RUSSIAN ACADEMY OF SCIENCES
INSTITUTE OF CYTOLOGY AND GENETICS SB RAS
NOVOSIBIRSK STATE UNIVERSITY

PROGRAM OF THE THIRD INTERNATIONAL SCHOOL "EVOLUTION, SYSTEMS BIOLOGY AND HIGH PERFORMANCE COMPUTING BIOINFORMATICS"

Novosibirsk
June 28 – July 2, 2008

INTERNATIONAL PROGRAM COMMITTEE

- Dr. Dmitry Afonnikov, Institute of Cytology and Genetics, Novosibirsk, Russia (Conference Scientific Secretary)
- Prof. Ralf Hofestadt, University of Bielefeld, Germany (Co-Chairman)
- Prof. Alexis Ivanov, V.N. Orekhovich Institute of Biomedical Chemistry, Russian Academy of Medical Sciences, Moscow, Russia
- Prof. Nikolay Kolchanov, Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia (Chairman)
- Dr. Olga Krebs, EML Research gGmbH, Heidelberg, Germany
- Prof. Thomas Ludwig, Ruprecht-Karls-Universitat Heidelberg, Institut fur Informatik, Heidelberg, Germany
- Prof. Victor Malyshkin, Institute of Computational Mathematics and Mathematical Geophysics, Novosibirsk, Russia
- Dr. Luciano Milanesi, National Research Council - Institute of Biomedical Technology, Italy
- Prof. Dmitry Sherbakov, Limnological Institute SB RAS, Irkutsk, Russia
- Dr. Alexandros Stamatakis, The Exelixis Lab Teaching and Research Unit Bioinformatics Department of Computer Science Ludwig-Maximilians-University Munich

LOCAL ORGANIZING COMMITTEE

- Ekaterina Denisova, Institute of Cytology and Genetics, Novosibirsk
- Dr. Galina Kiseleva, Institute of Cytology and Genetics, Novosibirsk;
- Sergey Lavryushev, Institute of Cytology and Genetics, Novosibirsk (Chairperson);
- Dmitry Oschepkov, Institute of Cytology and Genetics, Novosibirsk ;
- Dr. Natalia Sournina, Institute of Cytology and Genetics, Novosibirsk;
- Svetlana Zubova, Institute of Cytology and Genetics, Novosibirsk; and
- Yana Kolodyazhnaya, Institute of Cytology and Genetics, Novosibirsk.

ORGANIZERS

Bielefeld University

EML Research gGmbH

Institute of Cytology and Genetics,
Siberian Branch of the Russian Academy of Sciences

Novosibirsk State University

Laboratory of Theoretical Genetics

Chair of Information Biology

German/Russian Virtual Network of Bioinformatics
"Computational Systems Biology"

HP Technology Education and Research Center
German/Russian Virtual Network of Bioinformatics
"Computational Systems Biology"

Institute of computational technologies

PBSoft Ltd.

SPONSORS

The Klaus Tschira Foundation gGmbH

Russian Foundation for Basic Research

Federal Agency for Science and innovation

International Summer School 2008
Evolution, Systems Biology and High Performance Computing
Novosibirsk, June 28-July 2, 2008

Timetable

Saturday, 28 June

11.00 – 14.00 Registration at the Institute of Cytology and Genetics

14.00 Opening Ceremony

14.10 – 15.45 Lecture 1.

Prof. Falk Schreiber, IPK Gatersleben, Germany.

Analysis of biological networks and related data.

16.00 -17.45 Lecture 2.

Prof. Ralf Hofestädt, Bielefeld University, Germany.

Metabolic network analysis

18-00 – 19.35 Seminar 1.

Mr. Yury Vyatkin, Institute of Cytology and Genetics, Novosibirsk, Russia.

Introduction to HPC.

Sunday, 29 June

9.30 – 11.05 Lecture 3.

Prof. Dmitry Scherbakov, Limnological Institute SB RAS, Irkutsk, Russia.

Discrete models for molecular evolution simulation at the population level.

11.25-13.00 Lecture 4.

Dr. Luciano Milanesi, CNR-Institute for Biomedical Technologies, Milan, Italy

Bioinformatics applications by using the GRID computing technology.

Lunch

15.00 – 16.35 Lecture 5.

Dr. Srinivasan Ramachandran, Institute of Genomics and Integrative Biology, Delhi, India

Application Bioinformatics for infectious diseases

17-00 – 18.35 Seminar 2.

Dr. Daniil Naumoff, GosNIIgenetika, Moscow, Russia.

Practical course on the protein sequence analysis: searching for homologues and their hierarchical classification

20-00 – 24-00 Cultural Program. Evening Novosibirsk excursion. Fireworks festival dedicated to Novosibirsk 115th annivesary.

Monday, 30 June

9.30 – 11.05 Lecture 6.

Prof. Alexis Ivanov, V.N. Orechovich Institute of Biomedical Chemistry RAMS, Moscow, Russia.

Platform «From Gene to Lead Compound»: integration *in silico* and *in vitro* technologies

11.25-12.20 Lecture 7.

Dr. Alexey Pylkin , Sun Microsystems.

To be announced

12.30-13.45 Lecture 8.

Dr. Olga Krebs, EML Research gGmbH, Heidelberg, Germany.

Integration of reaction kinetics data: and modeling of metabolic networks: SABIO RK and SYCAMORE.

Lunch

15.00 – 16.35 Lecture 9.

Dr. Stefan Heinrich, EML Research gGmbH, Heidelberg, Germany.

PIPSA: Comparison of protein interaction properties.

17-00 – 18.35 Seminar 3.

Dr. Olga Krebs, Dr. Stefan Heinrich, EML Research gGmbH, Heidelberg, Germany.

Practical course on SABIO RK and SYCAMORE.

19-00 Cultural Program. Foolball match between ICG and non-ICG participants dedicated to EURO-2008.

Thursday, 1 July

9.30 – 11.05 Lecture 10.

Prof. Thomas Ludwig, Ruprecht-Karls-Universitat, Heidelberg, Germany.

Introduction to high-performance computing. (1) Architecture of high performance computers

11.25-13.00 Lecture 11.

