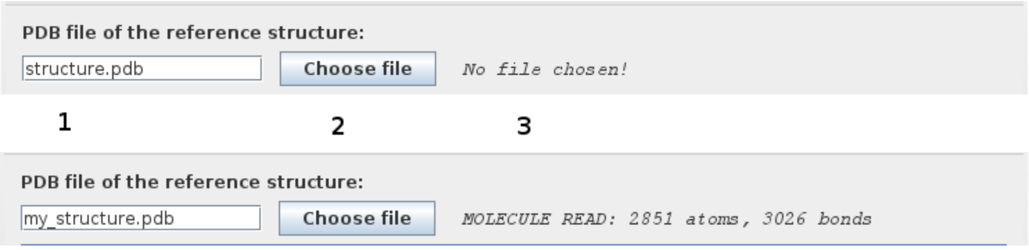
2. conformation set loader;

3. parameters for the module for finding dynamic domains;

4. controls and output for the module for finding dynamic domains.

2.2 Loading the structures

Prior to any calculations, the representative or starting structure of the molecule has to be loaded into the application.



PDB file of the reference structure: structure.pdb

No file chosen!

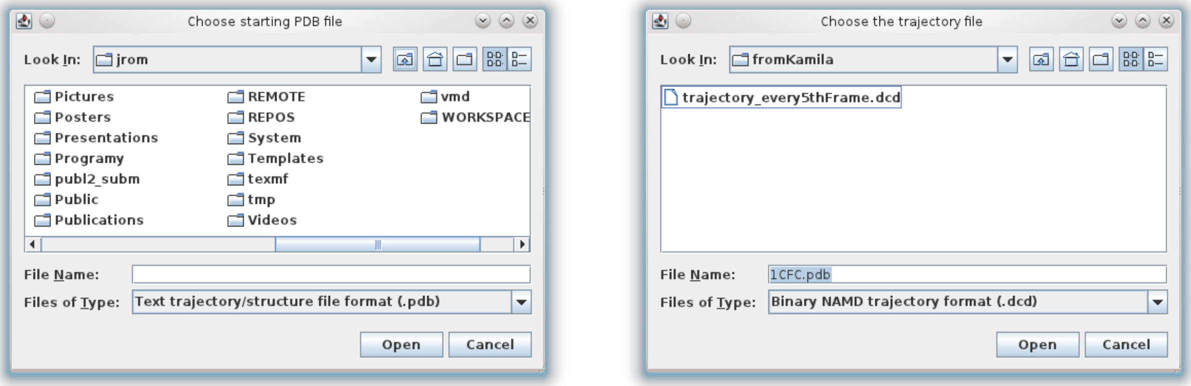
PDB file of the reference structure: my_structure.pdb

MOLECULE READ: 2851 atoms, 3026 bonds

Figure 2.2: Section of GUI before *(top)* and after *(bottom)* a successful load of the molecule.

The top part of GUI includes the text field where the name of the PDB file is displayed (**1**), the button *Choose file* (**2**) and the information label (reading *No file chosen* in the beginning) (**3**), as presented in Fugure 2.2, top. When the button is pressed, a file chooser window appears (Fugure 2.3a). After choosing the correct file, the application loads the molecule into the memory and displays information on the size of the molecule (Fugure 2.2, bottom).

NOTE:
⋆ only files with the specific enxtension are visible in the window!
⋆ if your PDB file contains water and/or ions, these will not be used in the following calculations, but the trajectory file should still match the whole PDB structure
⋆ check carefully whether the loaded molecule has the correct number of atoms, i.e., whether the displayed information matches the number of atoms in the PDB file — non-standard names of atoms (like in some topologies generated by Amber, e.g., the name LO13 describing the oxygen atom) may not be recognized by the application; in this case, you have to manually change the names of these atoms in the starting PDB file



Choose starting PDB file

Look In: jrom

File Name:

Files of Type: Text trajectory/structure file format (.pdb)

(a) PDB file chooser

Choose the trajectory file

Look In: fromKamila

File Name: 1CFC.pdb

Files of Type: Binary NAMD trajectory format (.dcd)

(b) DCD file chooser

Figure 2.3: The windows for choosing a PDB or DCD format file — for the ease of use only the files with extensions .pdb or .dcd are shown by the file chooser.

