The International School of Young Scientist "Evolution, Systems Biology and High Performance Computing Bioinformatics"

Scientific Program

Novosibirsk,
July 12-15, 2006

Organizers and sponsors

Organizers

- Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences
- Siberian Branch of the Russian Academy of Sciences
- Ruprecht-Karls-Universitat Heidelberg, Institut fur Informatik, Heidelberg, Germany
- Institute of Computer Science, Foundation for Research and Technology-Hellas, Greece
- Novosibirsk State University

Organizational support

- Council of Young Scientists of the Institute of Cytology and Genetics
- The Chair of Informational Biology of the Novosibirsk State University
- Laboratory of theoretical genetics, IC&G

Financial support

This Summer School has been held with the financial assistance of INTAS* and Russian Foundation for Basic Research (RFBR).

*The views expressed at the event, however, can in no way be taken to reflect the official opinion of INTAS.

SCHOOL TIMETABLE

TUESDAY, 11 JULY

15.00 – 18.00 Registration

WEDNESDAY, 12 JULY

9.00 – 9.30 Registration

9.30 Opening

9.30 – 11.05 Lecture 1

Architecture of High Performance Computers

Prof. Thomas Ludwig, Ruprecht-Karls-Universitat Heidelberg, Institut fur Informatik, Heidelberg, Germany

11.25 - 13.00 Lecture 2

Parallel Programming Principles

Prof. Thomas Ludwig, Ruprecht-Karls-Universitat Heidelberg, Institut fur Informatik, Heidelberg, Germany

Lunch

14.30 – 15.30 Seminar 1

Phylogenetic Analysis of a Protein Family

Dr. Daniil Naumoff, State Institute for Genetics and Selection of Industrial Microorganisms, Moscow, Russia

16.30- 21.30 Ob River Boat Trip

THURSDAY, 13 JULY

9.30 – 11.05 Lecture 3

Message Passing with MPI

Prof. Thomas Ludwig, Ruprecht-Karls-Universitat Heidelberg, Institut fur Informatik, Heidelberg, Germany

11.25 - 13.00 Lecture 4

Advanced Issues with Message Passing

Prof. Thomas Ludwig, Ruprecht-Karls-Universitat Heidelberg, Institut fur Informatik, Heidelberg, Germany

Lunch

14.30 – 16.05 Lecture 5

Computation of Large Phylogenetic Trees: Algorithmic and Technical Solutions

Dr. Alexandros Stamatakis, Swiss Federal Institute of Technology, Lausanne, Switzerland

16.30 – 17.30 Seminar 2

The Parallelization of Bioinformatics Problems: A Tutorial

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

18.00 - 20.00 Football Match Between the School Participants and Young Scientists from IC&G

FRIDAY, 14 JULY

9.30 – 11.05 Lecture 6

Inhibitors of Protein-Protein Interactions as Lead Compounds for New Drugs Generation

Prof. Alexis Ivanov, V.N. Orekhovich Institute of Biomedical Chemistry, Russian Academy of Medical Sciences, Moscow, Russia

11.25 - 13.00 Lecture 7

Transcription and Translation Regulations of Amino Acid Metabolism Genes in Actinobacteria and Intron-Containing Genes in Chloroplasts of Algae and Plants

Prof. Vassily Lyubetsky, Institute for Information Transmission Problems, Russian Academy Of Sciences, Moscow, Russia

Lunch

14.30 – 16.05 Lecture 8

Gene Expression Patterns: Methods for Visualization, Processing, and Quantification

Dr. Konstantin Kozlov, St. Petersburg State Polytechnic University, St. Petersburg, Russia

16.30 – 17.30 Seminar 3

Transcription and Translation Regulations of Amino Acid Metabolism Genes in Actinobacteria and Intron-Containing Genes in Chloroplasts of Algae and Plants

Dr. Alexander Seliverstov, Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia

18.00 – 19.00 Seminar 4

GenomeBrowser – Software for Visual DNA Analysis

Volod'ko V.B.[1], Shuvaev R.Yu.[1], Ulyashin A.V.[1], Oshchepkov D.Yu[2], [1]Novosibirsk Center of Information Technologies "Unipro", Novosibirsk, Russia; [2]Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

SATURDAY, 15 JULY

9.30 – 13.00 Session of the Young Scientist's Presentations

Lunch

14.30 – 16.05 Lecture 9

Distributed Applications, Web Services, Tools and GRID Infrastructures for Bioinformatics

Prof. Luciano Milanesi, National Research Council - Institute Of Biomedical Technology, Italy

16.30 – 17.30 Seminar 5

The Models Of Adaptive Dynamics As Tools For Studying Of Neutral Molecular Evolution

Dr. Yury Bukin, Limnological Institute SB RAS, Irkutsk, Russia

18-00 Closing Ceremony

LECTURES AND SEMINARS

LECTURE 1. ARCHITECTURE OF HIGH PERFORMANCE COMPUTERS

Prof. Thomas LUDWIG

Ruprecht-Karls-Universitat Heidelberg, Institut fur Informatik, Heidelberg, Germany

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

Prof. Thomas LUDWIG

Ruprecht-Karls-Universitat Heidelberg, Institut fur Informatik, Heidelberg, Germany

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.