Prof. Thomas Ludwig, Ruprecht-Karls-Universitat, Heidelberg, Germany.

Introduction to high-performance computing. (2) Parallel programming principles

Lunch

15.00 Young scientists presentations.

Wednesday, 2 July

9.30 – 11.05 Lecture 12.

Prof. Thomas Ludwig, Ruprecht-Karls-Universitat, Heidelberg, Germany.

Introduction to high-performance computing. (3) Message passing with MPI

11.25-13.00 Lecture 13.

Prof. Thomas Ludwig, Ruprecht-Karls-Universitat, Heidelberg, Germany.

Introduction to high-performance computing. (4) Advanced issues with message passing

Lunch

15.00 – 16.35 Lecture 14.

Dr. Alexandros Stamatakis, Ludwig-Maximilians-University, Munich, Germany

Models, Algorithms, and Parallel Computing for Large-Scale Phylogenetic Inference

17-00 Closing Ceremony

Young Scientists presentation program.

Oral presentations. (Time for presentation: 15 min + 5 min for questions)

I. Anischenko, Belarusian State University, Minsk, Belarus. HIV-1 gp120 v3-loop comparative structure analysis: search for the structurally conserved regions.

E. Chaplygina, Institute of Applied and Fundamental Medicine, Nizhny Novgorod State Medical Academy, Nizhny Novgorod, Russia. Phylogenetic classification of the HECT-domain ubiquitin-protein ligase family

N. Eltsov, Institute of cytology and genetics SB RAS, Novosibirsk, Russia. The new algorithm for phylogenetic reconstruction of non-recombining DNA sequences.

V. Fazalova, Limnological Institute SB RAS, Irkutsk, Russia. Individual-based modelling of adaptive speciation in spatially structured populations.

Yu. Grushetsky, United Institute of Informatics Problems of NASB, Minsk, Belarus. Prediction of protein interactions using homologous interfaces.

I. Kulakovskiy, Engelhardt Institute of Molecular Biology RAS, Mocsow, Russia. Incorporating different types of experimental data on DNA-protein binding into the single *in silico* model.

Yu. Medvedeva, GosNIIgenetika, Moscow, Russia. Reduced level of synonymous substitution in CpG containing codons suggests functional role of intragenic and 3' CpG islands in human genes.

A. Nyporko, Institute of Cell Biology and Genetic Engineering of NAS of Ukraine, Kiev, Ukraine. Influence of amino acid replacements associated with multidrug resistance on beta-tubulin molecular dynamics.

Poster presentations

V. Kovalyov, Institute of Applied and Fundamental Medicine, Nizhny Novgorod State Medical Academy, Nizhny Novgorod, Russia. UBIQUITOMIX database: a new resource on ubiquitin system.

N. Prakhov, University of Nizhni Novgorod, Nizhni Novgorod, Russia. Validation of viral EBNA1 protein as the perspective target for drug discovery.

LECTURES AND PRACTICAL COURSES

Analysis of biological networks and related data

Prof. Falk Schreiber
IPK Gatersleben, Germany

This talk will give an overview of the structural analysis of biological networks, located at the interface of biology and computer science. Biological networks represent processes in cells, organisms, or entire ecosystems. Large amounts of data which represents (or is related to) biological networks have been gathered in the past, not least with the help of the latest technological advances. Thus, the analysis of these networks is an important research topic in modern bioinformatics,and the analysis of biological networks is gaining more and more attention in the life sciences and particular in the growing field of systems biology.

Network analysis methods for biological networks are presented and discussed. This includes global network properties and network models, centrality analysis which helps in ranking network elements, and network motifs which can represent potentially important network parts and clustering methods. Furthermore we discuss the analysis of -omics data (e.g. transcriptomics, proteomics, and metabolomics). To support an integrative, systems biology directed approach, the interactions of the biological entities (e.g. DNA, RNA, proteins, metabolites) are important and the data has to be linked to relevant networks. We will discusses methods for the visualisation and analysis of networks with related experimental data and presents VANTED, a system implementing these methods. Different data such as transcript, enzyme, and metabolite data can be integrated and presented in the context of their underlying networks, e. g. metabolic pathways or classification hierarchies such as gene ontology. Statistical methods allow analysis and comparison of multiple data sets. Correlation networks can be automatically generated from the data and substances can be clustered according to similar behavior over time. Sophisticated visualisation approaches support an easy visual analysis of the data enriched networks. VANTED is available free of charge at http://vanted.ipk-gatersleben.de/.

Metabolic network analysis

Prof. Ralf Hofestaedt
Bielefeld University, Germany

Currently, there are about 1000 database and information systems and various analysis tools available via the internet. The challenge we have, is to integrate these list-parts from genomics and proteomics at novel levels of understanding. Integrative bioinformatics would be this new area of research using the tools of

computer science applied to biotechnology. Finally, these tools will represent the backbone of the concept of the virtual cell, which is both, a scientific vision and challenge of bioinformatics. This talk will present the architecture of a federated database concept for the integration of metabolic database systems. Moreover, behind the prediction of networks we will discuss the modelling and simulation of metabolic networks using automata and Petri nets.

Discrete models for molecular evolution simulation at the population level.

Prof. Dmitry Scherbakov
Limnological Institute SB RAS, Irkutsk

Modeling of genetic processes in populations facing different ecological challenges are widely used .in evolutionary studies. Still there are few advantages for using discrete models for the same purpose. Most of them are individual-oriented. This means that one may assign many different traits to objects. Therefore a simulation may result in any kind values.

For example, the objects may contain vectors mimicking nucleotide sequences, mutating according certain set of rules. The set of organisms resulting a simulation may be “sampled” and “sequenced” exactly the same way as it is done experimentally while studying real population. The set of aligned “sequences” may be processed with usual procedures depending on the exact problem. Important feature of this approach is that unlike the case of values which are next to impossible to measure in real world (like species of population numbers etc.), here we may tune simulation so that the results correspond directly to the experimental data.

Individual- oriented computer simulations may be used in order to study ecological interactions between several species or several populations of the same species distributed in space.