2.3 Loading trajectories

Next, you should provide the trajectory file (or simply a file containing different conformations of the loaded molecule) — the extensions that are currently recognized are .dcd and .pdb. The section of GUI responsible for loading the trajectory is presented in Fugure 2.4, top. This module contains the following elements (numbering as in the figure):

1 format chooser — depending on the choice of file format, the subsequent file chooser presents only files with specific extension;

2 a set of components similar to the structure loader: the text field showing the name of the chosen file, the *Choose file* button, which triggers opening of the file chooser (Fugure 2.3), and the label displaying information on the trajectory;

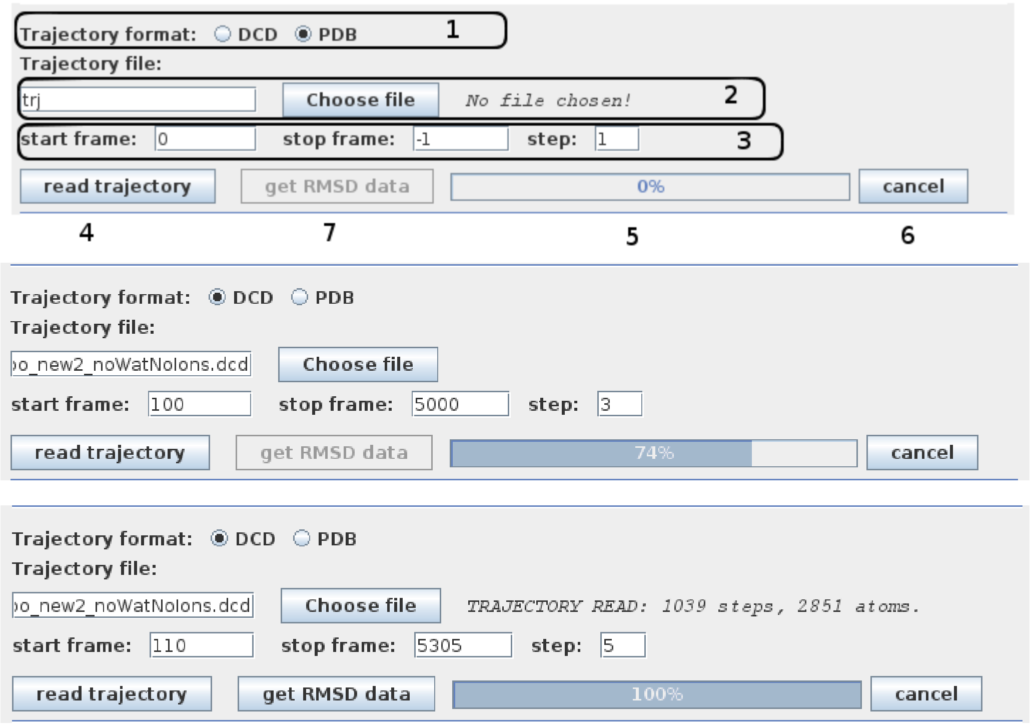
3 controls for setting the first and the last frame that should be read, and the interval; if the last frame is set to -1, the whole trajectory will be read;

4 *read trajectory* button — only after clicking this button, the file will be read and the trajectory will be loaded into memory;

5 the progress bar, showing how much of the trajectory has been already loaded (Fugure 2.4, middle);

6 throughout the loading process the *cancel* button can be pressed to stop the action;

7 *get RMSD data* button — after successful reading of the trajectory (Fugure 2.4, bottom) you can write the basic root mean square distance values for the whole molecule, per each frame, to a specified file on the disk.



