Dziedzinowo zorientowane
usługi i zasoby infrastruktury
PL-Grid dla wspomagania
Polskiej Nauki w Europejskiej
Przestrzeni Badawczej

Managing protein folding process as workflow model with wise data selection

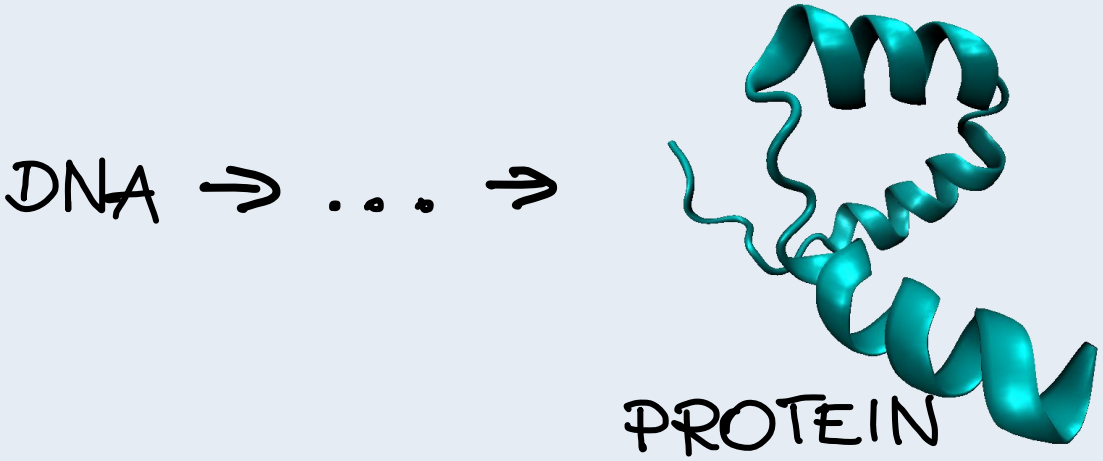
I. Roterman, M. Tomanek, M. Sterzel, T. Szepieniec, B. Kalinowska, Z. Baster, D. Dulak

ZBiT CM UJ, ACK Cyfronet AGH, WFAiIS UJ

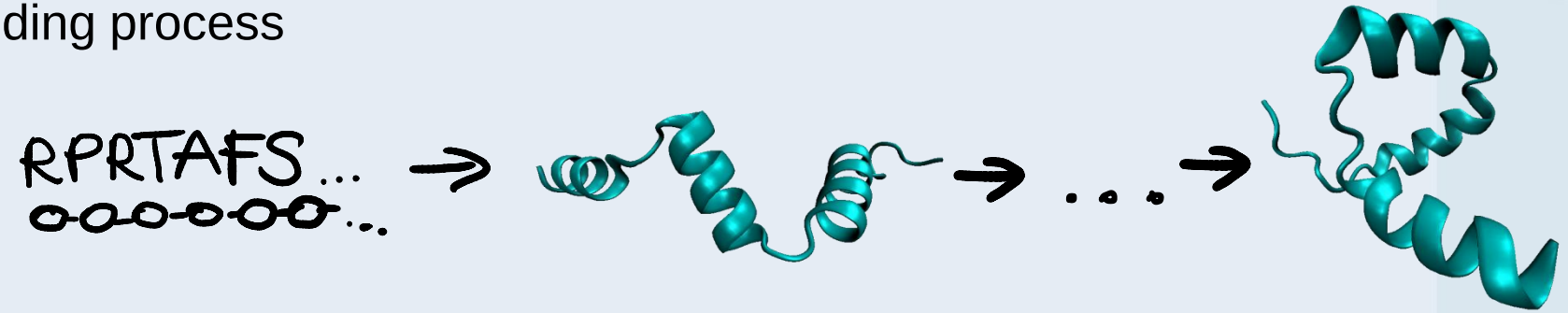
CGW 2012, Kraków, 22-24.10.2012

What is protein folding?