We give several examples of the use of individual-based models for elucidation of some features of molecular evolution in asexual organisms, host – parasite co-evolution and evolution under variable pressure of selection.

Distributed applications, web services, tools and grid infrastructures for bioinformatics

Dr. Luciano Milanesi
National Research Council - Institute of Biomedical Technology, Italy

Due to the increasing number of nucleotide and protein sequences produced by high throughput techniques, that have to be analyzed by bioinformatics tools, will be necessary to increase the actual calculation resources. Therefore, in order to face these new challenges successfully, it will be necessary to develop dedicated

supercomputers, parallel computer based on clustering technologies and high performance distributed platforms like GRID.

Next generation of GRID infrastructures, are trying to implement a distributed computing model where easy access to large geographical computing and data management resources will be provided to large multi/inter-disciplinary Virtual Organizations (VO) made of both research and user entities.

Indeed, computational and data Grids are "de facto" considered as the way to realize the concept of virtual places where scientists and researchers work together to solve complex problems in Bioinformatics, despite their geographic and organizational boundaries.

In these respects, then, Grid Computing is announcing another technological and societal revolution in high performance distributed computing as the World Wide Web has been since the last ten years for what concerns the meaning and the availability of global information. The aim is to operate this widely distributed computing environment as a uniform service, which looks after resource management, exploitation, and security independently of individual technology choices.

A general overview of the GRID technologies and computer cluster application to perform distributed bioinformatics applications for data mining, gene discovery, sequence similarity for searching of DNA and protein will be illustrated.

Platform "From Gene to Lead Compound": integration *in silico* and *in vitro* technologies

Prof. A.S. Ivanov

V.N. Orechovich Institute of Biomedical Chemistry RAMS, Moscow, Russia

Motivation and Aim. The pathway of drug discovery from idea to market consists of 7 basic steps: 1) disease selection, 2) target selection, 3) lead compound identification, 4) lead optimization, 5) preclinical trial evaluation, 6) clinical trials, 7) drug manufacturing. Two final stages are time- and money-consuming and their reduction is practically impossible owing to strict state standards and laws. Therefore, researchers paid special attention to increase the efficiency of drug development at earlier stages using computer modeling and bioinformatics integrated with new experimental methods. This methodology is directed at accelerating and optimizing the discovery of new biologically active compounds suitable as drug candidates (lead compounds). Recently these approaches have merged into a "from gene to lead compound" platform that covers the principle part of the pipeline. Several steps of this platform include computer modeling, virtual screening, and properties predictions`. Bioinformatics methods can reduce the amount of the compounds that are synthesized and tested by up to 2 orders of magnitude. Nonetheless, these approaches cannot completely replace the real experiments. The purpose of computer methods is to generate highly probable

hypotheses about new targets and/or ligands that must be tested later in real experiments.

Methods and Algorithms. The following methods and approaches are hilighted in lecture: 1) bioinformatics approaches in genome-based antiinfective targets selection [1]; 2) experimental technologies for target validation [2]; 3) solving of 3D structure of target - experimental and computer modeling technologies [3]; 4) strategy of computer-aided drug design [4]; 5) experimental testing of probable lead compounds.

Results. Some examples of passing execution of some bioinformatics steps of platform "from gene to lead compound" are presented: 1) targets selection in genome of *M. tuberculosis* and beyond [5]; 2) 3D modeling of cytochrome P450 1A2 and database mining for new leads using docking procedure [6]; 3) dimerization inhibitor of HIV protease: screening *in silico* and *in vitro* [7].

Conclusion. This lecture describes the integration of computer and experimental approaches in a complementary manner and some specific examples of the steps in implementing this platform.

Acknowledgments. This work was supported in part by Russian Foundation for Basic Research (grant 07-04-00575 and Russian Federal Space Agency in frame of ground preparation of space research).

REFERENCES.

1.A.V. DUBANOV, ET AL. (2001) VOPR. MED. KHIM. 47, 353-367. (IN RUSSIAN).
2. A.S. IVANOV, ET AL. (2005) BIOMED. CHEM. 51 (1), 2-18. (IN RUSSIAN).
3. A.S. IVANOV, ET AL. (2003) BIOMED. CHEM. 49 (3), 221-237. (IN RUSSIAN). 4. A.V. VESELOVSKY, A.S. IVANOV (2003) CURRENT DRUG TARGETS - INFECTIOUS DISORDERS, 3, 33-40.
5. A.S. IVANOV, ET AL. (2005) METHODS MOL. BIOL., 316: 389-432.
6. N.V. BELKINA, ET AL. (1998) VOPR. MED. KHIM. 44(5), 464-473. (IN RUSSIAN).
7. A.S. IVANOV ET AL. (2007) J. BIOINFORM. COMPUT. BIOL., 5(2B): 579-592.

Integration of reaction kinetics data: and modeling of metabolic networks: SABIO RK and SYCAMORE

Dr. Olga Krebs
EML Research gGmbH, Heidelberg, Germany

Systems biology involves analyzing and predicting the behavior of complex biological systems like cells or organisms. This requires qualitative information about the interplay of genes, proteins, chemical compounds, and biochemical reactions. It also calls for quantitative data describing the dynamics of these networks.

To provide quantitative experimental data for systems biology, we have developed SABIO-RK, a database system offering information about biochemical

reactions and their corresponding kinetics. It not only describes participants (enzymes, substrates, products, inhibitors, activators) and kinetic parameters of the reactions, but also provides both the environmental conditions for parameter determination and detailed information about the reaction mechanisms, including the mechanism type of a reaction and its related kinetic law equation defining the reaction rate with its corresponding parameters.

The SABIO-RK database is populated by merging information about biochemical reactions, mainly obtained from existing databases like KEGG (Kyoto Encyclopedia of Genes), with their corresponding kinetic data, manually extracted from literature. The kinetic data from articles are entered into the database using a web-based input interface, and subsequently curated, unified and systematically structured. The use of controlled vocabularies, synonymic notations and annotations to external resources offers the possibility of comparing and augmenting information about biochemical reactions and their kinetics.