Trajectory format: ○ DCD ● PDB
Trajectory file: trj
No file chosen!
start frame: 0 stop frame: -1 step: 1
0%

Trajectory format: ● DCD ○ PDB
Trajectory file: ›o_new2_noWatNoIons.dcd
start frame: 100 stop frame: 5000 step: 3
74%

Trajectory format: ● DCD ○ PDB
Trajectory file: ›o_new2_noWatNoIons.dcd
TRAJECTORY READ: 1039 steps, 2851 atoms.
start frame: 110 stop frame: 5305 step: 5
100%

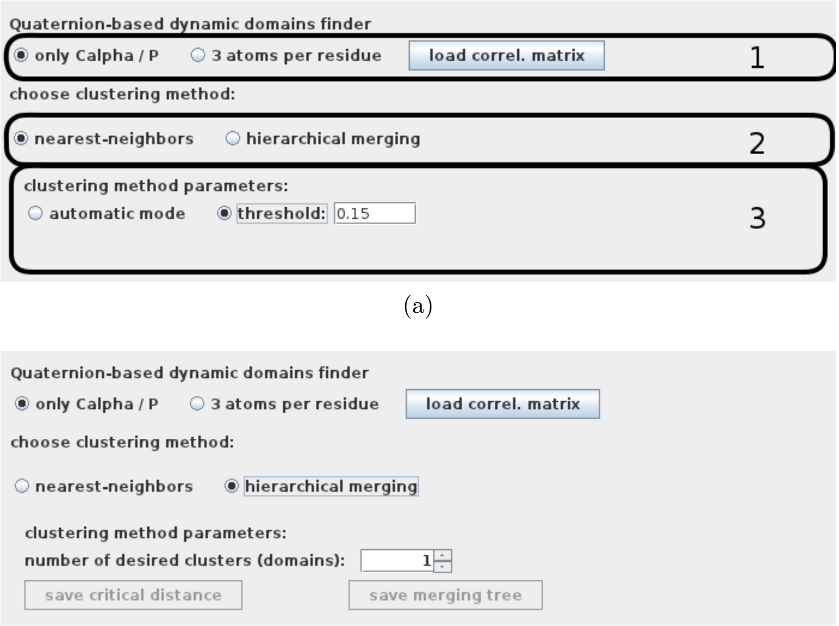
Figure 2.4: Section of the GUI before *(top)*, during *(middle)* and after *(bottom)* successful reading of the trajectory file.

NOTE:

- ⋆ the default values of the first frame, the last frame, and the interval are: 0, −1, and 1, respectively, which means “read from beginning till the end, every frame”;
- ⋆ if you fill the *start*, *stop* or *step* fields with the values that are larger than the number of available frames, or any combination of non-consistent values, you will be notified about it and asked to correct the input;
- ⋆ it is recommended to use binary dynamics format (.dcd) for larger files, however, any file format would take longer time to load above certain file size (ca. 300 MB).

2.4 Dynamic Domains

This procedure concentrates on individual atoms instead of entire conformations of a molecule. It extracts the movements of each atom and performs a geometric comparison of the traces of movements in a pairwise manner. The comparison is done with the use of quaterion representation of rotations and a search for the best superposition of the two traces. The calculations are implemented based on theory presented in Kneller, G. R., & Calligari, P. *Acta crystallographica. Section D*, **2006**, 62, 302-11. This results in construction of a matrix of similarity coefficients (called atomic movements similarity matrix, AMSM) of size N×N, where N is the number of atoms taken into consideration. These similarity coefficients are then clustered and the clusters are translated into atom groups. Currently, there are two clustering algorithms to choose, as described below. The detailed description of the method is given in [Romanowska, J., Nowiński, S. K., and Trylska, J., *J Chem Theory Comput*, submitted].


Quaternion-based dynamic domains finder

● only Calpha / P ○ 3 atoms per residue

choose clustering method:

● nearest-neighbors ○ hierarchical merging

clustering method parameters:

○ automatic mode ● threshold: 0.15

(a)

Quaternion-based dynamic domains finder

● only Calpha / P ○ 3 atoms per residue

choose clustering method:

○ nearest-neighbors ● hierarchical merging

clustering method parameters:

number of desired clusters (domains): 1

(b)

Figure 2.5: The views of the module for finding dynamic domains. The two algorithms for clustering of AMSM have different parameter sets: *(a)* common nearest-neighbors algorithm, *(b)* hierarchical merging algorithm.

