A Flexible Protein Folding Simulator for Cluster Environment

Malgorzata Tomanek, Tomasz Szepieniec,
Irena Roterman-Konieczna

ACK Cyfronet AGH - ZBiT CM UJ

Cracow Grid Workshop, Kraków, 27-29.10.2014

INNOVATIVE ECONOMY
NATIONAL COHESION STRATEGY

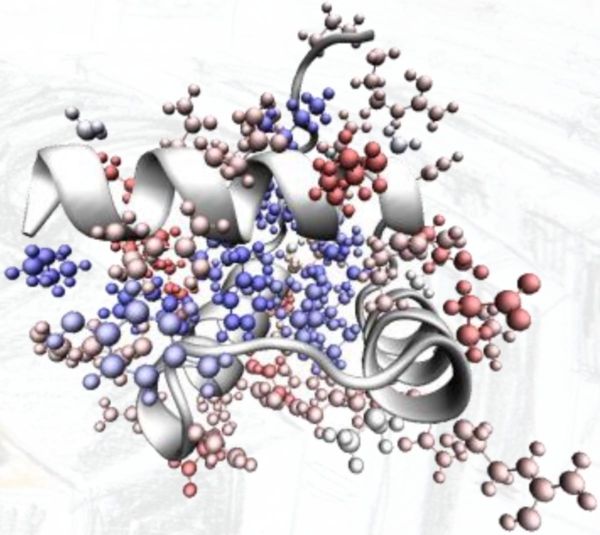
What is a protein folding?

- A biological process of proteins creation using information obtained from DNA

Amino acids sequence:

```
RPRTAFSSEQLAR
LKREFNENRYLTE
RRRQQLSSELGLN
EAQIKIWFQNKRAKI
```

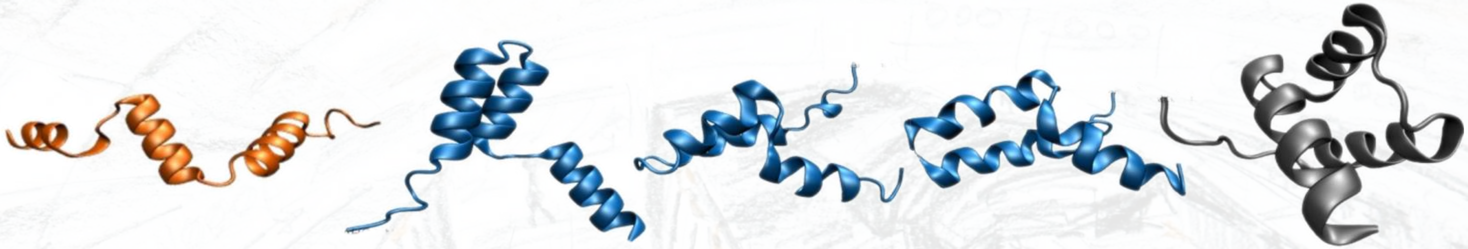
Protein folding



- Being able to predict result of this process using in silico methods is very crucial for drug design, especially in an aspect of personal therapy

PL GRID NG CYFRONET INNOVATIVE ECONOMY NATIONAL COHESION STRATEGY

DRIPPY ATTACK – THE PROTEIN FOLDING SIMULATOR

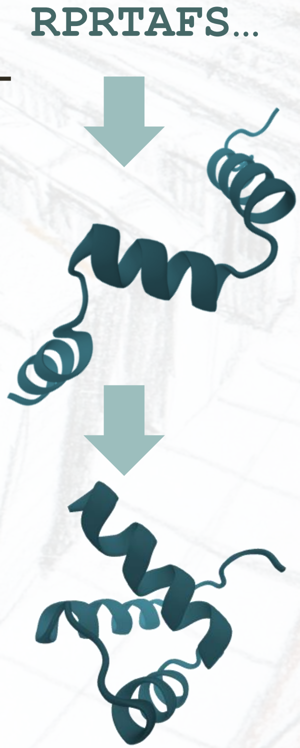


- simulates a protein folding process using a two-step model
- is designed as a workflow built from flexible modules. Each module represents some step of the folding process
- makes it possible to test a large amount of process parameters, collect results of the simulation and compare them
- is created for efficient usage with local queue system in PL-Grid infrastructure.