SABIO-RK can be accessed in two different ways: via a web-based user interface to browse and search the data manually, and, more recently, via web-services that can be automatically called up by external tools, e.g. by other databases or simulation programs for biochemical network models. In both interfaces, reactions with kinetic data can be exported in SBML (Systems Biology Mark-Up Language), a data-exchange format widely used in systems biology.

SYCAMORE is a browser-based application that facilitates construction, simulation and analysis of kinetic models in systems biology. Thus, it allows e.g. database supported modelling, basic model checking and the estimation of unknown kinetic parameters based on protein structures. In addition, it offers some guidance in order to allow non-expert users to perform basic computational modelling tasks. SYCAMORE provides an interface to the SABIORK database to permit the user to locate and select the relevant kinetic data for these two reaction steps.

Availability:
SabioRK : http:/sabio.villa-bosch.de/SABIORK
SYCAMORE: http://sycamore.eml.org.

PIPSA: Comparison of protein interaction properties

Dr. Stefan Heinrich
EML Research gGmbH, Heidelberg, Germany

The simulation of metabolic networks in quantitative systems biology requires the assignment of enzymatic kinetic parameters. Experimentally determined values are often not available for the specific enzyme and therefore computational methods to estimate these parameters are needed. Sometimes corresponding parameters for the enzyme of interest have been measured for other species, isoforms or under different environmental conditions. These parameters can be used to estimate the required value, but it is not obvious which one is the best to choose.

Enzymes' catalytic and binding properties are dependent their on molecular interaction fields (MIFs). PIPSA, Protein Interaction Property Similarity Analysis (1,2), permits quantification of the similarity in the electrostatic potentials or other interaction properties of homologous proteins and has been applied to a variety of protein types. PIPSA is available as standalone software (http://projects.villa-bosch.de/mcmsoft/pipsa/3.0) but has recently been made available online in a version of webPIPSA (http://pipsa.eml.org) as well as in the SYCAMORE webserver (http://sycamore.eml.org). In the webserver, it is combined in a workflow with automated protein homology model building and electrostatic potential calculation. The results of the similarity analysis between each pair of proteins are given as a colour coded matrix as well as a dendrogram. Based on the similarities, the proteins can be clustered and the known parameters of the proteins with the most similar MIFs to the one of interest can be used to estimate the unknown kinetic parameters.

In this talk, the methodology of PIPSA and its extension to quantitative PIPSA (qPIPSA) (3) will be explained. After presenting applications of PIPSA in SYCAMORE as well as for target-selective drug design, some examples of webPIPSA (4,5) will be shown.

1. Blomberg, N., Gabdoulline, R.R., Nilges, M. and Wade, R.C. (1999) Classification of protein sequences by homology modeling and quantitative analysis of electrostatic similarity. *Proteins*, **37**, 379-387.
2. Wade, R.C., Gabdoulline, R.R. and De Rienzo, F. (2001) Protein interaction property similarity analysis. *International Journal of Quantum Chemistry*, **83**, 122-127.
3. Gabdoulline, R.R., Stein, M. and Wade, R.C. (2007) qPIPSA: relating enzymatic kinetic parameters and interaction fields. *BMC Bioinformatics*, **8**, 373.
4. Henrich, S., Richter, S. and Wade, R.C. (2008) On the use of PIPSA to Guide Target-Selective Drug Design. *ChemMedChem*, **3**, 413-417.
5. Richter, S., Wenzel, A., Stein, M., Gabdoulline, R.R. and Wade, R.C. (2008) webPIPSA: a web server for the comparison of protein interaction properties. *Nucl. Acids Res.*, gkn181.

Introduction to high-performance computing.

Prof. Thomas Ludwig
Ruprecht-Karls-Universitat, Heidelberg, Germany

The supercomputer parallel calculations find an ever increasing number of applications to solving numerous typical problems in modern science and technology. Their use is governed by emergence of a new class of superlarge problems. The course of lectures will include the history of development of supercomputing, review of the most powerful supercomputers, description of the classes of problems requiring parallel computations, and the main trends in development of technologies. Architectures of the supercomputers with shared and

distributed memories will be described as well as distinctions between supercomputers and parallel clusters. Specific features of data storage during parallel computations will be considered as well as the technologies of parallel programming.

Lecture 1: Architecture of high performance computers

To start with, we will discuss the architectural principles of high performance computers and, in particular, of compute clusters. We will have a closer look to processors, interconnect technology, storage, and in particular, to the memory architecture. The latter defines the classes of shared and distributed memory computers. The lecture will also present some data from the current TOP500 list of the strongest computers in the world. Finally, an overview over operating system aspects will be presented.

Lecture 2: Parallel programming principles

We will now learn how parallel programs are characterized and how in principle we design and implement such programs. A good knowledge of compiler and hardware details is often necessary in order to get optimal performance of the program. The parallelization paradigm of data partitioning and message passing will be introduced. Two measures will be presented to evaluate the performance of the parallel program.

Lecture 3: Message passing with MPI

The first step into parallel programming will be done based on the Message Passing Interface (MPI). We will write a small program that distributes data to different compute nodes, calculates some data, and finally collects the results. A few basic library calls for message passing will be introduced, which are already sufficient to write a first parallel program. Problematic issues like debugging and performance analysis will be covered.

Lecture 4: Advanced issues with message passing

MPI offers a huge number of library calls, most of which do just combine several basic calls and thus realize complicated activities in a single call. We will have a look at collective calls and sophisticated communication patterns. As bioinformatics is particularly data intensive, a first introduction to parallel input/output via MPI will be given. We will present an outlook onto advances features in the MPI-2 standard and what they are used for.

Models, Algorithms, and Parallel Computing for Large-Scale Phylogenetic Inference

Dr. Alexandros Stamatakis
Ludwig-Maximilians-University, Munich, Germany

The computation of ever larger as well as more accurate phylogenetic trees with the ultimate goal to compute the "tree of life" represents one of the grand challenges in high performance computing (HPC) Bioinformatics. Statistical methods of phylogeny reconstruction such as Maximum Likelihood (ML) and Bayesian inference have proved to be the most accurate models for evolutionary tree reconstruction and are becoming increasingly popular.