First, the desired level of details of calculations has to be chosen (area **1** in Fugure 2.5a): whether to take into consideration only one atom per residue (i.e., the “only CA/P” mode, Cα or P atoms) or three atoms per residue (i.e., Cα-C-N in case of proteins, and P-C3′-C4′ in case of nucleic acids).

Next, the user can choose between two algorithms for clustering the AMSM (area **2** in Fugure 2.5a): *(i)* common nearest-neighbor graph algorithm (NN); and *(ii)* hierarchical merging of AMSM columns

(HM). More complete tests were performed with the use of the NN algorithm, and this clustering enables easy implementation of the automated mode. However, HM gives insight into the process of cluster merging, thus providing a means for the user to control the level of details that the division into dynamic domains gives.

2.4.1 Common Nearest-Neighbors clustering of AMSM

This clustering method includes two modes: automatic and manual. The automatic mode is the default one and it involves calculating the division into the domains for a set of different thresholds, and choosing the optimal division. The threshold set includes the following values: 0.3, 0.25, 0.2, 0.15, 0.1. For each threshold, the trajectory frames are superimposed with respect to each of the identified domains, and the overall RMSD spread is calculated. The division that yields the lowest RMSD spread is then taken as the optimal one.

The results in some cases depend on the “coarseness” of the representation, i.e., sometimes taking only Cα or P atoms is not enough to notice the subtle correlations, and only taking three atoms per residue gives the best division. Therefore, the user can choose between these two modes, independently of choosing between the automatic and manual modes.

Sometimes the automatic mode gives us too rough division into the domains, and one would like to investigate the differences in the movements in more detail, thus in the manual mode (Figure 2.5a) the user can simply set the desired threshold and observe how it influences the specific result.

2.4.2 Hierarchical Merging clustering of AMSM

In this method, the user has to explicitly set the desired number of clusters, i.e., into how many domains the molecule should be divided (Figure 2.5b). At first, it may seem like the worst type of parameter that could appear in this program, since it is also the very outcome of this analysis. However, hierarchical clustering gives us an easy approach to guess this number. According to [Shao J, Tanner SW, Thompson N, Cheatham TE, IIIrd. Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J Chem Theory Comput.* **2007**; 3(6):2312-2334], when the value of the distance of the currently merged clusters differs a lot from the previous one, these clusters should probably not be merged. The distance is called critical distance (CD). Therefore, when plotting CD vs. clustering step, the number of clusters that gives a reasonable division into dynamic domains, can be found. An example is given below.

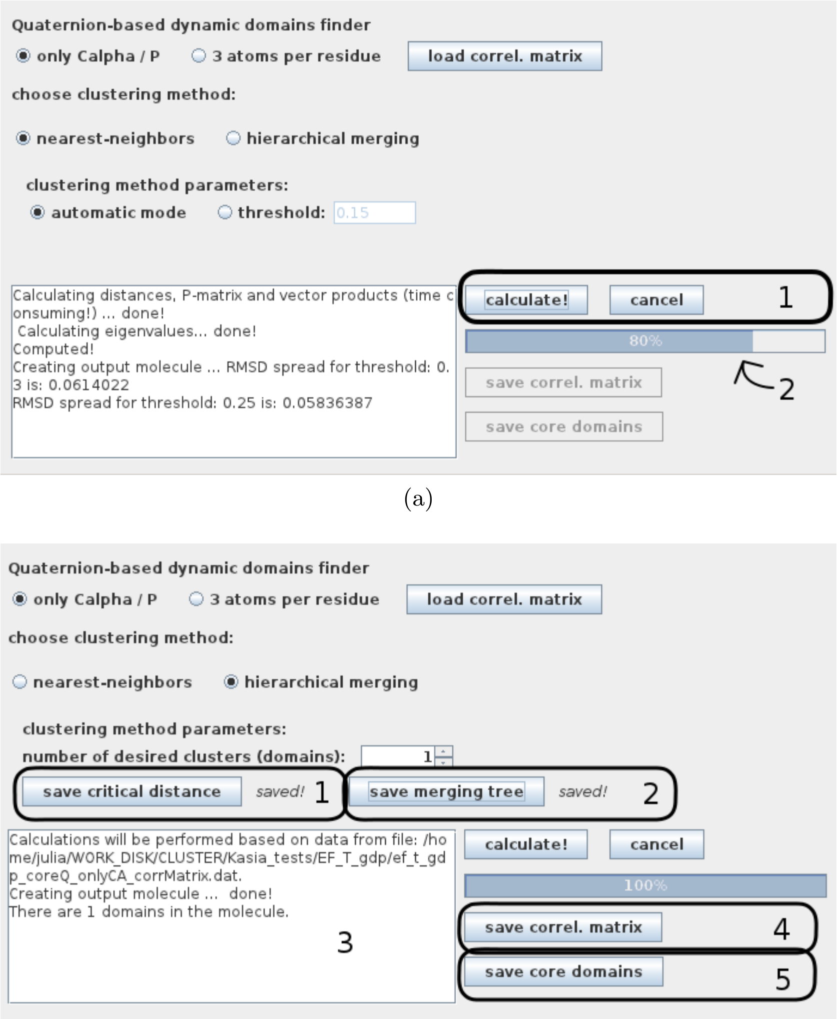
2.4.3 Starting the calculations and collecting the output