SEMINAR 1. PHYLOGENETIC ANALYSIS OF A PROTEIN FAMILY

Dr. Daniil NAUMOFF

State Institute for Genetics and Selection of Industrial Microorganisms, Moscow, Russia.

I am going to present a complete procedure of a protein family analysis, viz. from database searching to visualization of the phylogenetic tree. A special attention will be paid for solving problems of multi-domain protein structure and for clarifying the phylogenetic status of 'atypical' members of a protein family. Results of the phylogenetic analysis of several glycosidase families will be shown as examples. The methods and programs suggested can be applied to any protein family but they would work more effectively with globular solving proteins. Protein family analysis can be started using any protein sequence as a query. It does not matter if the protein has been studied enzymatically or corresponds to a biochemically uncharacterized ORF. A preliminary version of the lecture (in Russian) has been published on-line in Zbio journal (http://zbio.net/bio/001/003.html).

LECTURE 3. MESSAGE PASSING WITH MPI

Prof. Thomas LUDWIG

Ruprecht-Karls-Universitat Heidelberg, Institut fur Informatik, Heidelberg, Germany

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

Prof. Thomas LUDWIG

Ruprecht-Karls-Universitat Heidelberg, Institut fur Informatik, Heidelberg, Germany

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.

LECTURE 5. COMPUTATION OF LARGE PHYLOGENETIC TREES: ALGORITHMIC AND TECHNICAL SOLUTIONS

Dr. Alexandros STAMATAKIS

Swiss Federal Institute of Technology, Lausanne, Switzerland

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 phylogenetic analysis such as maximum likelihood and Bayesian inference have proved to be the most accurate models for evolutionary tree reconstruction.

Unfortunately, the size of trees which can be computed in reasonable time is limited by the severe computational cost induced by these methods. There exist two orthogonal research directions to overcome this challenging computational burden: Firstly, the development of novel, faster, and more accurate heuristic algorithms. Secondly, the application of high performance computing techniques, the deployment of supercomputers, and Grid-computing to provide the required computational power, mainly in terms of CPU hours.

The field has witnessed significant algorithmic advances over the last 2-3 years which allow for inference of large phylogenetic trees containing 500-1000 sequences on a single PC processor within a couple of hours using maximum likelihood. On the other hand, the main problem which high performance computing implementations of maximum likelihood analyses faces is that technical development lags behind algorithmic development, i.e., programs are parallelized that do not represent the state-of-the-art algorithms any more.

Within this context, the talk initially aims to provide a brief overview of the computational challenges large-scale phylogenetic inference face concerning both algorithmic as well as supercomputing aspects. The benefits of simultaneous algorithmic and technical development are outlined by example of the program RAxML (Randomized Axelerated Maximum Likelihood). The sequential version of RAxML has been used to compute the largest maximum likelihood tree to date (comprising 25.000 organisms) on a single CPU.

In addition, recent algorithmic developments including novel genetic search algorithms and search techniques will be discussed. Finally, an overview over possible future HPC implementations of those novel algorithms is provided including Grid-based solutions, implementations for hybrid supercomputer architectures, and exploitation of vector-like peripheral processors like for example Graphics Processing Units (GPUs).

SEMINAR 2. 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.

LECTURE 6. INHIBITORS OF PROTEIN-PROTEIN INTERACTIONS AS LEAD COMPOUNDS FOR NEW DRUGS GENERATION

Prof. Alexis IVANOV

V.N. Orekhovich Institute of Biomedical Chemistry, Russian Academy of Medical Sciences, Moscow, Russia

Protein-protein interactions represent a new and extremely attractive class of molecular targets for creation of essentially new drugs generation. The reason is that contact areas of protein molecules in complexes are very conservative regarding mutational changes and, hence, the probability of mutational drug resistance is low for drugs targeted to these areas.

Laboratory of authors works in the area of computer-aided design and experimental testing of inhibitors of protein-protein interactions. The computer technologies include methods of 3D molecular modeling, methods of molecular mechanics, molecular dynamics simulation, molecular docking, analysis of intermolecular interactions, virtual alanine screening, molecular database mining, de novo design, etc. The basic experimental approach is technology of intermolecular interactions analysis in vitro using optical biosensor Biacore-3000 utilizing the effect of surface plasmon resonance. Particular examples of approaches and results will be presented based on the study of tetramer of bacterial L-asparaginase and inhibitors of HIV-1 protease dimerization.