Unfortunately, the size of trees which can be computed in reasonable time is limited by the severe computational cost induced by these methods coupled with the explosive accumulation of sequence data, and the increasing popularity of large "gappy" multi-gene alignments.

There exist two orthogonal research directions to overcome this challenging computational burden which will be covered in this lecture:

Firstly, the development of faster and more accurate heuristic search algorithms as well as the implementation of efficient data-structures for multi-gene alignments.

Secondly, the application of high performance computing techniques to provide the required computational power, mainly in terms of CPU hours.

Initially, I will provide an introduction to phylogenetic inference under ML and outline the major computational challenges. Thereafter, I will discuss some of the basic search techniques as well as recent algorithmic advances in the field, especially with respect to rapid inference of support values. In the second part of my talk I will describe how the ML function can be adapted to a large variety of hardware architectures, ranging from multi-core processors to the IBM BlueGene supercomputer.

I will conclude with an overview of future challenges in the field.

The parallelization of bioinformatics problems: a tutorial

Yury VYATKIN,
Institute of Cytology and Genetics, Novosibirsk, Russia

In this tutorial we are going to follow the entire path from the serial program to its completely parallel version to learn how to use the features of modern high performance computing systems in full measure. This tutorial could be useful to everyone who knows C language a little bit and wants to learn how to solve bioinformatics problems with modern tools. We are going to cover the next topics:
What is High Performance Computing?

- Modern computers and supercomputers. Their types and features.
- What is parallelization and how to use it?
- Models of programming on supercomputers.

Problems that could be solved on HPC systems.

- Is my problem worth parallelization and how to determine that?
- The usage of profiler tool.
- Sample Plato program.

Parallelization with Message Passing Interface.

- The most frequently used places in programs to make parallelization.
- How to find a place in program to make parallelization?
- The way of parallelization.
- The most frequently used MPI operators.
- Let's insert some code to Plato program.

Further practice with Plato.

Practical course on the protein sequence analysis: searching of homologues and their hierarchical classification

Dr. Daniil Naumoff
GosNIIgenetika, Moscow, Russia

Using PSI-BLAST allows to reveal during several iterations a lot of potentially homologous proteins for almost any query. However, it is very difficult to analyze the obtained data manually. I will show how the analysis can be automated by a very simple algorithm. Using a new program – PSI Protein Classifier – allows to classify the obtained by PSI-BLAST proteins into known families. Proteins that remain unclassified can be grouped into new families. Also, this program can distinguish compact protein subgroups inside each family and compare distances between evolutionary related families. As a result, a preliminary hierarchical classification of the analyzed proteins can be proposed. Practical course will include various protein sequence comparisons.

PARTICIPANT'S ABSTRACTS

HIV-1 GP120 V3-LOOP COMPARATIVE STRUCTURE ANALYSIS: SEARCH FOR THE STRUCTURALLY CONSERVED REGIONS

I.V. Anishchenko

Belarusian State University, Minsk, Republic of Belarus
e-mail: anishchenko.ivan@gmail.com

Motivation and aim: The object of the current study is the third hypervariable V3 region of the HIV-1 gp120 protein, which is responsible for many aspects of viral infectivity. It is remarkable for its sequence diversity and, hence, structural diversity, which brings sufficient complications in the study of the V3-loop. At the same time information on the V3-loop conserved regions within its preferred conformations could be a significant tool for anti-AIDS drug design. On account of the V3-loop sequence diversity the conservative sequences of the HIV-1 group M subtypes are taken into consideration. On the basis of the V3-loop structures of definite isolates obtained before 3D modeling for each subtype sequence is performed. The study then is sighted at the search for structurally invariant regions in the models obtained, which may be regarded as drug targets.
Methods and Algorithms: 3D structure modeling is performed in MODELLER, also supporting functions of de novo modeling of loops in protein structures (http://www.salilab.org/modeller/). Several V3-loop structures obtained from the NRM and X-ray studies are taken as templates [1-2]. The consensus sequences of the V3-loop, being targets for modeling, are supported by the HIV Sequence Database at Los Alamos (http://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html). Structural analysis is held in Bio3d package for the R environment (http://mccammon.ucsd.edu/~bgrant/bio3d/). Both comparison in the geometric spaces of Cartesian coordinates and dihedral angles are performed.
Availability: In the result of the study we pay attention to the regions, which preserve their conformational states within the structures under review. The most probable ones are drawn out to be a starting point for further drug design.
References:
1. A.M. Andrianov, V.G. Veresov (2006) Determination of Structurally Conservative Amino Acids of the HIV-1 Protein gp120 V3 Loop as Promising Targets for Drug Design by Protein Engineering Approaches, *Biochemistry (Moskow)*, **71**: 906-914
2. Huang C.-C., M. Tang, M. Y. Zhang, S. Majeed, E. Montabana, R. L. Stanfield, D. S. Dimitrov, B. Korber, J. Sodroski, I. A. Wilson, R. Wyatt, and P. D. Kwong (2005) Structure of a V3-containing HIV-1 gp120 core, *Science*, **310**: 1025–1028

PHYLOGENETIC CLASSIFICATION OF THE HECT-DOMAIN UBIQUITIN-PROTEIN LIGASE FAMILY

E.V.Chaplygina*, A.S.Zhabereva, M.R.Gainullin

Institute of Applied and Fundamental Medicine, Nizhny Novgorod State Medical Academy, Nizhny Novgorod, Russia
e-mail: anastasia@gma.nnov.ru
*Corresponding author