After choosing the clustering method and setting relevant parameters, the calculations are started upon pressing the *calculate* button and aborted upon pressing *cancel* button (**1** in Figure 2.6a). The progress bar updates accordingly (**2** in Figure 2.6a).

When choosing the HM algorithm, we recommend to first run the clustering till the complete merge (i.e., setting the *number of desired clusters* to “1”). Next, save the CD vs. clustering step (by clicking on the *save critical distance* button; **1** in Figure 2.6b), and visualize it in your favorite plotting software. An exemplary graph is presented in Fugure 2.7. The visual inspection helps to determine the number of clusters (and so the number of domains) to choose for the next run of the calculations.

If the user is interested also in the process of cluster merging, the clustering tree can be analyzed for this purpose. The tree is saved by clicking on the *save merging tree* button (**2** in Figure 2.6b). Here, a Newick tree format is used to describe the tree. Wikipedia provides a clear description of this format (en.wikipedia.org/wiki/Newick_format). The text output can be visualized e.g., in Drawgram[1],

[1] www.phylogeny.fr/version2_cgi/one_task.cgi?task_type=drawgram



Quaternion-based dynamic domains finder

● only Calpha / P ○ 3 atoms per residue

choose clustering method:

● nearest-neighbors ○ hierarchical merging

clustering method parameters:

● automatic mode ○ threshold: 0.15

```
Calculating distances, P-matrix and vector products (time c
onsuming!) ... done!
 Calculating eigenvalues... done!
Computed!
Creating output molecule ... RMSD spread for threshold: 0.
3 is: 0.0614022
RMSD spread for threshold: 0.25 is: 0.05836387
```

(a)

Quaternion-based dynamic domains finder

● only Calpha / P ○ 3 atoms per residue

choose clustering method:

○ nearest-neighbors ● hierarchical merging

clustering method parameters:

number of desired clusters (domains): 1

```
Calculations will be performed based on data from file: /ho
me/julia/WORK_DISK/CLUSTER/Kasia_tests/EF_T_gdp/ef_t_gd
p_coreQ_onlyCA_corrMatrix.dat.
Creating output molecule ...  done!
There are 1 domains in the molecule.
```

(b)

Figure 2.6: Module for finding dynamic domains: *(a)* the view during the calculations with the NN algorithm and automatic mode set; and *(b)* after the calculations (HM clustering algorithm) and with the correlation matrix loaded beforehand.

which is part of a simple but powerful server described in [Dereeper A, Guignon V, Blanc G, *et al.* Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic acids res.* **2008**; 36:W465-9].

When the calculations are successful, the output information is presented in the text area (**3** in Figure 2.6b) and the user has two types of output to save: the correlation matrix and the molecule divided into the dynamic domains (**4** and **5** in Figure 2.6b). The matrix can be written to a text file, and used later as an additional input. The molecule can be saved in the PDB format, with different domains described by different chain names. This file can be later visualized in an external program.

