AN APPROACH TO PROTEIN FOLDING ON THE GRID – EUCHINAGRID EXPERIENCE

M. Malawski**, T. Szepieniec**, M. Kochanczyk***, M. Piwowar***, I. Roterman***.

**Institute of Computer Science, AGH, Al. Mickiewicza 30, 30-059 Krakow, Poland
***Academic Computer Center CYFRONET, ul. Nawojki 11, 30-950 Krakow, Poland
***Department of Bioinformatics and Telemedicine, Jagiellonian University, Collegium Medicum, ul. Sw. Anny 12, 31-008 Krakow, Poland

Abstract: Contemporary pharmacology in its quest for more relevant and effective drugs needs to examine large range of biological structures to identify biological active compounds. We consider large grid environment the only platform to face such a computational challenge.

In our project, the search is focused on peptide-like molecules containing about 70 amino acids in a single polypeptide chain. The limited number of proteins existing in the nature will be extended to those, which have not been recognized in any organisms (“never born proteins”). The assumption is that those which do not exist in the nature may also render biological activity, which directed on pharmacological use may correct some pathological phenomena.

As the function results from the structure, two approaches are applied to predict cartesian coordinates of proteins’ atoms: sophisticated Monte Carlo structure creation, elimination and refinement using the Rosetta program and our own program for simulation of the protein folding process.

As a computing platform we use the EuChinaGRID project resources, which are currently a part of EGEE infrastructure and are expanding to include Chinese resources as well. We describe the approach for porting the application to the grid and the prototype portal developed for simulation management and results analysis.

Key words: protein folding simulation, grid system, pharmacology, drug design

Introduction

Fundamental research for individualized therapy

Contemporary pharmacology, which is expected to be ready to design the therapy in the individual manner for each patient, is facing a large challenge. The fast pace of drug design, which is assumed to satisfy all specific expectations related to particular disease and to particular patient, is the critical issue. The know-how in chemical disciplines seems to be developed on the satisfactory level. Computer equipment and software are also ready to be applied. The only missing link in simulation of biological processes is theoretical and then numerical ability to predict three-dimensional structure of protein. These are the molecules responsible for most of the processes in each living organism. It was evidenced that the function of protein molecule is determined entirely by its structure. This is why the search for reliable numerical models allowing correct structure prediction on the basis of known amino acids sequence is necessary to make the progress in fast new drug design aimed to correct disfunctional proteins. All steps of so called “central dogma of molecular biology”: from nucleic acids to biological function, seem to be recognized to the extent of understanding the mechanism of disfunction called disease. Instead of “drug design” understood as correction of proteins’ activity, the “therapy design” extends all over the steps of biological dogma and is placed in the focus of modern pharmacology. The processes of larger and larger systems need to be simulated in silico. The experiments (forbidden in vivo) are possible in silico and seem to be unlimited. The solution of this problem in its complete form seems to attainable in the relatively close future.

Simulation of the protein folding process

The only important step: accurate automated prediction of the three-dimensional structure is still unavailable in the in silico form. This is why many efforts are undertaken to solve this problem. This is also why the very large computer resources of grid-like size are exploited for protein structure prediction programs. Despite of 30-years long history, the correct model able to recreate the path according to which the unique spatial structure of polypeptide chain is formed, is still missing. The problem seems to be a hot one for life sciences nowadays. There are some tools like Rosetta which applied to particular amino acid sequence are able to suggest (only suggest) the native conformation of protein. The outcome of this method (treated so far as the best one) is the structure of limited confidence. The question: How do the proteins fold? remains still unanswerred. The probability based models are not able to give the reliable answer to this question. As long as the mechanism of folding is unrecognized, the models for its modification are also unavailable.
The “never born proteins” project
The task of massive protein structure prediction for 70 amino acids long polypeptides (10^7 of them) has been undertaken by the international team within IST EUChinaGrid Project [1]. The team joins experts from two disciplines: biochemistry specialists in protein structure prediction and computer science specialist in grid system. O predict the structure, the Rosetta method is applied and also a technique elaborated recently in JUCM [2], which is an attempt to simulate the protein folding process rather than protein structure prediction, is harnessed. Within the project the application was prepared to run on European and Chinese grid-resources.

In this paper we describe the protein folding application, focusing on the step which we needed for porting it to the grid. We also give the overview of the additional tools based on a portal, which were developed to simplify the process of management of running such a large-scale application on the grid and to aid biochemists interested in result analysis.

Related work
Distributed processing infrastructures such as grids or peer-to-peer systems has been used for protein folding since a relatively long time. The examples include early experiments using CHARMM software and the grid infrastructure [3]. There are also widely recognized projects that exploit the power of thousands of PC machines voluntarily offered by the participants via the BOINC platform clients [4]: pioneering Predict@home [5], Folding@home [6], that performs distributed molecular dynamics simulations, and Rosetta@home, which is powered by Rosetta software [7]. Predictor@home was recently taken offline, Folding@home concentrates on simulations of the folding pathways of single proteins etiologically related to specific diseases, and Rosetta@home is devoted to bringing to perform their jigsaw-puzzle-game-like method. The initiative of Human Proteome Folding Project [8], running on the infrastructure of Grid.org and World Community Grid and also using Rosetta software, produced a database of predicted structures of all human protein domains that were not yet resolved experimentally.

Running the application on the grid
The initial application comprised software developed by the JUCM team were prepared to run on a single machine or local cluster. When porting it to the grid such as EGEE, where the basic processing unit is a batch job, it is necessary to analyze application workflow in order to identify basic tasks and their data dependencies. The task should be possibly coarse-grained, since the overhead of job submission and batch system execution is considerable.