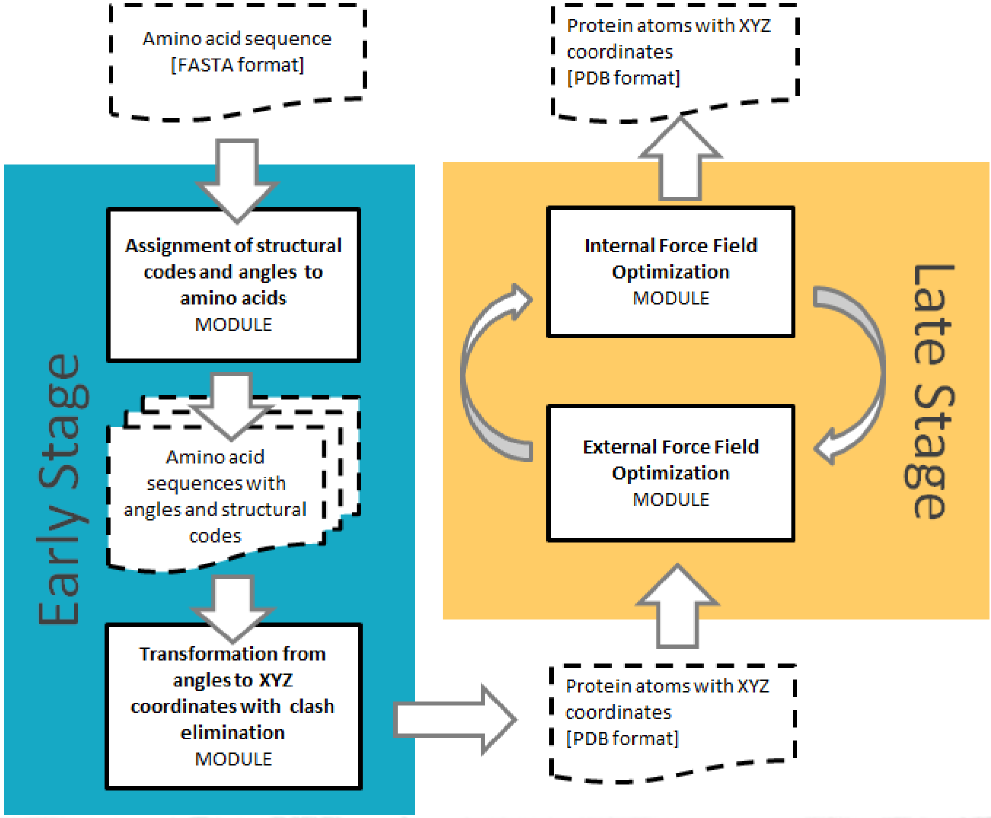
PL GRID NG
CYFRONET
INNOVATIVE ECONOMY
NATIONAL COHESION STRATEGY

Two-step model of the protein folding

- In this model the protein structure is generated during two main stages: the Early and the Late
- The Early Stage: an amino acid sequence in form of an one-letter code sequence is enriched with an information about dihedral angles and a specific structural codes. In the next step a 3D form of the protein is generated and clashes between atoms are eliminated
- The Late Stage: pre-folded protein is optimized in terms of hydrophobicity space distribution (external force field influence) and inner energy of the compound.



A WORKFLOW OF THE SIMULATOR


Amino acid sequence
[FASTA format]
Early Stage
Assignment of structural codes and angles to amino acids
MODULE
Amino acid sequences with angles and structural codes
Transformation from angles to XYZ coordinates with clash elimination
MODULE
Protein atoms with XYZ coordinates
[PDB format]
Late Stage
Internal Force Field Optimization
MODULE
External Force Field Optimization
MODULE
Protein atoms with XYZ coordinates
[PDB format]