2.4.4 Providing additional input

The calculation of the correlation matrix can be time consuming, therefore the user can decrease the time of the run by loading the file with the correlation matrix prior to calculations of the domains. This can be done by pressing the *load correl. matrix* button (**1** in Figure 2.5a) and pointing to the specific file. The user is notified if the reading of the information in the file was successful (Figure 2.6b).

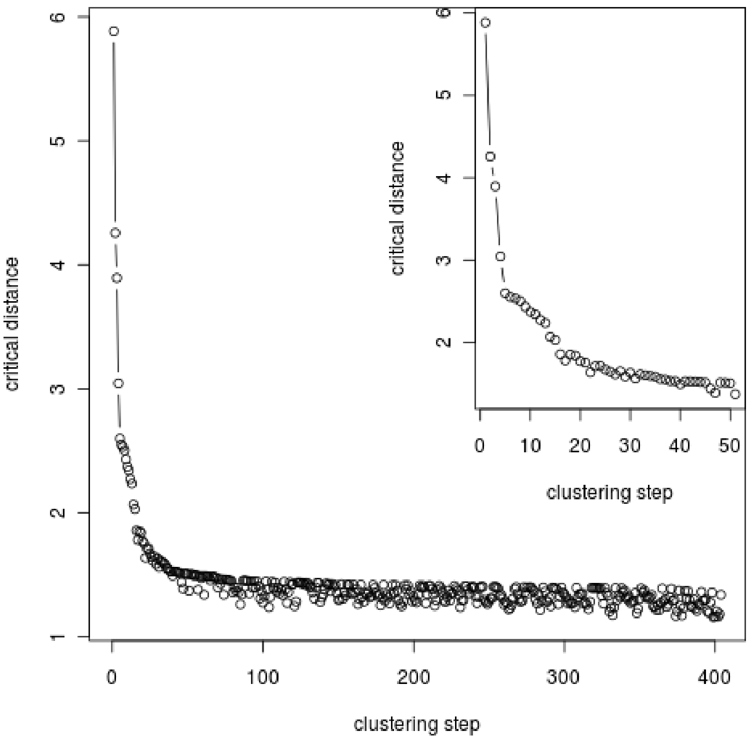


Figure 2.7: Critical distance (CD) versus merging step — the large jumps in CD indicate merging of two clusters that are unrelated and most probably should stay separated. The inset presents a zoomed region where the largest changes are present.

NOTE:

⋆ after successful loading of correlation matrix, subsequent calculations of dynamic domains will be performed based on the values from the file; if one wants to reset this, a new trajectory has to be loaded

⋆ when setting the threshold manually, we recommend to use relatively small numbers (i.e., between 0.3 to 0.05), however the calculations will be performed for any threshold

⋆ when saving the output, the user can provide only the beginning of the file name — the extension is added automatically, based on the type of output

Chapter 3

Usage cases / tutorials

3.1 CASE 1 – NMR ensemble; automatic mode

For this tutorial you will need only one PDB file, with an ensemble of conformations from NMR, downloadable from the PDB website (www.pdb.org), named 1CFC. This file contains 25 conformations, so it is very light and GeoStaS should complete the job fast.

1. Load the 1CFC.pdb file both, as a starting structure and as a trajectory. Leave the default “start”, “stop” and “step” values.
2. Leave the default settings for dynamic domain finder: “nearest-neighbors” clustering algorithm, “automatic mode” and “onlyCA” should be marked. Press “calculate” button.
3. ... and *voilá*! Save the output: similarity matrix and the molecule.

Figure 3.1 shows the visualized output. The different dynamic domains are marked with different chain names in the output PDB, therefore you can visualize the division of the molecule by uploading the PDB file to your favorite program and making it color the structure by chain names.

You can repeat the calculations for the three-atom-per-residue representation in the automatic mode and compare the minimal RMSD spreads from both representations. You will see that the Cα mode gives smaller RMSD spread, therefore it should better define the dynamic domains.

3.2 CASE 2 – NMR ensemble; manual mode

In some cases you will need to manually set the threshold, e.g., when you would like to focus on some detailed division into dynamic domains or when the automatic mode provides too crude or maybe too detailed division. The only difference between the manual and automatic modes is in explicitly providing the threshold, i.e., the value determining how compact the dynamic domains are. If the value is set higher than 0.2, a warning message will appear, but you may continue to perform calculations.