- Proteins are created based on information gathered in DNA

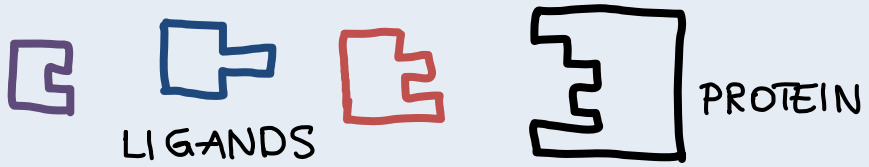


- During protein generation process, sequence of amino acids decoded from DNA shapes into 3D structure – this is the protein folding process

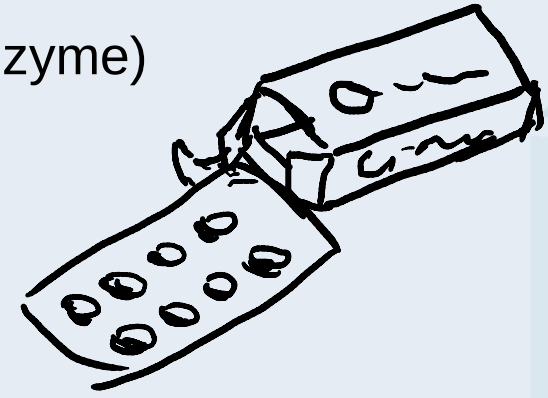


Why prediction of 3D structure is so important?

- Function of each protein depends on its 3D structure - **3D shape determines the specificity of protein** (ligand binding, protein complexation etc.)
- If 3D structure of protein is known:
  - Ligand/protein complexation can be predicted