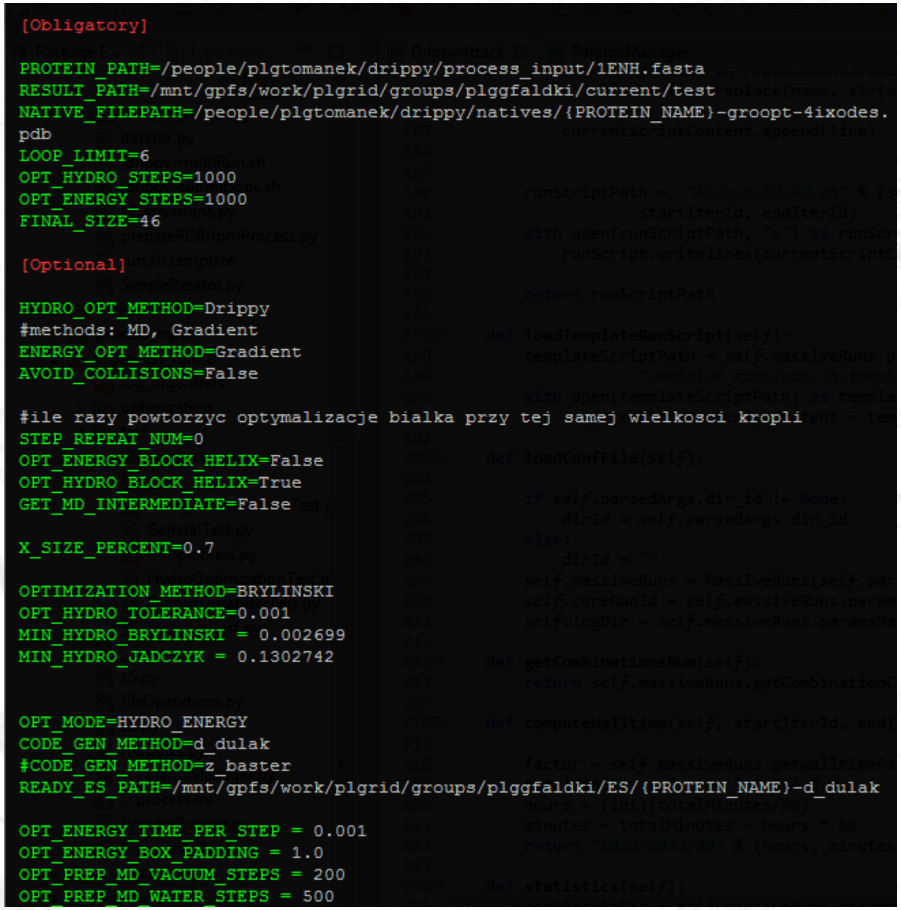
- Each step of the process is designed as a module in a workflow
- The Early and the Late Stage can be run separately
- Each module can be replaced or enhanced by the user provided that the module interface is kept

External software used in the simulator

- The Early Stage:
  - Script written by Barbara Kalinowska
  - Programs written by Dawid Dulak and Zbigniew Baster
- The Late Stage:
  - Gromacs (www.gromacs.org)
- Assesment of the structures:
  - Maxcluster (www.sbg.bio.ic.ac.uk/~maxcluster)
- Visualizations:
  - VMD (www.ks.uiuc.edu/Research/vmd)
  - gnuplot (www.gnuplot.info)