Motivation and Aim: Ubiquitylation proceeds as a hierarchical cascade of reactions catalyzed by enzymes of three different types: E1 (an ubiquitin-activating enzyme), E2 (ubiquitin conjugating enzyme) and an ubiquitin-protein ligase E3, that directly binds ubiquitin to the lysine residue of substrate protein. E3 proteins serve as major substrate recognition component during ubiquitylation reaction. There are three classes of ubiquitin protein ligases: RING-finger, U-box, and HECT-domain containing E3s. Members of HECT E3 ubiquitin-protein ligases family are big proteins and range from approximately 80 kDa to more than 500 kDa. They are characterized by the presence of a HECT (homologous to E6-AP C-terminus) domain at a C-terminal region. The major goal of the present work was to build a new classification of HECT-domain E3 ligase family. This classification was elaborated based on phylogenetic analysis and domain architecture of 138 HECT-domain proteins from seven eukaryotic organisms.
Methods and Algorithms: A search for homologous sequences was performed with the PSI-BLAST, pairwise alignment was carried out manually using BioEdit program, multiple alignment was processed by ClustalW 1.83. The phylogenetic tree was generated using the neighbor-joining algorithm of ClustalW (distance method), with correction for multiple substitutions, and 2000 bootstrap calculations. Graphic representations of trees were obtained with the TreeView Win32 program. The analysis of proteins domain organization was executed by NCBI Conserved Domain Database.
Results: Basing on the phylogenetic analysis, all examined members of the HECT E3 family were divided into 9 subfamilies, integrated into 3 phylogenetic groups. The common evolutionary origin of subfamilies relating to respective group has been proposed. The first phylogenetic group includes three subfamilies: RSP5, TOM1, K0317. The K0317 subfamily represents the earliest branch of this group. At late stages of evolution there was a division in two branches - RSP5 and TOM1. Each of them are characterized by complicated multi-domain structure. For members of TOM1 subfamily the presence of two DUF domains is obligatory. Proteins of RSP5 subfamily contain C2 and several WW domains and in K0317 subfamily Filamin domain is revealed. The second phylogenetic group includes HUL5, HUL4, S-HERC subfamilies. The majority of this HECT-domain protein group members - HUL5 and HUL4 subfamilies - are characterized by mono-domain structure. Representatives of S-HERC subfamily have multi-domain organization containing RCC1 and ATS1 domains. In our opinion it occur because

this subfamily has appeared during evolution greatly later than HUL5 and HUL4 subfamilies. The third phylogenetic group includes UFD4, L-HERC, K1333 subfamilies. The UFD4 subfamily is an early appeared group of proteins. At more late stages of evolution L-HERC and K1333 subfamilies were formed. The multi-domain architecture of those subfamilies is different. In particular, members of UFD4 subfamily contain PolyA, ZnF and ARM domains. The APC10 domain is revealed in L-HERC subfamily, containing representatives of K1333 subfamily PHD domains.

THE NEW ALGORYTHM FOR PHYLOGENETIC RECONSTRUCTION OF NON-RECOMBINING DNA SEQUENCES

N.P. Eltsov
Laboratory of Human Molecular Genetics, Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: eltsovnp@bionet.nsc.ru

Motivation and Goals: The mitochondrial genome has been the most widely used system for the investigation of the evolutionary history of our species. It has become a system of choice because of its high rate of sequence divergence and because of its uniparental, maternal inheritance. With the advent of human population genomics [1] and rapid accumulation of complete mtDNA sequences, it has become increasingly important to quickly and comprehensively analyze the data available.
Methods and Algorithms: We propose a novel maximum parsimony-based algorithm for reconstruction of phylogeny of non-recombining DNA sequences. This algorithm includes three consecutive steps. 1) sorting and identification of recurrent mutations (the ones that do not allow for unambiguous phylogeny reconstruction); 2) analysis of recurrent mutations and identification of the most plausible parallel mutations; 3) parallelization of these mutations. When aligning the DNA sequences, we also applied a novel optimized algorithm for weighted alignment.
Results: The algorithm designed was used in mtPhyl. This software package allows analyzing rapidly human entire mtDNA sequences. The mtPhyl identifies the mutated region, aminoacid replacements, and calculates the coalescence time for the most recent common ancestor. In addition, it sorts out sequences in accord with parameters as outlined by user, and enables us to estimate the natural selection. The output can be easily converted into any formats used by popular programs such as Arleqin, DnaSP, etc. The mtPhyl appears to be a unique package which can be used as a standalone tool and as an accessory program for preliminary data analysis.
Conclusion: mtPhyl based on a new algorithm to reconstruct molecular phylogeny represents a timely advance, since the advent of cheaper sequencing methods has

generated an excess of sequence data, and there is an urgent need to perform their automatic analysis.
Availability: Demo version of mtPhyl is available from the authors upon request and at http://www.bionet.nsc.ru/labs/mtgenome/programs.html.
References:

1. S.B. Hedges (2000) Human evolution. A start for population genomics, *Nature*, **408**: 652-653.

INDIVIDUAL-BASED MODELLING OF ADAPTIVE SPECIATION IN SPATIALLY STRUCTURED POPULATIONS

***V. Fazalova*[1,2] [*], *U. Dieckmann*[2]**
[*]Corresponding author

[1]Limnological Institute, Siberian Branch of the Russian Academy of Sciences, Irkutsk 664003, Russia
[2]International Institute for Applied Systems Analysis, Laxenburg 2361, Austria

Motivation and Aim
Traditional theoretical studies of sympatric speciation were based on mean-field models. Such models assume that all individuals in a population experience the same environment. More recently, spatially explicit models of speciation have been analyzed, in an effort to better account for properties characteristic of real populations. The purpose of this study was to investigate how spatial patterns resulting from the self-structuring of population could influence the course and pace of phenotypic adaptive evolution. We focused on asexual reproduction to investigate how ecological interactions among individuals can result in phenotypic diversification through frequency-dependent disruptive natural selection.

Methods and Algorithms
We investigated the phenotypic evolution of asexual organisms inhabiting spatially continuous and intrinsically homogeneous two-dimensional environments, using an individual-based modeling approach [1]. We used individual-based simulations for our study since these are able to incorporate realistic spatial population structure and the stochasticity inherent to biological interactions. In our model, individuals compete, mate, reproduce, and disperse locally. New phenotypes are occasionally created through mutations of small effect. Since the biological species concept does not apply to asexual organisms, we based our investigation of speciation dynamics on the detection of distinct phenotypic clusters arising in the evolving population. Algorithmically, the model was implemented using the minimal process method.