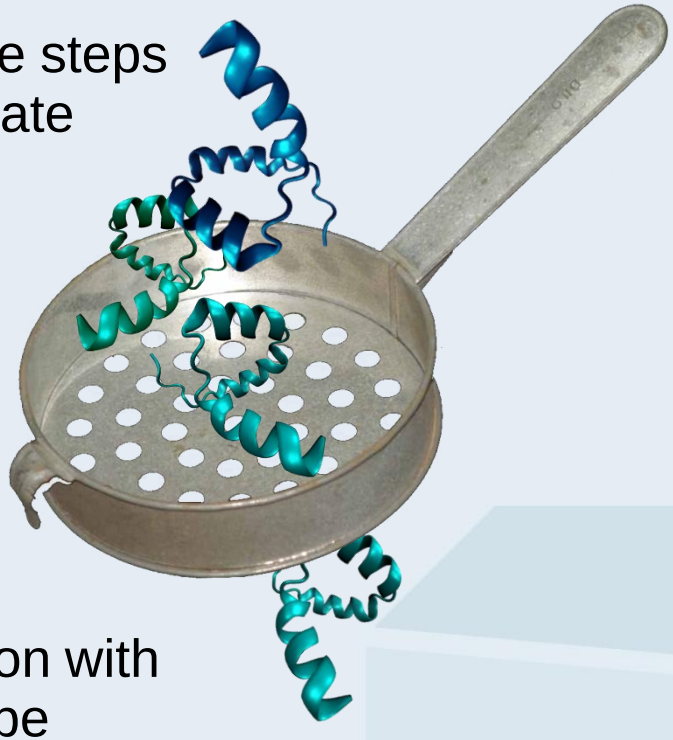


  - Biological function can be predicted (enzyme)
  - Drugs desing is possible



Model of protein folding process

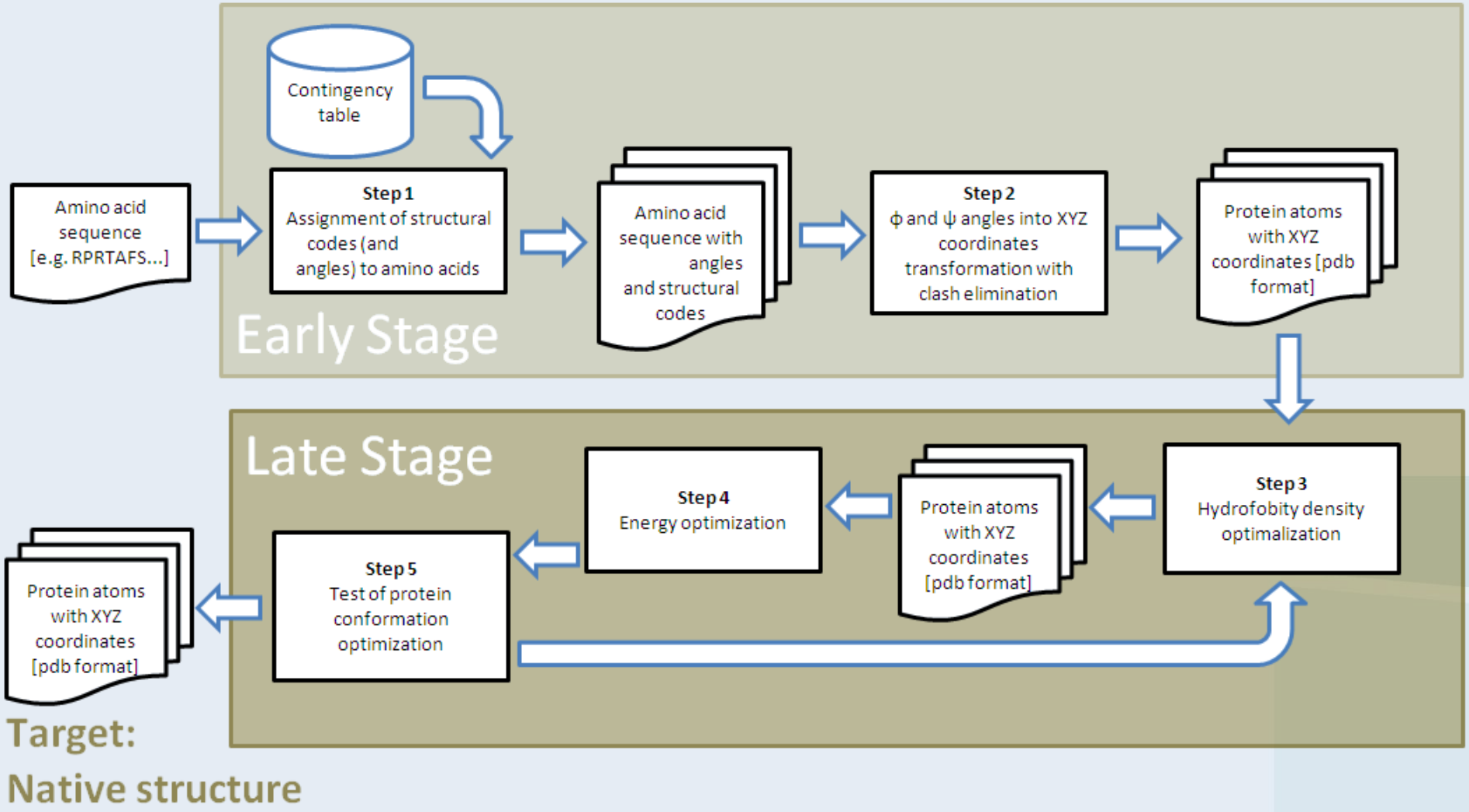
- …bases on the multi-step character of biological protein folding process
- …is composed of many steps and at each of these steps many (sometimes hundreds or thousands) candidate structures can be generated



One of the most crucial issuses in the model is prediction if a given structure is potentially a good candidate. So it is important to have:

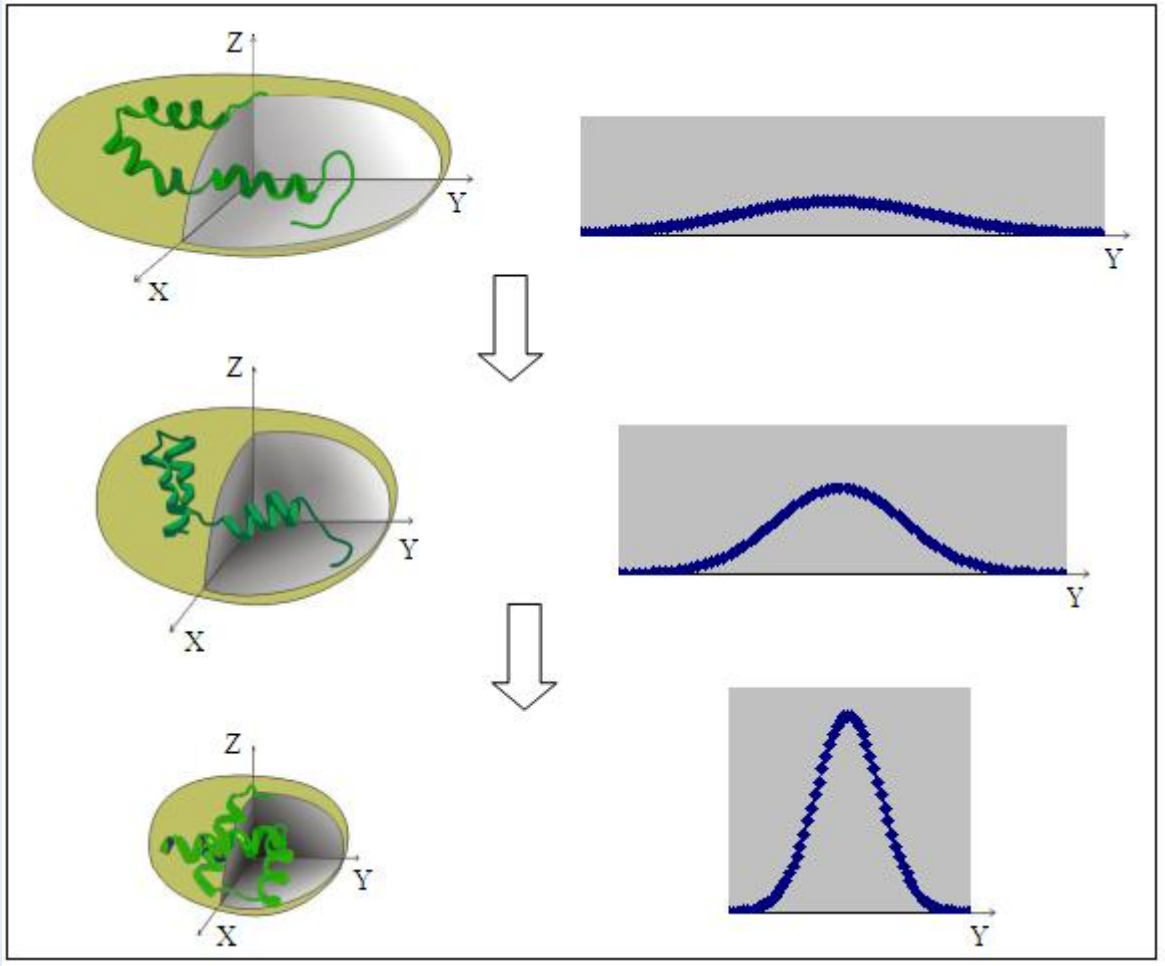
- adequate models of automatic data filtering
- convenient and fast candidate proteins presentation with manual tagging (because a part of data needs to be filtered manually)

Protein folding process flowchart



First results

- Hydrofobisity density optimization (Late Stage):



Problem: we do not know the best parameters for process at the beginning

We can quess and estimate good parameters for the specific protein but probably we can choose better parameters after process observation.

Sometimes process parameters should change according to current results according to some function.

Solution

- Readable visualisation of process products and configuration at each step with possibility of this configuration change
- Parameters not only as simple values but also as functions

MAX_ITERATION_NUMBER = BASE_ITERATION_NUMBER
+ (LAST_END_VALUE - LAST_START_VALUE) * MODIFIER

Problem: too many protein candidates

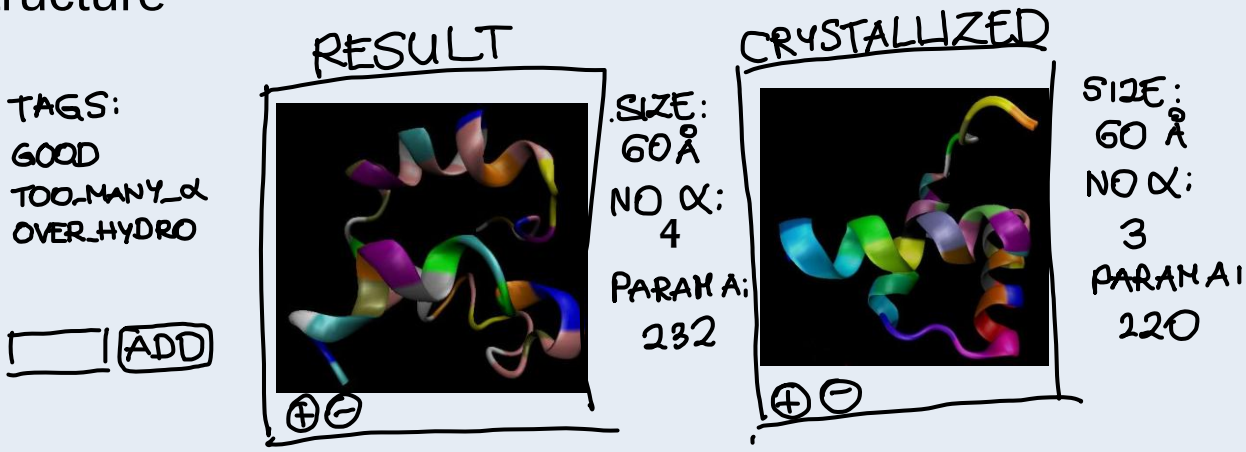
- Large amount of structure candidates generates even larger amount of stuctures in the next step