PL GRID NG CYFRONET INNOVATIVE ECONOMY NATIONAL COHESION STRATEGY

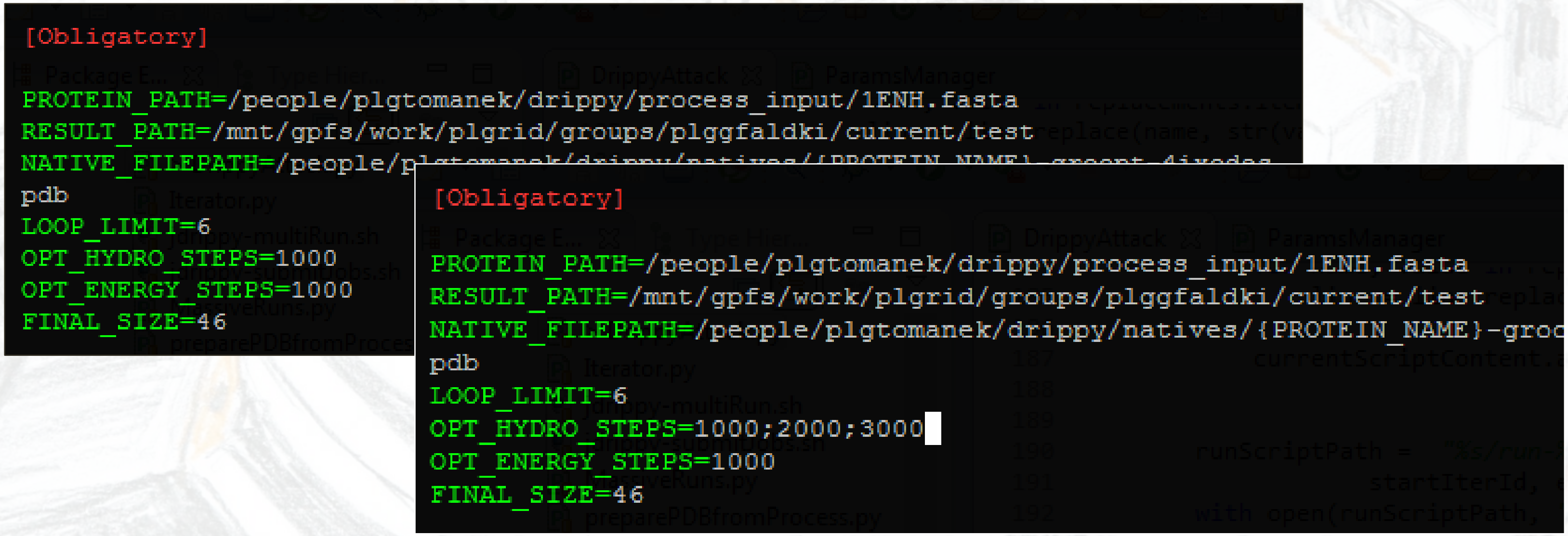
Simulation process can be parametrized in details



```
[Obligatory]

PROTEIN_PATH=/people/plgtomanek/drippy/process_input/1ENH.fasta
RESULT_PATH=/mnt/gpfs/work/plgrid/groups/plggfaldki/current/test
NATIVE_FILEPATH=/people/plgtomanek/drippy/natives/{PROTEIN_NAME}-groopt-4ixodes.
pdb
LOOP_LIMIT=6
OPT_HYDRO_STEPS=1000
OPT_ENERGY_STEPS=1000
FINAL_SIZE=46

[Optional]

HYDRO_OPT_METHOD=Drippy
#methods: MD, Gradient
ENERGY_OPT_METHOD=Gradient
AVOID_COLLISIONS=False

#ile razy powtorzyc optymalizacje bialka przy tej samej wielkosci kropli
STEP_REPEAT_NUM=0
OPT_ENERGY_BLOCK_HELIX=False
OPT_HYDRO_BLOCK_HELIX=True
GET_MD_INTERMEDIATE=False

X_SIZE_PERCENT=0.7

OPTIMIZATION_METHOD=BRYLINSKI
OPT_HYDRO_TOLERANCE=0.001
MIN_HYDRO_BRYLINSKI = 0.002699
MIN_HYDRO_JADCZYK = 0.1302742


OPT_MODE=HYDRO_ENERGY
CODE_GEN_METHOD=d_dulak
#CODE_GEN_METHOD=z_baster
READY_ES_PATH=/mnt/gpfs/work/plgrid/groups/plggfaldki/ES/{PROTEIN_NAME}-d_dulak

OPT_ENERGY_TIME_PER_STEP = 0.001
OPT_ENERGY_BOX_PADDING = 1.0
OPT_PREP_MD_VACUUM_STEPS = 200
OPT_PREP_MD_WATER_STEPS = 500
```

Testing of multiple process parameters

- A particular process parameter can be defined in configuration file as a single value or a list of values
- Drippy Attack executes set of folding processes covering combinations of all declared values using parallel runnings on cluster queue system