Results
By varying the dispersal rates and/or distances of individuals, we obtained different spatial patterns of phenotypes, as well as different speciation dynamics. Two extremes need to be distinguished: at one end of the spectrum, there are

highly structured populations in which individuals form distinct spatial and phenotypic clusters, while at the other end, there are well-mixed populations in which individuals and phenotypes are distributed randomly in space. We found that, under a wide range of parameter conditions, adaptive speciation occurs faster in well-mixed populations than in spatially structured populations. We could show how this finding can be explained by two effects arising in spatially structured populations. First, in different spatial clusters the fitness minima induced by frequency-dependent disruptive selection occur at different phenotypes. This results in a shallower global fitness landscape and, therefore, in weaker disruptive selection pressures. The second effect arises from source-sink dynamics among spatial clusters, with some clusters expanding while others are shrinking. This causes wider variation around mean fitness values and thus blurs the global fitness landscape, which in turn reduces the average survival probability of mutants and thereby slows down the population's response to selection.

References

1. Doebeli, M. and Dieckmann, U. (2003). Speciation along environmental gradients. *Nature*, **421**: 259-264.

PREDICTION OF PROTEIN INTERACTIONS USING HOMOLOGOUS INTERFACES

T.V. Kirys[1,2], A.V. Tuzikov[1], D.K. Voytekhovsky[1,3], Y.E. Grushetsky*[1,3]
[1]United Institute of Informatics Problems BAS, Minsk, Belarus
[2]Belarusian State University, Minsk, Belarus
[3]Moscow Institute of Physics and Technology, Moscow, Russian Federation
e-mail: {kirys, tuzikov, grushetsky}@newman.bas-net.by
*Corresponding author

Motivation and Aim:

The importance of protein in all living systems is immense. At the protein level most biological mechanisms are based on shape-complementarity, so that proteins present particular concavities and convexities that allow them to bind to each other and form complex structures. In general, proteins perform their functions by forming complexes. Therefore the knowledge of interactions between proteins is essential for understanding the molecular mechanisms of biological systems and drug design. The aim of our research is the prediction of protein interactions using homologous interfaces. We address two problems. First, given an interface database and a protein database the task is to predict possible protein interactions. Second, given a target protein, an interface database and a protein database the task is to find possible interacting partners.

Methods and Algorithms:

The assumption behind the homology-based approaches is that interaction information can be extrapolated from one complex structure to homologs of the interacting proteins. Define an interface as a pair of interacting binding sites (patch) of two protein. Each patch consists of all residues of a protein that are located within a distance of 6 $\mathring{A}$ from the other protein. From such an interface definition it is clear that in a patch there are continuous segments of protein polypeptide chain. We use that observation for searching similar patches in proteins. We also assume that if a protein spatial feature is similar to a patch spatial feature then the corresponding distance matrices are highly correlated.
The proposed algorithm consists of the following steps:

1. select continuous segments in the interface;
2. find similar continuous segments in proteins using dynamic programming on distance matrix;
3. superpose proteins by the correspondence found;
4. check modeled interaction (steric clash, similarity score);

Taking into account residue types makes prediction more accurate.
Results:
This algorithm was implemented in C++. Our algorithm is quite fast, depending on the protein and interface sizes it takes several seconds to find the solution, what is important in screening task.
Conclusion:
We have proposed a novel homology-based algorithm for prediction of protein interactions, which employs dynamic programming on distance matrix. Finding particularly spatially similar continuous segments reduces searching time drastically in comparison with hashing algorithms and adequately reflects the nature of protein interactions.
Availability:
The software is available on request from the authors.

INCORPORATING DIFFERENT TYPES OF EXPERIMENTAL DATA ON DNA-PROTEIN BINDING INTO THE SINGLE *IN SILICO* MODEL

I.V. Kulakovskiy[1,2] ****, A.V. Favorov***[2,3]***, V.J. Makeev***[2,1]
[1]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia, [2]Institute of Genetics and Selection of Industrial Microorganisms, FGUP GosNIIgenetika, Moscow, Russia, [3]The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore, MD, USA
e-mail: ikulakovsky@inbox.ru
[*]Corresponding author

Motivation and Aim: Genome wide location of transcription factor binding sites (TFBS) at ChIP-chip tiled arrays and even footprints can bring about rather

extended DNA segments, which makes a challenge of binding motif identification with traditional techniques. On the other hand the data on protein binding to DNA is available from many different sources of experimental information. Simultaneous analysis of data obtained from such sources as SELEX, ChIP-chip, footprints etc can result in a much clearer signal for DNA-protein binding than usage any of the data sources alone. For instance, the oligos yielded by SELEX strictly correspond to the binding protein, but they are usually short and in practice the binding motif is often distorted. At the same time ChIP-chip arrays give functional binding motifs, often *in vivo*, but the resulting sequences are long and can contain binding signals for proteins different from the test one. Our objective was to make an integrated tool for incorporating different types of experimental data into the single protein binding model.

Methods and Algorithms: For a binding model we have selected the Positional Weight Matrix (PWM) which is traditional motif model for transcription factor binding sites (TFBS) at DNA. The core of the algorithm is SeSiMCMC Gibbs sampler which is used to construct the anchored optimal multiple local alignment (MLA) of raw sequence data. "The anchored" means that any sequence included into MLA should overlap with the anchor sequence initially seeded into the data. This layout allowed incorporating the data of ChIP-chip and SELEX simultaneously. SELEX data was used to place anchors in ChIP-chip sequences. The resulting MLA corresponds to the binding signal for the correct protein.

Results: We paid a particular attention to identify the length of a binding signal, the problem, which is not solved in many signal identification tools. We have tested our system for several TFBS of Human and Drosophila fly and resulting motif models have better selectivity than those built using one source of experimental data.

Conclusion: We created a tool designed to construct a binding motif model from different types of experimental data on DNA-protein binding. Now we can map specific site occurrences at genome sequences within mapped ChIP-chip resulting regions. We can detect genome wide putative TFBS rich regions, which were not covered by ChIP-chip results. This opens a view to compare ChIP-chip results obtained in different experimental environment and study tissue-specific gene expression.

Availability: The source code is available by request. Web-based version of the software tool is planned for release.