3.3 CASE 3 – MD simulation

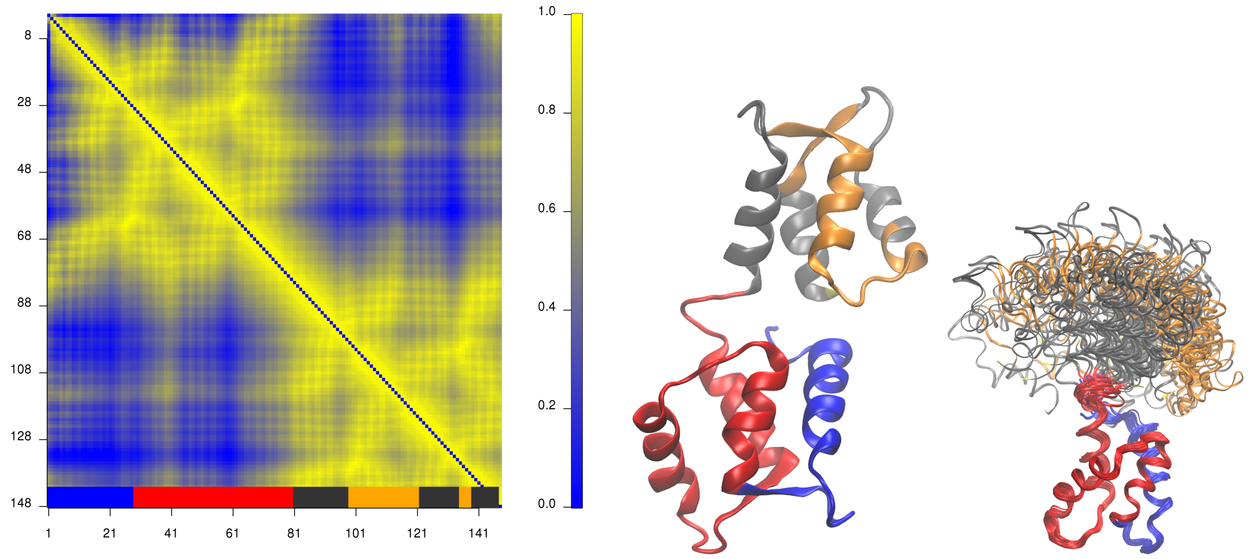
Currently, GeoStaS accepts only DCD files as binary trajectories, therefore if your trajectory is in some other format, please convert it beforehand. Also, if your initial structure is in other format than PDB, the starting PDB file should be prepared. One robust tool that enables conversion between many text and binary formats is VMD.

When analyzing the dynamics for the first time, you are advised to use the “automatic mode” with the NN algorithm or clustering till complete merge with the HM algorithm. In the latter case, follow a simple rule, described in the previous chapter, section 2.4.3 *Starting the calculations and collecting the output*. When choosing the NN algorithm, check the divisions found with the automatic mode for one- and three-atoms-per-residue representations. If one of them has significantly lower RMSD spread than the other, you should probably choose it for subsequent studies. Sometimes, when the domains found with the automatic mode are too coarse, you should try manually setting a different threshold. Since the calculation of the similarity matrix takes a long time, be sure to save it to a file when calculating for the first time.

You can download exemplary trajectory from the GeoStaS webpage. The files include:

- the starting structure and trajectory of elongation factor EF-Tu:GDP complex [Kulczycka, K., Długosz, M., & Trylska, J. *European biophysics journal*, **2011**, 40(3), 289-303]:
  `ef_t_gdp_1stFrame_noGDP.pdb`, `trajectory_25ns_10thFrame_noGDP.dcd`,
  zipped to `ef_t_gdp_every10thFrame_KKulczycka.zip` (90 MB);
- the starting structure and trajectory of elongation factor EF-Tu:tRNA complex [Kulczycka, K., Długosz, M., & Trylska, J. *European biophysics journal*, **2011**, 40(3), 289-303]:
  `ef_t_trna_1stFrame.pdb`, `trajectory_25ns_10thFrame.dcd`,
  zipped to `ef_t_trna_every10thFrame_KKulczycka.zip` (126 MB);

When analyzing these trajectories, you will notice that the lower RMSD spread correlates with visually more “reasonable” division — Figure 3.2 compares the results for the EF-Tu:GDP complex.