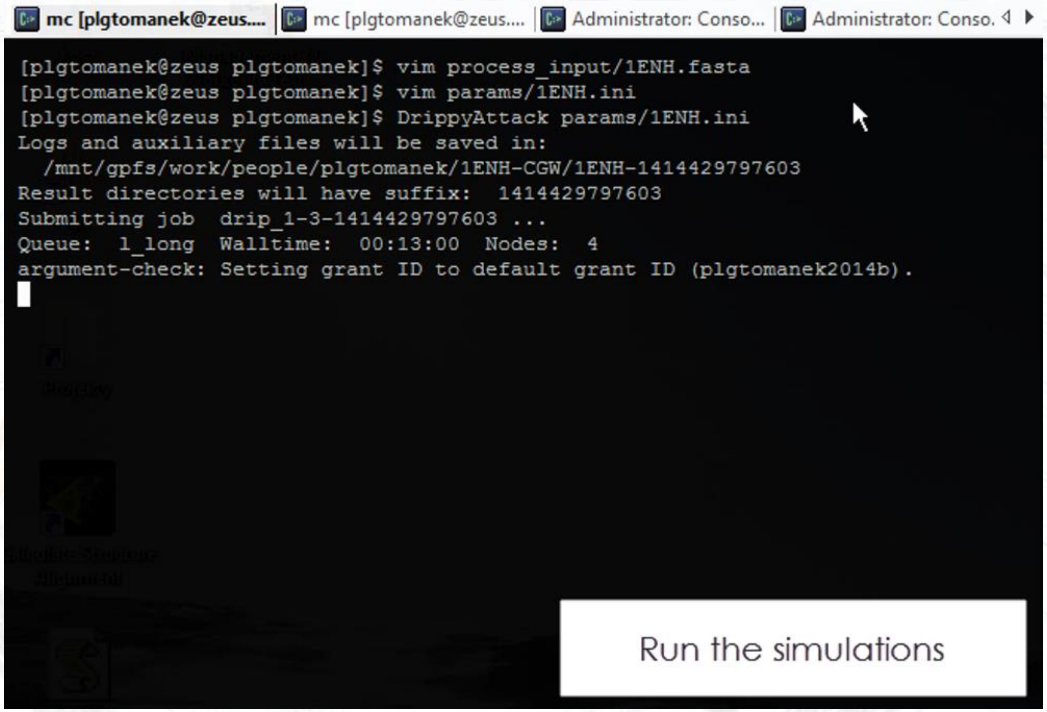


```
[Obligatory]
PROTEIN_PATH=/people/plgtomanek/drippy/process_input/1ENH.fasta
RESULT_PATH=/mnt/gpfs/work/plgrid/groups/plggfaldki/current/test
NATIVE_FILEPATH=/people/p
pdb
LOOP_LIMIT=6
OPT_HYDRO_STEPS=1000
OPT_ENERGY_STEPS=1000
FINAL_SIZE=46
```

```
[Obligatory]
PROTEIN_PATH=/people/plgtomanek/drippy/process_input/1ENH.fasta
RESULT_PATH=/mnt/gpfs/work/plgrid/groups/plggfaldki/current/test
NATIVE_FILEPATH=/people/plgtomanek/drippy/natives/{PROTEIN_NAME}-groc
pdb
LOOP_LIMIT=6
OPT_HYDRO_STEPS=1000;2000;3000
OPT_ENERGY_STEPS=1000
FINAL_SIZE=46
```