Stages of simulation
The JUCCM protein folding application consists of three main stages shown in Fig.1: early and late stage folding followed by the active site recognition. Given a sequence, creation of the early stage is entirely polypeptide backbone dependent [9] and requires a large contingency table with precomputed locations of tetrapeptides i a limited conformational subspace [10]. Additionally at the step, eventual steric clashes between distant amino acids are detected and resolved. The late stage works with such an intermediate structure by the introduction of side chain interactions that are extended by an external force field expressing hydrophobic character of some amino acids [11]. Its impact on the structure is evaluated as the discrepancy between actual and expected hydrophobicity, dependent on the distance from the centroid according to the gaussian distribution and assigned due to the own normalized scale distribution ("fuzzy-oil-drop"). Alternating with internal energy minimizations in the ECEPP/3 force field prevents atoms from overlapping. Distance relations between rigid elements of such small peptides hinder the molecule to cover itself thoroughly with hydrophylic residues providing hint for the location of the active site [12].

Fig.1. Stages of the protein folding process simulation

Steps for porting to the grid
After identification of logical stages of the application, it is necessary to consider also the technical side of the software, such as executables, library dependencies, and input/output files. The following steps were needed to grid-enable the folding application.
1. For all programs used in workflow all the required packages, which were not available on the grid worker nodes, were collected. For example, the library of sequences required for early-stage and code dependencies for late-stage were included.
2. The main script was created for running the application. It is responsible for proceeding with workflow execution checking if results of each stage are available. The parameters of this script are a sequence string and an identifier of the sequence.
3. It was decided to register results of computation for single sequence in a separate file on the grid storage, namely in LFC catalogue. The file name includes the sequence identifier and the resulting protein is stored in PDB format.
4. A self-containing bundle of programs and libraries needed for executing the application was created. This bundle was also registered in LFC catalogue.
5. A script was created to prepare installation of application on site each time when job is started and to spawn the main scripts with appropriate parameters.
6. Finally, the JDL (Job Description Language) file for gLite middleware was created. Performing these steps resulted in a self-contained application, which can be executed on the grid infrastructure without any pre-installation required. This is especially convenient for running it on many virtual organizations within EGEE (such as Euchina and VOCE VOs) and also on Chinese grid infrastructure.

UI/Portal

As the number of simulation tasks and produced structures is huge and the way they are processed and finally interpreted is homogenous, we have developed a portal for job submission and monitoring and for data analysis, that appreciably simplifies the interaction of the average user with the complex infrastructure of the grid. Portal was developed in GridSphere Portal [13] using GridwiseTech LCG-API package [14] to cooperate with grid infrastructure.

Job submission and monitoring

In Fig. 2 the most important features of the portal and data-flow in the application were presented. Using the application portal job submission is performed (step 1). Typically this is done by uploading a file, in which up to several thousand sequences are listed with their identifiers. The portal creates a separate grid job for each sequence and it adds them to the submission queue. Jobs from the queue are submitted to grid using LCG-API Job Monitor. This is done according to specified policies that could prevent from flooding VOs with to many grid jobs. A single job running on grid computing element download an application package to the working node (step 2), compute the results and save them to the grid storage system (step 3). Results of the jobs are validated by the portal with a post-processing analysis routines. In case of positive validation the results are registered in the results database, otherwise decision what to do next are left for the portal operator. Portal services analyze also the results of jobs failures and decide whether to resubmit a job or rather to ask the operator what to do next.

At runtime the operator can monitor computations using the application portal. Monitoring in the portal was designed to face the large amount of jobs running at the same time. The portal implements features like grouping, browsing by various criteria, viewing statistic and listing of current problems-to-solve.

Finally, basing on database, the results of computations can be browsed and accessed (step 4) for analysis by a set of tools described in next paragraphs.

Result analysis

In the part of the portal devoted to the result analysis we provide conventional tools that are familiar to biochemists and biophysicists dealing with proteins. After choosing the id of a resulted structure, remote secondary structure assignment is performed remotely in DSSP [15] and presented graphically.

Using JUMC Structural Bioinfo Toolkit, that operates on the server-side and generates images to the virtual framebuffer using Java2D, the A/A map with preferred areas and contact map with different distance cut-offs are displayed.

We also reengineered the MBT Protein Workshop [16] in order to enable immediate visualization in the classic cartoon-like representation. On the basis of our Toolkit we developed a specialized molecular viewer that is able to point the location...
of the probable active site using a color scale. Molecular surface is computed remotely in MSMS [17] and retrieved via Java RMI. If a protein was synthesized in a wet biology laboratory and has undergone a 2D electrophoresis, within the portal it is also possible to get an estimate location of the molecule in the gel (portlet with a Curl wrapper to the ExPASy.org service).

Fig. 3. Part of the portal for result analysis. Molecular viewers can be launched via JavaWebStart.

Summary and future work

The possibility to fold the proteins on such a scale applying two different methods is the great opportunity to test both of them. The mutual comparison of obtained results (according to Rosetta in cooperation with University Roma Tre [18] and basing on our mechanistic approach) is assumed to help to understand the nature of proteins in respect to their behavior in natural environment.

Moreover, the possible synthesis of the protein of assumed as pharmacologically active allows (recognized on the basis of predicted structure) immediate verification of obtained computational results (experimental partners in the Beijing University) and laboratory tests of harnessing them as potential new drugs.

The approach to running the application on the grid was tested on a sample batch comprising 10000 sequences and the prototype portal was used for demonstration purposes. Current work focuses on the development of a database for management of simulations and improving the usability of portal. Performing more tests will allow to verify both the simulation model and our portal toolkit.

Acknowledgements

This work was partly funded by the European Commission, Project EUChinaGRID and by the related Polish SPUB-M Grant. Maciej Malawski kindly acknowledges the support from the Foundation for Polish Science.

References