Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Editorial**

E-infrastructure technologies triggering of Bioinformatics development

Irena Roterman[1,] *

[1]Department of Bioinformatics and telemedicine, Collegium Medicum – Jagiellonian University
31-530 Krakow, Lazarza 16, Poland; Irena Roterman* - Email: myroterm@cyf-kr.edu.pl; * Corresponding author

received November 13, 2007; published online December 05, 2007

Editorial Message:

Bioinformatics, the history of which is abort 20 years long, has been initiated as the discipline supporting the genome sequencing project. Bioinformatics was planed to collect and store the nucleotide sequences deposited by laboratories engaged in sequencing the nucleotides in genetic materials of selected organisms including *Homo sapiens*. New e-infrastructure technologies triggered significant steps in bioinformatics technologies. Therefore, we shall look carefully at trends in new e-infrastructure while they would bring new opportunity and simplify the every-day work in bioinformatics. The probability calculus and information theory based discipline – genomics – appeared to be the independent scientific field raising independent scientific problems, creating tools to solve them and even influencing the experimental work to speed up the progress in the research producing the complete human genome earlier than expected.

The NCBI database available for all people on the Earth is equipped with special tools to perform the similarity search for nucleotide sequences, localizing them in particular positions in genome under consideration and selecting the diseases related to particular mutations. The proteomics focused on the world of proteins seems to require more differentiated specializations to collaborate with. The aminoacids sequence comparison is based on similar tools as genomics. Although the protein structure prediction requires the knowledge of optimization techniques to search for low energy structural forms, the differential equations to run the molecular dynamics simulation and matrix calculus to perform any structural conformational changes.

The protein structure prediction is the most exciting specialization in protein science. The polypeptide of particular aminoacids sequence accepts in the cell the unique three-dimensional structure, which is able to demonstrate its biological activity. This phenomenon is still the secret of nature. Despite of many efforts to recognize the method of polypeptide folding, this process is unrecognized. The results of CASP (critical assessment of protein structure prediction) experiments organized every second year (starting in 1995), which focuses top teams specialized on protein structure prediction are unsatisfactory. After some successes of *ab initio* methods (called Boltzmann model nowadays), the comparative techniques (called Darwin-theory-based nowadays) are on the top of ranking list estimating the similarity of model versus the crystal structure. The success in comparative modeling is unfortunately not optimistic due to high uncertainty of prediction and high sequence similarity necessary to recognize the type of fold. The Darwin-based approach works on the premise that sequences of common ancestor fold to similar structural forms.

The suggestion is that the “protein structure prediction” shall rather change to the “protein folding simulation”. The difference between these two approaches emphasizes the complexity of protein structure formation as the result of the multi-step process rather than one-step procedure oriented solely on energy optimization. Integral part of protein structure analysis is the biological function recognition. Each protein in living organism is created to play particular highly specific role. The specific three-dimensional structure seems to be aim-oriented to ensure the high specificity biological activity. This is why the automatic method for recognition of biological function is highly expected for basic reasons (recognition of the Nature's secret) as well as for practical purposes to identify the biological function of proteins produced according to genes recognized *in silico*.

The proteomics focused on individual protein changed its view recently to the whole set of proteins in living organisms. The new discipline called *Systems Biology* is oriented on the construction of the system simulating the complete set of processes with all couplings and mutual dependencies between processes and proteins playing particular roles in the organism. The creation of proteome (the complete set of proteins together with their biological functions and mutual relations) simulating the perfect harmony in the living organism (including bacteria and highly developed organisms) could give the excellent possibility to simulate the pathological processes without the necessity of experimental examination of these phenomena. The analysis of results of the aim-oriented introduction of disorder is assumed to shade the light on highly complexed pathological effects including cancer.

The next discipline incorporated into bioinformatics is the computer aided drug design. The simulation of interaction between protein and molecule assumed to play the pharmacological role makes possible the large scale search for biologically active compounds. For the last few years, we are witnessing the birth of another epoch in computations. The large computing centers connected together provide integrated service for computations and data storages. Available though uniform interfaces, computing grids, aims to provide as easy

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Editorial**

access to computation as electric grid provide access to electricity to socket in our wall. Idea of joining resources provide for the user larger computational power available as an effect of better management of the pool. Additionally, in the grid, it is more convenient to organize computational projects that would be not feasible without a unified environment because of organizational overheads. Bioinformatics is the consumer of large computing resources running the calculations oriented on protein structure prediction enabling finding the method to project the protein of expected structure and preliminary defined biological activity. The pharmacology is expecting the individual pharmacological therapy for each patient individually to minimize the site effects due to the specificity of each human organism. The individual pharmaco-therapy is the task for modern pharmacology. The developed tools for *in silico* simulation of biological processes on molecular level (protein-drug interaction) together with complete proteom simulation make possible the design of individual therapy. The large scale computing is necessary to reach all these aims. The availability of large scale computer resources is critical for developing these tools and run the quite complexed calculations.

The importance instance of grid environment is the grid maintained by EU Project EGEE. The grid gathers more than 200 installations distributed world wide, and provides stable resources at level of 40k CPUs and 50PB (1 PetaByte equals 1 Mega GB). In this environment users are grouped in, so called, Virtual Organization, with possibility of sharing the same pool of resources as well as data stored in the grid. This could open a new dimension of collaboration between researchers groups located in different places.

Undeniably, grids could change the way of making researches in bioinformatics, however transferring the specific software and tools, as well as every day habits is not straightforward. Some effort in software adaptation is required, hopefully EU support this kind of activity funding projects that supports new applications in grid environment. Example of this kind of project is EUChinaGrid, which focus in specific issues on technical integration between European and Chinese computational grid, as well as on supporting some application that researches from both countries show interest in collaboration.

The search for new drugs, as was presented above, which is highly computing time consuming discipline, was also implemented into grid system. The project oriented on “Never Born Proteins” treated as potential storage of potential biological activity together with the grid infrastructure construction was the aim of EuChinaGrid project financially supported by European Commission (FW6 STREP LSH 2003-1.1.0-1 contract 026634). Two tasks were assumed to be accomplished: unification of the grid system working in China and Europe and the search of new pharmacologically active proteins (70 aminoacids in polypeptide). The 27 months lasting project (the project is planned to be accomplished in March 2008) resulted in cooperation of two grid systems (China and Europe). The biological project assumed to fold (*in silico*) the 70 aminoacids containing polypeptides of sequence which has not been created by evolution (Never Born Proteins). Since the method to predict the structure of proteins is not available yet, two methods were applied to construct the three-dimensional structure of polypeptides: ROSETTA and one newly elaborated, which assumes to simulate the folding process introducing the early- and late-stage of folding (calculation performed by European partners). The biological function of these proteins (for each structural form independently) Proteins (there are 10 million sequences generated) of similar 3-D structures were selected to be experimentally expressed and structurally analyzed by NMR. The experimental part is assumed to be performed by Chinese group.

The partners participating in the project are waiting for experimental measurements which will clarify, whether any of the two used methods is able to approach the real proteins for particular polypeptide and to demonstrate the predicted (expected) biological activity. The conclusion for bioinformatics that emerge is rather obvious: we should observe the development of e-infrastructure and e-science technologies like grid environments not only for using them as it would be ready, but also to actively influence their development to meet bioinformatics requirements. Development of the environment could open new perspective for bioinformatics, like the individually projected therapy. Taking advantages from grid technology it would be possible to approach this task in a relatively close future.

Editorial: I. Roterman

Citation: Roterman, Bioinformation 2(4): 126-127 (2007)