LECTURE 7. TRANSCRIPTION AND TRANSLATION REGULATIONS OF AMINO ACID METABOLISM GENES IN ACTINOBACTERIA AND INTRON-CONTAINING GENES IN CHLOROPLASTS OF ALGAE AND PLANTS

Prof. Vassily LYUBETSKY

Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia

Formation of alternative structures in mRNA in response to external stimuli, either direct or mediated by proteins or other RNAs, is a major mechanism of regulation of gene expression in bacteria. This mechanism has been studied in detail using experimental and computational approaches in proteobacteria and Firmicutes, but not in other groups of bacteria. Comparative analysis of amino acid biosynthesis operons in Actinobacteria resulted in identification of conserved regions upstream of several operons. Classical attenuators were predicted upstream of trp operons in *Corynebacterium* spp. and *Streptomyces* spp., and trpS and leuS genes in some *Streptomyces* spp. Candidate leader peptides with terminators were observed upstream of ilvB genes in *Corynebacterium* spp., *Mycobacterium* spp. and *Streptomyces* spp. Candidate leader peptides without obvious terminators were found upstream of cys operons in *Mycobacterium* spp. and several other species. A conserved pseudoknot (named LEU element) was identified upstream of leuA operons in most *Actinobacteria*. Finally, T-boxes likely involved in the regulation of translation initiation were observed upstream of ileS genes from several Actinobacteria. The metabolism of tryptophan, cysteine and leucine in Actinobacteria seems to be regulated on the RNA level. In some cases the mechanism is classical attenuation, but in many cases some components of attenuators are missing. The most interesting case seems to be the leuA operon preceded by the LEU element that may fold into a conserved pseudoknot or an alternative structure. A LEU element has been observed in a transposase gene from *Bifidobacterium longum*, but it is not conserved in genes encoding closely related transposases despite a very high level of protein similarity. One possibility is that the regulatory region of the leuA has been co-opted from some element involved in transposition. Analysis of phylogenetic patterns allowed for identification of ML1624 of M. leprae and its orthologs as the candidate regulatory proteins that may bind to the LEU element. T-boxes upstream of the ileS genes are unusual, as their regulatory mechanism seems to be inhibition of translation initiation via a hairpin sequestering the Shine-Dalgarno box.

A short description of the originally developed algorithms of searching for conservative protein-RNA binding sites will be provided. One of these algorithms is applied to analyze chloroplast genes. Candidate protein-RNA binding sites were detected upstream of atpF, petB, clpP, psaA, psbA and psbB genes in many chloroplasts of algae and plants. We surmise that some of these sites are involved in suppressing translation until splicing is completed.

The lecture includes results of the two original publications and describes several novel algorithms in bioinformatics.

LECTURE 8. GENE EXPRESSION PATTERNS: METHODS FOR VISUALIZATION, PROCESSING, AND QUANTIFICATION

Dr. Konstantin KOZLOV

St. Petersburg State Polytechnic University, St. Petersburg, Russia

High-quality and high-resolution images of gene expression patterns become available for developmental biology due to confocal scanning microscopy technique. Extraction of quantitative information is important to get insights into underlying regulation, construct mathematical models, and plan new experiments. We introduce a new image processing software package ProStak integrated into distributed computing environment. ProStak includes all operations needed to extract quantitative information from 2D and 3D biological images. The chain of processing steps can be visually constructed using graphical user interface that provides convenient environment for digital image processing for all groups of scientists: beginners, non-programmers, and experts, for which the speed of the result acquisition is critical. All processing methods can be accessed by a user through the command line interface, as well as through shared and static libraries. The combination of features mentioned above distinguishes ProStak from other image processing packages such as commercial systems Matlab and VisiQuest, and freely available SIVIL, SCIRun, and TiViPe.

SEMINAR 3. TRANSCRIPTION AND TRANSLATION REGULATIONS OF AMINO ACID METABOLISM GENES IN ACTINOBACTERIA AND INTRON-CONTAINING GENES IN CHLOROPLASTS OF ALGAE AND PLANTS (ACCOMPANYING THE LECTURE 7)

Dr. Alexander SELIVERSTOV

Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia

SEMINAR 4. GENOMEBROWSER – SOFTWARE FOR VISUAL DNA ANALYSIS

VOLOD'KO V.B.[1], SHUVAEV R.YU.[1], ULYASHIN A.V.[1], OSHCHEPKOV D.YU[2]

[1]Novosibirsk Center of Information Technologies "UniPro",
Novosibirsk, Russia; [2]Institute of Cytology and Genetics SB RAS,
Novosibirsk, Russia

GenomeBrowser is a software tool aimed at interactive visualization, analysis, and annotation of DNA sequences, which allows user to navigate, zoom and create annotations for selected parts of sequences. The options of advanced search, set of contextual statistics, support of major DNA sequence data formats and pluggable modules interface increase efficiency, simplify and, in a number of cases, make the process of analysis automatic. Final results as well as intermediate ones are graphically displayed and can be saved and used in a future analysis. UniPro has been developing this instrument in close collaboration with scientists of Institute of Cytology and Genetics SB RAS.