- Large amount of stuctures requires large amount of storage and computational space
- It is more difficult to check manually large amount of data

Solution for: too many protein candidates

Accurate automatic filtering methods and handy protein selection:

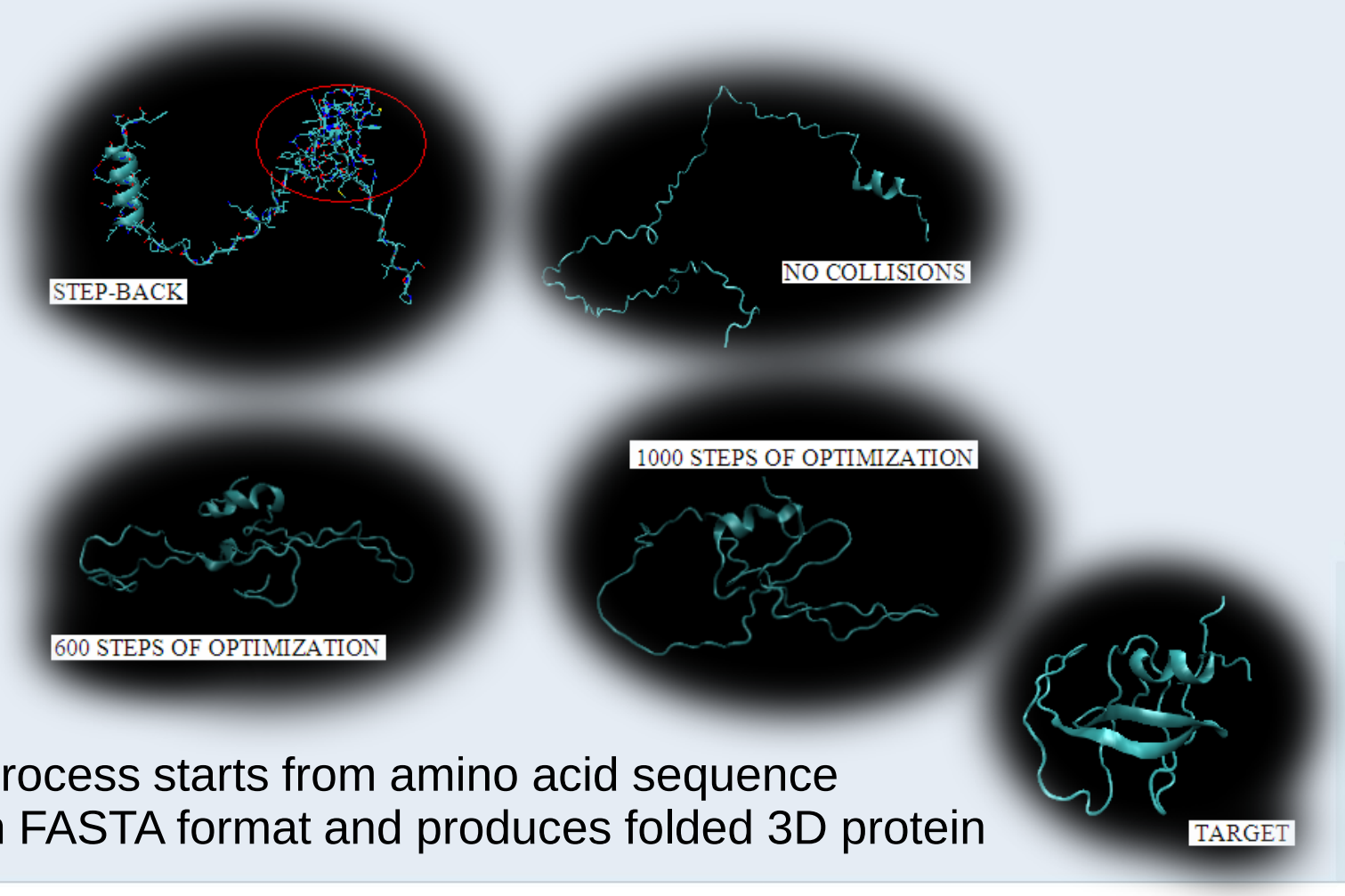
- User interface for compare protein finally received from the process with a crystallized form of the protein or with protein family using numerical parameters and visual form of 3D structure