An Example Running

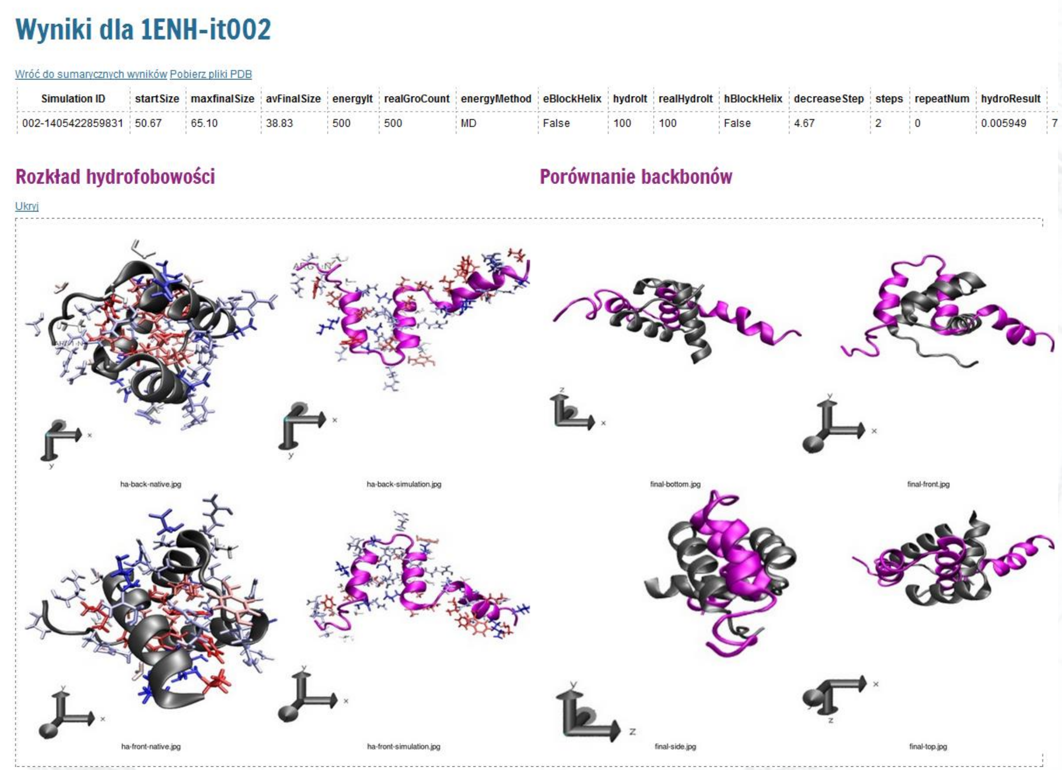
- Visit http://tinyurl.com/DrippyAttack-start to see short movie with preparation to simulation and launch of the Drippy Attack



```
[plgtomanek@zeus plgtomanek]$ vim process_input/1ENH.fasta
[plgtomanek@zeus plgtomanek]$ vim params/1ENH.ini
[plgtomanek@zeus plgtomanek]$ DrippyAttack params/1ENH.ini
Logs and auxiliary files will be saved in:
  /mnt/gpfs/work/people/plgtomanek/1ENH-CGW/1ENH-1414429797603
Result directories will have suffix:  1414429797603
Submitting job  drip_1-3-1414429797603 ...
Queue:  l_long  Walltime:  00:13:00  Nodes:  4
argument-check: Setting grant ID to default grant ID (plgtomanek2014b).
```

Prototype of the result viewer

- Visit http://tinyurl.com/DrippyAttackViewer to see short movie with result viewer presentation



Wyniki dla 1ENH-it002

Wróć do sumarycznych wyników Pobierz pliki PDB

Simulation ID	startSize	maxfinalSize	avFinalSize	energyIt	realGroCount	energyMethod	eBlockHelix	hydroIt	realHydroIt	hBlockHelix	decreaseStep	steps	repeatNum	hydroResult	
002-1405422859831	50.67	65.10	38.83	500	500	MD	False	100	100	False	4.67	2	0	0.005949	7