Main features of GenomeBrowser as well as additional plugins capabilities will be presented in interactive studying process.

LECTURE 9. DISTRIBUTED APPLICATIONS, WEB SERVICES, TOOLS AND GRID INFRASTRUCTURES FOR BIOINFORMATICS

Prof. 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.

SEMINAR 5. THE MODELS OF ADAPTIVE DYNAMICS AS TOOLS FOR STUDYING OF NEUTRAL MOLECULAR EVOLUTION

Dr. Yury BUKIN

Limnological Institute SB RAS, Irkutsk, Russia

PARTICIPANT'S ABSTRACTS

INVESTIGATION OF ALTERNATIVE SPLICING FEATURES, DNA HETEROGENEITY INFLUENCE ON EXPRESSION RATE FOR GENES AND PROTEINS RESPONSIBLE FOR CORONARY ARTERY DISEASE AND HYPERTROPHIC CARDIOMYOPATHY

Galina BOLDINA

The Kazakh National University named after Al-Farabi, Almaty, Kazakhstan
e-mail: gboldina@mail.ru

Cardiovascular disease (CVD) is the leading cause of death in Europe. Despite major advances in treatment, coronary artery disease (CAD) remains the biggest cause of mortality and morbidity in the developed world. Hypertrophic cardiomyopathy (HCM) is an inherited disease of the heart muscle, characterized by unexplained left ventricular hypertrophy. HCM is also one of the major causes of sudden cardiac death, sometimes occurring in young asymptomatic people. Based on this we've chosen genes and proteins responsible for CAD and HCM development for disclosure of molecular mechanisms development. Genetic labs provide researches for determination of genes and proteins susceptible to the modifications causing CVD. However, the molecular mechanisms of the development of these diseases are still largely unknown. The disclosure of this mechanism may help to determine targets for gene therapy.

The novelty of this research is the comprehensive approach to researching particular gene and protein properties such as alternative splicing features, gene organization peculiarities, and change of translation rate at nucleotide replacements. This comprehensive approach to investigation might become the first step to understanding molecular mechanisms for cardiovascular disease development.

We investigate DNA heterogeneity influence on expression rate, as protein quantity; nucleotide repeats characteristic for CAD and HCM, which can be use as molecular markers for symptomatic and asymptomatic cases of HCM and CAD and predict mRNA alternative variants with splicing disorders and their molecular markers.

Evaluate influence of DNA changing on translation rate and protein quantity allows tracking influence of DNA changing on gene expression that it can help in explaining molecular mechanisms of cardiovascular disease development. This method is based on codons division by growing and lowering translation rate.

An abnormal alternative splicing (AS) leads to different types of body's development pathologies, too. Thus, molecular markers for abnormal AS cases may assist in CAD and HCM diagnostics. Prediction of disorders in alternative splicing is based on exon-inron hydropathy determination. Hydropathy index is estimated by the ratio between purines and pyrimidines of codons.

The most frequent nucleotide repeats may appear to be very promising, too, as some CVD, affecting heart (e.g. myotonic dystrophy) are accompanied by multiplication of nucleotide repeats. We find out nucleotide repeats characteristic for CAD and HCM, which can be used as molecular markers for symptomatic and asymptomatic cases of HCM and CAD.

DYNAMIC PROGRAMMING ALGORITHM PARALLIZATION FOR PROTEIN FOLDING

Vladimir DULKO, Sergey FERANCHUK

United Institute of Informatics Problems of the National Academy of Sciences of Belarus, 6 Surganov Str., Minsk, Belarus
e-mail: dulko@inbox.ru

Dynamic programming algorithms are widely used in bioinformatics. In some situations they are very time-consuming and it is necessary to use parallel computations. We consider a problem of protein folding under constrains on the combinatorial search applied. Parallelization of these algorithms is not obvious. We presented an approach to this problem. The first and obvious step is to separate accumulating process from computation processes. An accumulating process should collect all results from computation processes and find a series of best results to be calculated next. But the problem is that the load of the only accumulating process can be high. To solve this problem, a notion of a *front* is introduced as a set of best results for some level of progress of the whole task. This front can be separated for several regions, and one accumulating process is assigned to each region. Computation processes are assigned to each accumulating process and take data from it, while accumulating processes can take results from each computation process. This solution is implemented on supercomputer SKIF K-1000 using MPI library and shows a good performance.