- Tagging mechanizm to asign protein to class
- Tracking the previous form of achieved protein structures

First results

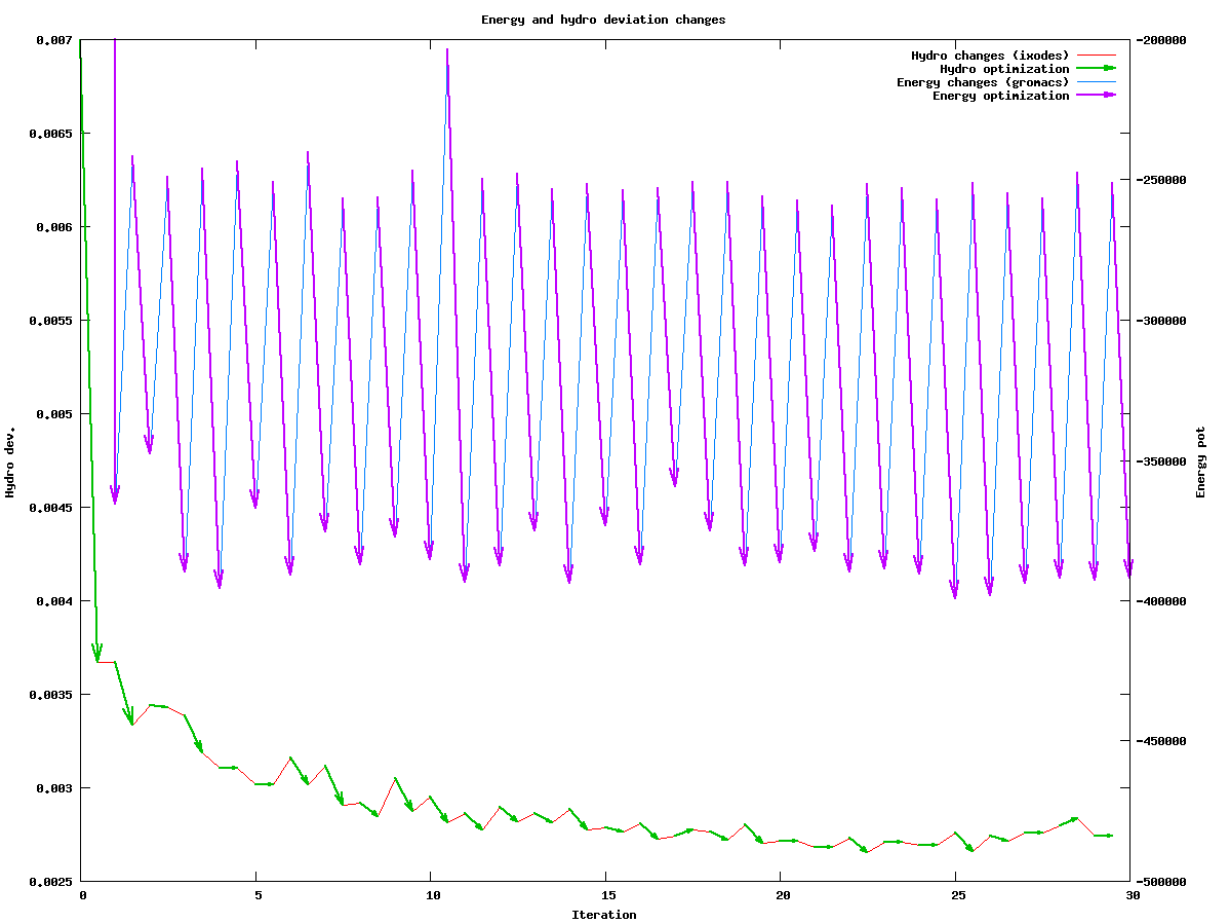
- Using existing applications, such as Gromacs, and application developed especially for this process purposes, a prototype of the process was created



