Managing Protein Folding Process as Workflow Model with Wise Data Selection

Irena Roterman[1], Malgorzata Tomanek[2], Mariusz Sterzel[2], Tomasz Szepieniec[2], Barbara Kalinowska[3], Zbigniew Baster[3], Dawid Dułak[3]

[1] Department of Bioinformatics and Telemedicine – Collegium Medicum – Jagiellonian University, Lazarza 16, 31-530 Krakow, Poland
[2] ACC Cyfronet AGH, Nawojki 11, 30-950 Krakow, Poland
[3] Faculty of Physics, Astronomy and Applied Computer Science – Jagiellonian University, Reymonta 4, 31-059 Krakow, Poland
emails: myroterm@cyf-kr.edu.pl,m.tomanek@cyfronet.pl

Keywords: protein folding, hydrophobicity, protein structure prediction, process workflow

1. Introduction

Prediction of three dimensional protein structures is a major challenge to modern molecular biology. In order to simulate how a protein achieves three dimensional structure starting from an amino acid sequence, a model is created which bases on the multi-step character of biological protein folding process. In this abstract we describe a model of a system which enables such process realization.

2. Description of a problem solution and results

Process of protein folding – according to experimental observations [1] – is the step-wise process. The early and late step was introduced in our model as described in [2] and [3] respectively. During the early stage an amino acid sequence in form of one-letter code sequence (e.g. RPRTAFS…), using contingency table, is enriched with information about φ and ψ dihedral angles as defined in limited conformational sub-space what gives the protein 3D forms. In the next step clashes between atoms are eliminated. The protein in its pre-folded form is submitted to late stage model where its hydrophobicity density and energy is optimized. The late stage minimizes these two parameters keeping balance between them.

In order to realize the described process a dedicated environment is needed. The environment aims at arranging tools as presented on Fig.1 with possibility of checking and filtering results of each step. Implementation of biological and chemical processes in a workflow environment is a known issue which was discussed e.g. in [4]. Some existing frameworks, like Taverna, GridSpace2 or InSilicoLab, dedicated for development of services workflows, can be used as a base for our environment.

It is necessary to collect knowledge how to filter generated data to achieve accurate automatic filtering methods. Additionally, manual filtration can never be completely automated so handy protein selection interface is significant. Beginning from step 2 results can be presented in 3D space. User should have a possibility to compare the protein finally received from the process with a crystallized form of protein (if exists) or with a protein family using numerical parameters such as RMS-D and visual form of 3D structure.

Tracking the previous forms of achieved protein structure (its provenance) can provide important information. Grid resources from PL-Grid Plus [5] project allow store huge amounts of data files therefore results from each step can be gathered. User can tag intermediate forms of protein by assigning them to specific classes. Statistics from the process must be collected and presented to the user, e.g. which protein classes from previous steps had the best scores in the end and with which process parameters. This means that the process is flexible –

parameters critical to protein selection can be changed. After the process statistics are aggregated and analyzed in order to draw conclusion from them, assignment of proteins to classes can be partially automated or semi-automated using elimination of some portion of candidates from later process steps.

The aim is to create a process automated as much as possible, allowing to minimize the factor of user engagement.

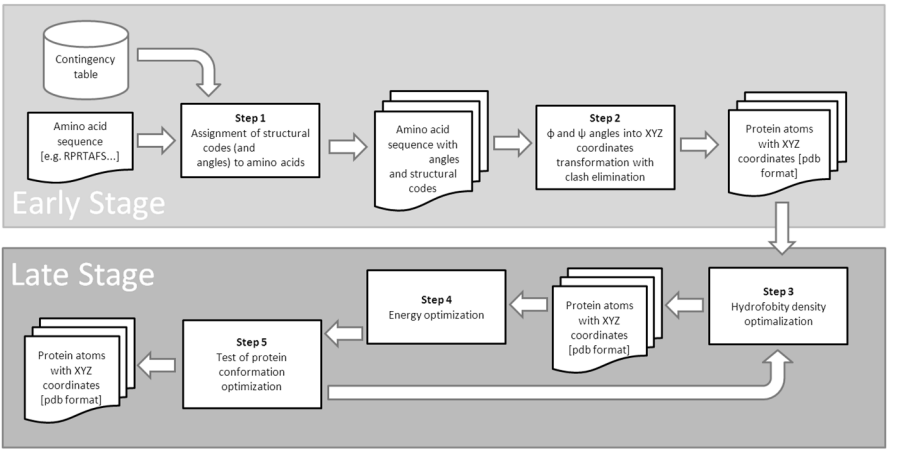

Fig. 1. Schema of protein folding process.

Using existing and especially developed applications a prototype of the process was created which produces folded 3D proteins starting from amino acid sequence in FASTA format.

3. Conclusions and future work

After analysis the problem model of process workflow and necessary environment features were defined. A simple prototype of the process was created. In the next step a user portal will be developed and will be continuously improved using statistics gathered in the process.

The correct definition of 3-D structure is an interesting problem by itself. However the solution of protein structure prediction plays critical role for computer aided drug design. This issue is of significant importance particularly now when individual personalized therapy is highly expected by medicine.

References

1. Creighton TE (1990) Protein folding. Biochem J 270, 1-16.
2. Bryliński M, Konieczny L, Kononowicz A, Roterman I (2008) Conservative secondary structure motifs already present in early-stage folding (in silico) as found in the serpine family. J Theor Biol 251, 275-285.
3. Roterman I, Konieczny L, Jurkowski W, Prymula K, Banach M (2011) Two-intermediate model to characterize the structure of fast-folding proteins. J Theor Biol 283(1):60-70.
4. Jadczyk T, Malawski M, Bubak M, Roterman I (2012) Examining Protein Folding Process Simulation and Searching for Common Structure Motifs in a Protein Family as Experiments in the GridSpace2 Virtual Laboratory. In: Bubak M, Szepieniec T, Wiatr K (Eds) Building a National Distributed e-Infrastructure - PL-Grid - Scientific and Technical Achievements. Springer 2012, ISBN 978-3-642-28266-9, 252-264.
5. PL Grid Plus, `http://www.plgrid.pl/plus`