The work was partially supported by the INTAS project N 04-77-7178.

References

1. A. Godzik (2003) Fold recognition methods, In: *Structural bioinformatics,* P.E.Bourne, H.Weissig (Eds.), 525-546 (Wiley-Liss).
2. C. Guerra, S. Istrail (Eds.) Protein Structure Comparison: Algorithms and
Applications, In: *Protein Structure Analysis and Design*, C. Guerra, S. Istrail (Eds.), pp. 1-33, 2003.
3. C.Ferrari, C.Guerra Geometric Methods for Protein Structure Comparison, In: *Protein Structure Analysis and Design*, C. Guerra, S. Istrail (Eds.), pp. 57-82, 2003.

COMPUTATIONAL ANALYSIS OF EGF-LIKE DOMAINS

Ralph Arnold S. LASALA

Department of Biology, Ateneo de Manila University, Katipunan Road, 1108 Quezon City, Philippines
e-mail: rlasala@ateneo.edu

EGF-like domains, domains that are homologous to the epidermal growth factor (EGF), are important building blocks for extracellular proteins. They occur frequently in animal proteins and have been shown to function in diverse processess such as blood coagulation, complement activation, and developmental determination of embryonic cell fates. This paper describes the sequence analysis and homology modeling of EGF-like domains. Sequences from Bos taurus (cow), Canis familiaris (dog), Felis catus (cat), Homo sapiens (human), Mus musculus (mouse), Ovis aries (sheep), Pan troglodytes (chimpanzee), Rattus norvegicus (rat), Sus scrofa (pig), and species were studied in order to obtain insights into the biological significance of the EGF-like domain. An EGF-like domain consists of 30 to 40 amino acids, present in a conserved form. The EGF is a polypeptide of about 50 amino acids with three internal disulfide bridges. It binds to cell-surface receptors, induces dimerization, and initiates signal transduction that results in DNA synthesis and cell proliferation. A common feature of EGF-like domains is the inclusion of six cysteine residues involved in disulfide bonds. The fold consists of two-stranded β-sheet followed by a loop to a C-terminal short two-stranded sheet. This paper describes the structural and evolutionary conservation of EGF-like domains as results of computational analysis, using various methods such as sequence alignment and structural considerations. We present that most of the residues can be reasonably modelled using Swisss-Model. Results show that sequence analysis combined with comparative modelling can make useful predictions based on sequence and structural data.

References

1. E. Appella, I.T. Weber, F. Blasi (1988) Structure and function of epidermal growth factor-like regions in proteins. *FEBS Letters,* **231:** 1-4.
2. R.F. Doolittle, D.F. Feng, M.S. Johnson (1984) Computer-based characterization of epidermal growth factor precursor. *Nature,* **307:** 558-560.

DEVELOPMENT OF PROGRAM FOR PREDICTION OF FUNCTIONAL SITES IN PROTEIN SPATIAL PATTERNS

Alexandr MAGDYSYUK

Institute of Cytology and Genetics, Novosibirsk, Russia
e-mail: machine@gorodok.net

The theme of my current work is development of program for prediction of functional sites in protein spatial patterns. And now I will try to briefly examine some problems were set up in this work. First of all I want to say some worlds about computation methods for calculation of vibrational frequencies and normal modes of large systems.

During past 30 years some numerical methods for the harmonic analysis of large systems were developed. For example, the methods Normal Mode Analysis and Driven Molecular Dynamics were developed for determination of vibrational frequencies and normal modes of large systems, like molecules of protein, in the full conformational space (including all degrees of freedom) and in a reduce conformational space (reducing the number of freedom degrees). The underlying principle is that from the atomic fluctuation, an effective harmonic force field can be determined relative to the dynamic average structure.

Normal Mode Analysis (in future NMA) has long been used as a tool for interpreting vibrational spectra of small molecules. Although NMA is a quasi-quantum method, it is approximate, because only the harmonic motion of the system around a single potential minimum is taken into account. Moreover, there are a number of bottlenecks associated with an application of NMA to biomolecular systems that contain more than ten thousand atoms because of resource or time limitation.

Driven Molecular Dynamics (in future DMD) is resonance based method which has considerable potential for the study of large systems, where the Hessian-based method NMA is not feasible. Another important advantage of DMD over the Hessian-based NMA is the ability to study the molecular dynamics away from the harmonic region.