UBIQUITOMIX DATABASE: A NEW RESOURCE ON UBIQUITIN SYSTEM

V.A.Kovalyov[3]*, M.R.Gainullin[1,3], E.V.Eremin[2], A.Garcia[1]
[1]Nizhny Novgorod State Medical Academy, [2]Institute of Applied Physics RAS,
[3]Nizhny Novgorod State University, Nizhny Novgorod, Russia
e-mail: vladlen-85@list.ru
[*]Corresponding author

Motivation and Aim: Ubiquitylation is a process of great importance for many vital cell functions (proteolysis, signal transduction, control of gene expression, DNA repair, etc). Ubiquitin system consists of complex of enzymes wich catalyse covalent attachment of ubiquitin to a target protein (E1 - ubiquitin activating enzymes, E2 – ubiquitin conjugating enzymes, E3 – ubiquitin protein ligases), particular proteins serving as substrates for ubiquitin modification, deubiquitylation enzymes (DUB) and proteins containing ubiquitin binding domains (UBP). Biological significance of ubiquitin dependent regulation makes it attractive object of research, using methods of systems biology. However among specialized biological Internet-resources (SwissProt/TrEMBL, GO, KEGG, Reactome, etc.) there are no resources correctly and completely describing ubiquitin system. Therefore Internet-resource is required for accumulation and ordering of knowledge on ubiquitin system. The aim of our work was to create specialized object-oriented database, collecting data on all members of ubiquitylation system and their interactions in various organisms.
Methods and Algorithms: For development of Ubiquitomix database BioUml platform has been chosen. It is open source software, being used for formalization of biological systems, as well as for their visualization (http://www.biouml.org).
Results: Main principles of formalization and graphic display of all compounds of ubiquitin system and their interactions have been developed. Proteins were formalized according to their properties. Basing on protein structure we had divided them in two different kinds of entities: simple entities (monomeric or homo-oligomeric proteins) and modular entities (hetero-oligomeric proteins and protein complexes). According to their function, all components of ubiquitin system have been divided into: 1) enzymes of conjugation (i.e. attachment of ubiquitin molecule to target protein); 2) deubiquitylating (the reversal of conjugation) enzymes; 3) proteins which recognize a specific ubiquitin signal; 4) ubiquitin target proteins. For modular entities, components have been divided into 5 types: possessing catalytic activity, carrying out ubiquitin binding function, substrate binding proteins, adaptor proteins and effector proteins. Function and structure of proteins are displayed by form and color of pictogram. Each protein is characterized by name and synonyms, gene name, MW, length, posttranslational modifications, function in ubiquitin system. Additionally, ubiquitylated proteins are characterized by type of ubiquitin chain, site of ubiquitylation and

ubiquitylating machinery. If ubiquitylation mechanism of target protein is described completely, respective reaction is displayed on own pathway diagram. For partially characterized ubiquitin binding processes semantic diagrams are chosen. Information on 383 human and 508 yeast proteins has been collected in Ubiquitomix database and their interactions have been analyzed.
Conclusions: General information on ubiquitin system members and their interactions has been collected in developed database. We expect that further development of Ubiquitomix database will be of common interest for different research groups involved in studies of ubiquitin system.

VALIDATION OF VIRAL EBNA1 PROTEIN AS THE PERSPECTIVE TARGET FOR DRUG DISCOVERY

***N.D. Prakhov*[2], M.R. Gainullin*[1,2]**
[1]Nizhny Novgorod State Medical Academy, [2]Nizhny Novgorod State University, Nizhny Novgorod, Russia
e-mail: n.prakhov@yahoo.com
*Corresponding author

Motivation and Aim:
The ubiquitylation is the crucial signal mechanism in eukaryotic cells. The ubiquitin system comprises a set of enzymes, which recognize several cellular target proteins and attach to them highly conservative protein ubiquitin as well as catalyzing reverse reaction - deubiquitylation. Mdm2 is an ubiquitin-protein ligase, which specifically ubiquitylates p53 protein. USP7 is a deubiquitylating enzyme, which is capable to interact with both p53 and Mdm2.
It was shown, that USP7 is one of the basic regulators, capable to raise p53 level in the cell in such a way that p53 induces an apoptosis [1]. Ubiquitylation may be utilized by viruses to enhance pathogenesis. In particular, Epstein-Barr nuclear antigen 1 (EBNA1) protein of Epstein-Barr virus interacts with USP7 and blocks its activity [2]. The purpose of the present work is to evaluate one of indicated proteins (i.e. p53, USP7, Mdm2 and EBNA1) as a promising potential target for further drug discovery.
Methods and Algorithms:
The computer analysis of proteins was done in the free program ViewerPro.
Results:
It was revealed that USP7 possesses single binding site, interacting with all partner proteins (EBNA1, Mdm2 and p53). Respectively, EBNA1, Mdm2 and p53 have similar areas, interacting with USP7. A binding affinity towards USP7 decreases in row EBNA1 - Mdm2 - p53 [3]. Thus, targeting EBNA1 with small ligands can be rated as preferable strategy of further investigation. However, some difficulties should be overcome, in particular, possible cross-reaction of compound with another physiological binding partners. To solve this problem we have superimposed 3D-structures of EBNA1, Mdm2 и p53. We have shown that

binding interfaces of three analyzed protein-protein complexes i.e. USP7 protein with EBNA1, MDM2 and p53 respectively are differ significantly. We have found a polypeptide fragment of EBNA1 protein, comprising side chain groups which were able to form H-bonds, while Mdm2 and p53 did not possess such groups. We propose that those differences are enough to reveal a number of chemical groups on surface of EBNA1, that may specifically interact with small ligand, without cross-reaction with other two proteins. It is planned to lead a virtual screening of small ligands to zone of 442-444 amino acids of the protein EBNA1.

References:

1. M.Li et al. (2002) Deubiquitination of p53 by HAUSP is an important pathway for p53 stabilization, *Nature*, **416**: 648-653.
2. M.N.Holowaty et al. (2003) Protein interaction domains of the ubiquitin-specific protease, USP7/HAUSP, *J Biol Chem*, **278**: 47753-47761.
3. M.Hu et al. (2006) Structural Basis of Competitive Recognition of p53 and MDM2 by HAUSP/USP7: Implications for the Regulation of the p53-MDM2 Pathway, *PLoS Biol*, **4:** e27.