Rozkład hydrofobowości

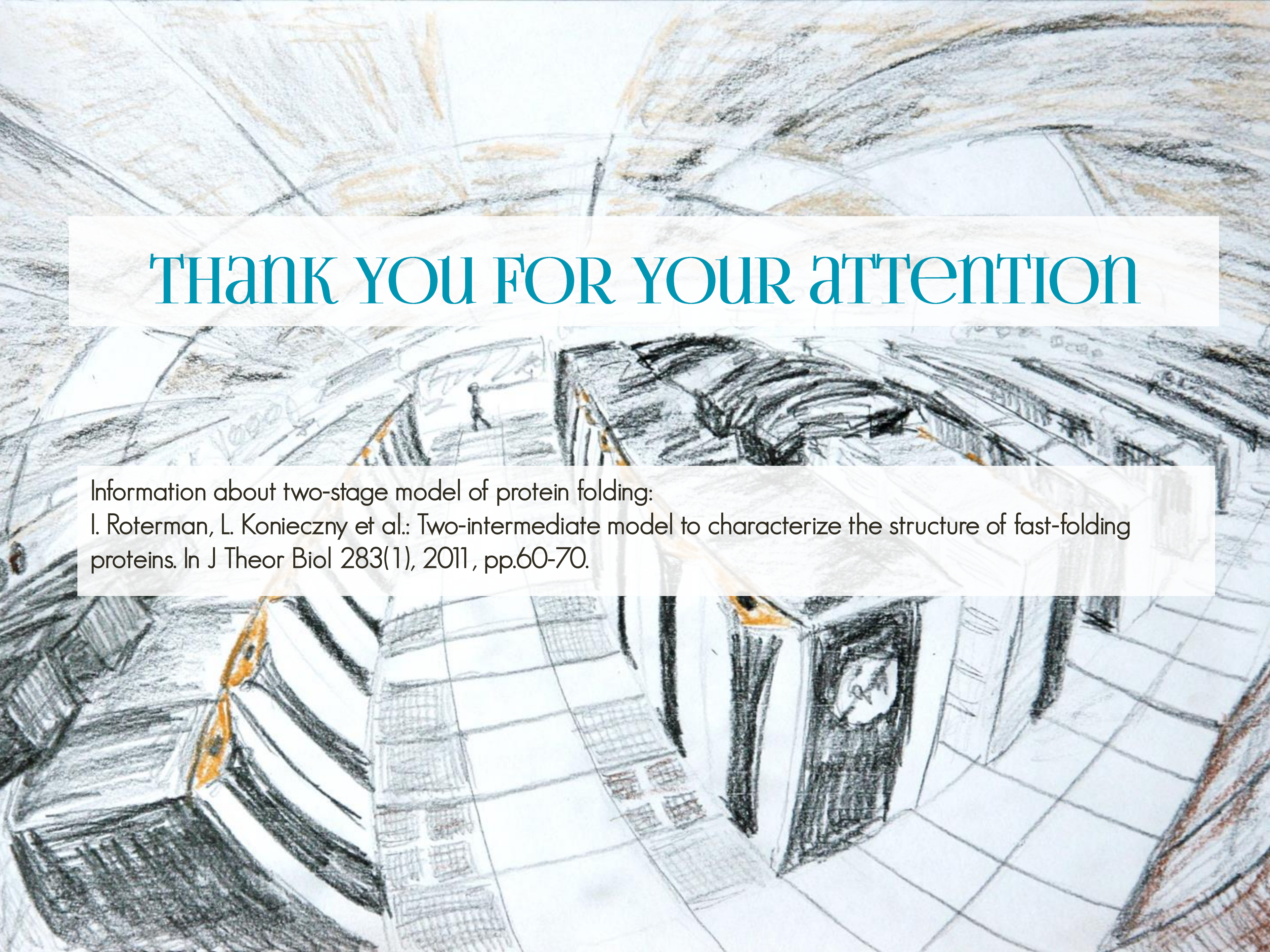
Ukryj

Porównanie backbonów

Conclusions and Future work

- Despite the fact that there are lots of general in silico experiment frameworks and portals, Drippy Attack can be an interesting option for the scientists who would like to test their hypothesis based on mentioned two-step protein folding model without lots of implementation work
- The user can use PLGrid infrastructure without knowledge of how to run jobs efficiently
- More comfortable module customization is currently being developed. The website for results visualization and results collector will be modified
- Drippy Attack will be available soon for PLGrid infrastructure users as a module on zeus.cyfronet.pl

PL GRID NG CYFRONET INNOVATIVE ECONOMY NATIONAL COHESION STRATEGY



Thank you for your attention

Information about two-stage model of protein folding:
I. Roterman, L. Konieczny et al.: Two-intermediate model to characterize the structure of fast-folding proteins. In J Theor Biol 283(1), 2011, pp.60-70.