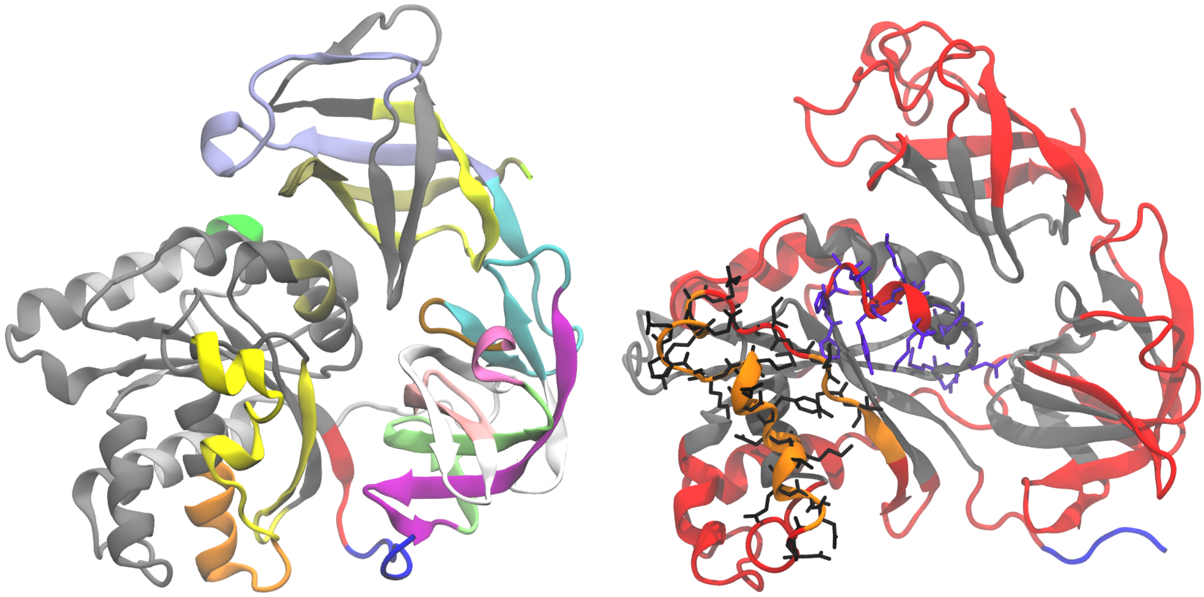


(a) AMSM

(b)

(c)

Figure 3.1: Exemplary visualization of the GeoStaS output for the 1CFC NMR ensemble. *(a)* The atomic movements similarity matrix (AMSM) showing pairwise correlations between atomic trajectories; the color bars on the x axis show different domains; visualized in R, with the use of custom script. *(b–c)* The structure colored by domains and identified by GeoStaS; visualized in VMD; *(b)* one conformation, cartoon representation; *(c)* whole ensemble superimposed with respect to the red-colored chain.



(a) "3 atoms" representation

(b) "Cα" representation

Figure 3.2: Comparison of the assignement into dynamic domains done by GeoStaS for the EF-Tu:GDP complex, for the *(a)* three- and *(b)* one-atom-per-residue representations. Only the initial structures are presented, colored by the dynamic domains. The sticks in *(b)* mark the so-called "switches" — regions that were identified as involved in biologically important mechanisms and were shown to be flexible [Kulczycka, K., Długosz, M., & Trylska, J. *European Biophysics Journal*, **2011**, 40(3), 289-303]. Minimal RMSD spreads are: *(a)* 0.0208 and *(b)* 0.0192.

Chapter 4

Contact information

If you have any questions and/or suggestions, write to me! I should respond...

Julia Romanowska
web page: http://bionano.icm.edu.pl/People/JuliaRomanowska
e-mail: `jrom@icm.edu.pl`