- Process starts from amino acid sequence in FASTA format and produces folded 3D protein

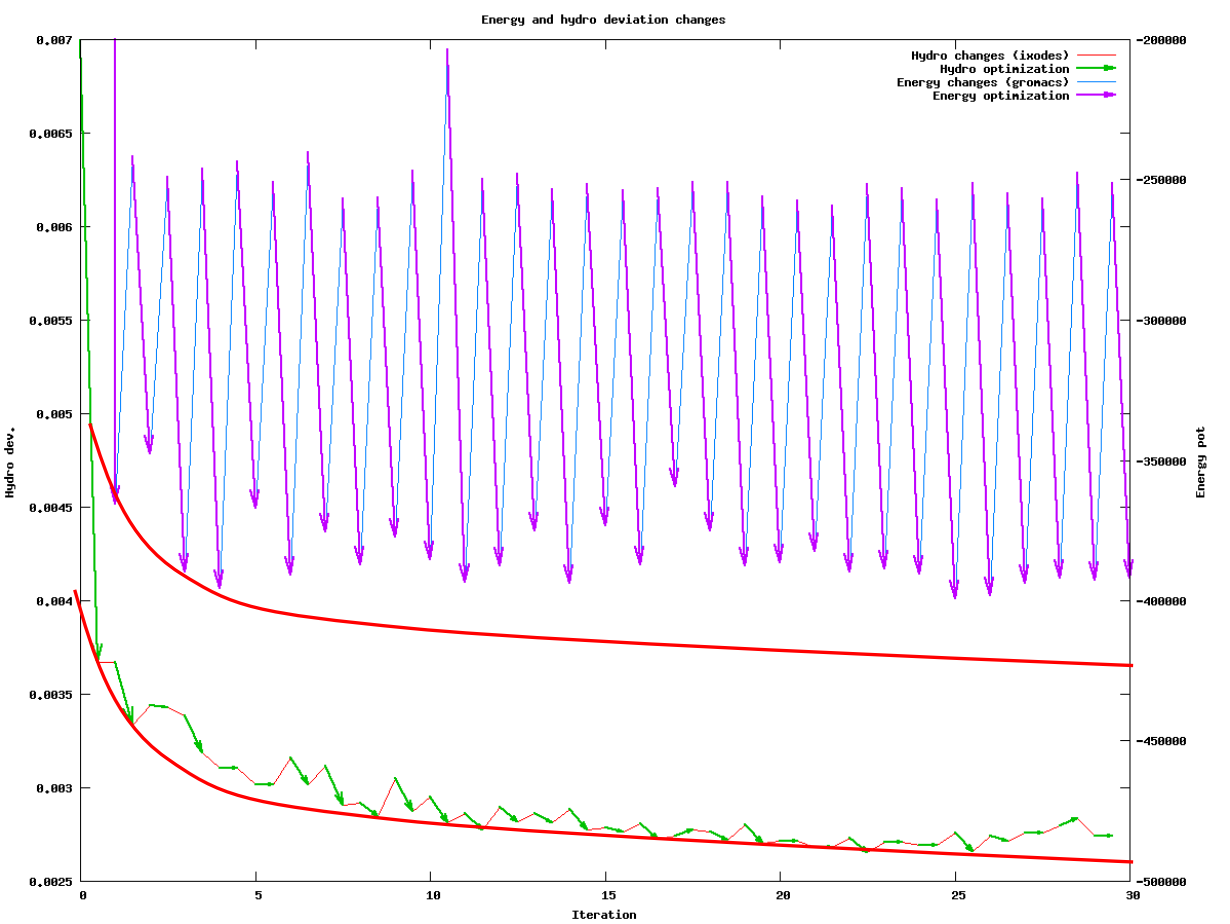
First results

- Protein optimization during Late Stage:



Target results

- Protein optimization during Late Stage – our target:



Important features of environment

- Visualization of the final and intermediate forms
- Interface for comparison of results with crystallized protein or with protein family properties
- Tracking results provenance
- Tagging mechanizm
- Possibility of process parameter modification
- Gathering process statistics and displaying them in clear summary form
- Automation of assignment of proteins to classes and protein filtration

Future work

- Development of user portal, which implements features important to the process
- …maybe using existing frameworks designed for development of services workflows (e.g. InSilicoLab, GridSpace2 from PL-Grid Plus project)
- Improvement of process using statistics from process


The End