My scientific adviser (Vladimir Ivanisenko) sets up a working hypothesis, that vibrational frequencies and normal modes of molecules of proteins can be applied for prediction of functional sites in protein spatial patterns. Because of impossibility to obtain analytical results, the decision to develop some program was made. And my current work consists of some program development: programs for calculation of vibrational frequencies and normal modes of proteins using NMA and DMD methods and program for prediction of functional sites in protein spatial patterns, which will be used frequencies and normal modes of proteins.

Programs for calculation will be used for calculation of some important characteristic of well known protein functional sites from data base PDBSite, which has been developed in Institute of Cytology and Genetics. And then (use heuristic methods, for example, bar charts of vibrational amplitudes, frequency range etc.) we (that is my scientific adviser and I) will try to describe protein functional sites in mathematical terms, which can be used for prediction of new protein functional sites.

After this, if the last program can predict protein functional sites with a high probability, the hypothesis will be confirmed. Otherwise, the hypothesis will be refuted, in this case the implementations of the methods for calculation of normal modes and vibrational frequencies and parser of PDB and PDBSite data bases can be used in future as a part of protein structure visualization program (for vibrational motion visualization).

And, of course, it is necessary to say some worlds about current work and future plans. Now I finish my work on implementation common part of both methods (like some numerical routines, multithreading subsystem and special parser of PDB and PDBSite data bases) and NMA implementation. After that I'm going to begin calculation of normal model and study of results for protein functional sites from data base PDBSite use NMA method.

EXTRACTION OF QUANTITATIVE GENE EXPRESSION DATA FROM THE IMAGES OF GENE EXPRESSION PATTERNS IN DROSOPHILA EMBRYO WITH PROSTACK AND ISIMBIOS

Anna MATVEEVA, Konstantin KOZLOV, Maria SAMSONOVA

Department of Computational Biology, Center for Advanced Studies, St. Petersburg State Polytechnical University, 29 Polytechnicheskaya ul., St. Petersburg, 195251, Russia
e-mail: mdespb@mail.ru

The development of multicellular organisms involves the differential expression of many genes. Thus knowledge about spatial and temporal patterns of gene expression is crucial for understanding development.

In this work, we apply our new software package ProStack (**Pro**cessing **Stack**s) integrated with information management system iSIMBioS (Integrated Service Infrastructure for Molecular Biology Systems) to accurately acquire quantitative data from the images of segmentation gene expression patterns in early *Drosophila* embryo at syncytial blastoderm stage. Images were obtained by confocal laser scanning microscopy of fixed embryos. The 16 embryos were stained with fluorescence tagged antibodies to visualize gene expression patterns of

Even-skipped protein, *lacZ* mRNA of the *even-skipped* promoter-reporter construct and histones.

We built three workflows to quantify gene expression levels per nucleus and outside nucleus area in these embryos without loss of spatial information.

The first workflow combines three optical sections, obtained in each channel, into one image and puts the resultant images into standard orientation. The second workflow finds the area occupied by the embryo in the image. Third workflow builds a nuclear mask from the nuclear channel (histones) image and averages fluorescence intensities in each nucleus and outside nuclei area.

We also introduce the method to estimate the quality of nuclear mask by considering two classes of pixels. The pixels of the first class come from the area occupied by a nucleus and outlined with nuclear mask, while the second class encompasses all the pixels outlined by the watershed domain with the exception of pixels that are "on" in the nuclear mask. To estimate the accuracy of segmentation we calculate the ratio of variances in pixel values between and inside these two classes. For correctly segmented image this measure should be larger than for the incorrectly segmented one and consequently can serve as an accuracy estimator.

The support of the study by the NIH Grants RUB1-1578 and CRDF GAP award RBO1286, RFBR-NWO grant 047.011.2004.013 and the Russian Federal Contract № 02.467.11.1005 is gratefully acknowledged.

References

1. J. Jaeger, M. Blagov, D. Kosman, K. N. Kozlov, Manu, E. Myasnikova, S. Surkova, C. E. Vanario-Alonso, M. Samsonova, D. H. Sharp, J. Reinitz (2004a) Dynamical analysis of regulatory interactions in the gap gene system of *Drosophila* melanogaster, *Genetics*, vol. 167, pp. 1721–1737.

2. J. Jaeger, S. Surkova, M. Blagov, H. Janssens, D. Kosman, , K. N. Kozlov, Manu, E. Myasnikova, C. E. Vanario-Alonso, M. Samsonova, D. H. Sharp, J. Reinitz (2004b) Dynamic control of positional information in the early Drosophila embryo, *Nature*, vol. 430, pp. 368–371.

3. H. Janssens, D. Kosman, C. E. Vanario-Alonso, J. Jaeger, M. Samsonova and J. Reinitz (2005) A high-throughput method for quantifying gene expression data from early *Drosophila* embryos, *Development Genes and Evolution,* vol. 215, pp. 374-381.

4. D. Kosman, J. Reinitz, D. H. Sharp (1998) Automated assay of gene expression at cellular resolution, In: *Proceedings of the 1998 Pacific Symposium on Biocomputing*, pp. 6-17.

ANALYTICAL METHODS IN PROBLEMS OF RECOGNITION THE STRUCTURAL AND FUNCTIONAL ORGANIZATION OF GENETIC SEQUENCES

Ruslan TETUEV, Florencz F. DEDUS, Lyudmila KULIKOVA, Sergey MAKHORTIKH, Anton PANKRATOV, Nafisa NAZIPOVA

Institute of Mathematical Problems of Biology RAS, Pushchino, Moscow Reg., Russia
e-mail: radja@impb.ru

We represent here a new method for the decision of a problem of searching tandem repeats in the sequences of DNA. Tandem repeats is characterized by length of the repeated sample and frequency rate. Each copy of the sample is independently exposed to the further mutations (mismatches, inserts etc.). Through certain time each copy of repeats becomes strongly diverged, and exact tandem repetition becomes unobservable. The problem consists in recognition such diverged tandem repeats of unknown length. This approach finds periodicity in DNA sequence using the generalized spectral-analytical method [1]. First we calculate the profiles of percentage of $(G{+}C)$-content and $(G{+}A)$-content in the sliding window of the length N along the given DNA sequence $S{=}\{\, S_1, S_2, \ldots, S_n, \ldots, S_L\}$. These are two discrete functions:

$$f_i^{(G,C)} = 1/N \cdot \sum_{n=i+1}^{i+N} \left|(S_n = G) \cup (S_n = C)\right|,$$

$$f_i^{(G,A)} = 1/N \cdot \sum_{n=i+1}^{i+N} \left|(S_n = G) \cup (S_n = A)\right|,$$

where $i{=}1,2,\ldots,L{-}N{+}1$. Then we consider the product of these functions as an approximation:

$$f_i = f_i^{(G,C)} \quad f_i^{(G,A)} \approx \overset{K}{\underset{j=0}{}} C_j\varphi_j(t_i),$$

where $\{\varphi_j(t)\}$ - is a system of orthogonal polynomial functions, $K \le N$ - is the depth of the approximation. Using coefficients C_j of the approximation of f_i it is possible now to construct a functional of the special form characterizing a degree of periodicity of function f_i in i-nth position. Thus under certain conditions we could easy denote localization of periodicity in DNA.

It is very important to distinguish our approach from the other spectral methods [2,3]. The basic idea of our methods is to convert data into a signal (a discrete function), then to find an appropriate signal transformation that could lead us to right solution and, finally, implement all the necessary signal transformation in its spectrum. Note: we are concerned about certain biological data transformation using spectra, not in spectra themselves. Why polynomial basis instead of DCT (Discrete Cosine Transform)? In some cases it could be necessary to neglect signal trend, curvature etc. It is non-trivial when you using trigonometric functions as the basis of FFT (Fast Fourier Transform). Moreover it is much applicable to use polynomial bases if you had to obtain reasonable values of signal derivative and some other analytical operation. In cooperation with the other well-known advantage of FFT, noise reduction, this approach looks very promising and pretends to be a really powerful instrument in case of huge data handling. Some original schemes of signal transformation in spectra computing on parallel processors systems were proposed and implemented.

The work was supported partially by the grants No 04-01-00756, 04-02-17368, 04-01-00814, 04-07-90402 and 06-07-89274 of the RFBR and of the Russian Science Support Foundation.

References

1. F.F.Dedus et al. (1999) *Generalized spectral-analytical method of data processing. Problems of image analysis and pattern recognition.* Moscow, Mashinostroenie Publishing (In Russian).
2. D.Gusfield (1997) In: *Algorithms on String, Trees, and Sequences.* New York: Cambridge University Press, pages 255-257.
3. G. Benson (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucl. Acids Res.*, volume 27, pages 573-580.

PREDICTION OF ALTERNATIVE SPLICING FOR THE GENES RELATED TO GASTROINTESTINAL TRACT CANCER DEVELOPMENT

Aizhan TURMAGAMBETOVA

The Kazakh National University named after Al-Farabi, Almaty, Kazakhstan
e-mail: aichyck@mail.ru

It is known that gastrointestinal tract cancer (GITC) is an extremely frequent pathology. Risk of disease development may reach 5-6 %. Annually, nearly 1 million new GITC cases are recorded in the world. Half-decade survival rates of GITC have achieved approximately 60 % in developed countries and remain less than 40 % in developing countries. Primary risk factor for GITC development (GITCD) is the age. Adults over 55 years old may happen to suffer from that type of cancer most frequently.

Alternative splicing patterns may show whether changes in modes of alternative splicing are the reason or the consequence of the disease. For example, determination of mRNA coding protein with improper metabolic functions will allow to find out molecular markers for GITC diagnostics.

Alternative splicing prediction is based on measurements on intron and exon hydropathies. Exon and intron hydropathies depend on a nucleotide structure. Usually, exons are hydrophilic and introns are hydrophobic. Exon and intron hydropathies may vary with point mutations. This increases the risk of appearance of mRNA giving a product which may not execute metabolic functions. Computer analysis of genes determining exon and intron hydropathies conducted using SeqAnalyst program.

It is very important to find a correlation between properties of gene and availability for this gene alternative splicing. After search of genes with alternative splicing there are 2 important issues:

1. To find principled differences between properties of genes with alternative splicing and without it;
2. To reveal discrepancies in group of genes having alternative splicing.

For example, it is possible to look, on what properties and as hardly exons and introns differ for genes having and haven't alternative splicing.

Therefore, the analysis of corresponding genes and proteins will allow not only improving statistics of treatment, but may serve as a preventive source for proper diagnostics. As treatment of cancer directly relates to early diagnostics, and with achievements in the field of deciphering of genetic origin to tumor development at various molecular levels.

Acknowledgements. I gratefully acknowledge my scientific chef Ivachshenko Anatoly. Also, Aitasheva Zaure who help me with my English.

1. Kim, T.M, et al. (2005) Determination of genes related to gastrointestinal tract origin cancer cell using a cDNA microarray, *Clinical cancer research*, **Vol.11:** 79-86
2. Stamm, S., et al. (2005) Function of alternative splicing, *Gene,* **344:** 1-20
3. Lee, C., Atanelov, L., Modrek, B., Xing, Y. (2003) ASAP: the alternative splicing annotation project, *Nucleic Acids Research,* **Vol.31:** 1, 101-105

PARTICIPANT'S PRESENTATIONS

ORAL PRESENTATIONS

The Dynamics of Expression Patterns of Fushi Tarazu Gene in the Drosophila Blastoderm
Anna Matveeva
Department of Computational Biology, Center for Advanced Studies, St. Petersburg State Polytechnical University, St. Petersburg , Russia

Computational Analysis of EGF-like Domains
Ralph Arnold Lasala
Department of Biology, Ateneo de Manila University, Quezon City, Philippines

The Modified Fuzzy C-Means Method for Clustering of Microarray Data
Anna Taraskina, Evgeny Cheremushkin
Novosibirsk State University, Novosibirsk, Russia

Analytical Methods In Problems Of Recognition The Structural And Functional Organization Of Genetic Sequences
Ruslan Tetuev, Florencz Dedus, Lyudmila Kulikova, Sergey Makhortikh, Anton Pankratov, Nafisa Nazipova
Institute of Mathematical Problems of Biology RAS, Pushchino, Moscow Reg., Russia

Comparative Analysis of Sequences of Silicic Acid Transporters in Diatom and Chrysophycean Algae
Julia Masyukova
Limnological Institute Sb RAS, Irkutsk, Russia

Novel Mechanism of Polysaccharide Extension in Atomic Force Microscopy. Molecular Dynamics and Density Functional Calculations
Igor Neelov
Institute of Macromolecular Compounds RAN, St. Petersburg, Russia / University of Leeds, UK

POSTER PRESENTATIONS

Time Scale of Poxvirus Evolution
Igor Babkin
SRC VB "VECTOR", Koltsovo, Russia

Investigation of Alternative Splicing Features, DNA Heterogeneity Influence on Expression Rate for Genes and Proteins Responsible for Coronary Artery Disease and Hypertrophic Cardiomyopathy
Galina Boldina
The Kazakh National University named after Al-Farabi, Almaty, Kazakhstan

Dynamic Programming Algorithm Parallization For Protein Folding
Vladimir Dulko, Sergey Feranchuk
United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus

Molecules Versus Morphology in Oligochaeta Systematics
Victoria Liventseva
Limnological Institute Sb RAS, Irkutsk, Russia

Dynamical Multi-State Population Model for Description of Protein Folding/Unfolding
Elena Neelova
University of Leeds, UK

Population Genetic Polymorphism of Endemic Molluscs Baicalia Carinata (Mollusca:Caenogastropoda)
Tatiana Peretolchina
Limnological Institute Sb RAS, Irkutsk, Russia

Prediction of Alternative Splicing for the Genes Related to Gastrointestinal Tract Cancer Development
Aizhan Turmagambetova
The Kazakh National University named after Al-Farabi, Almaty, Kazakhstan

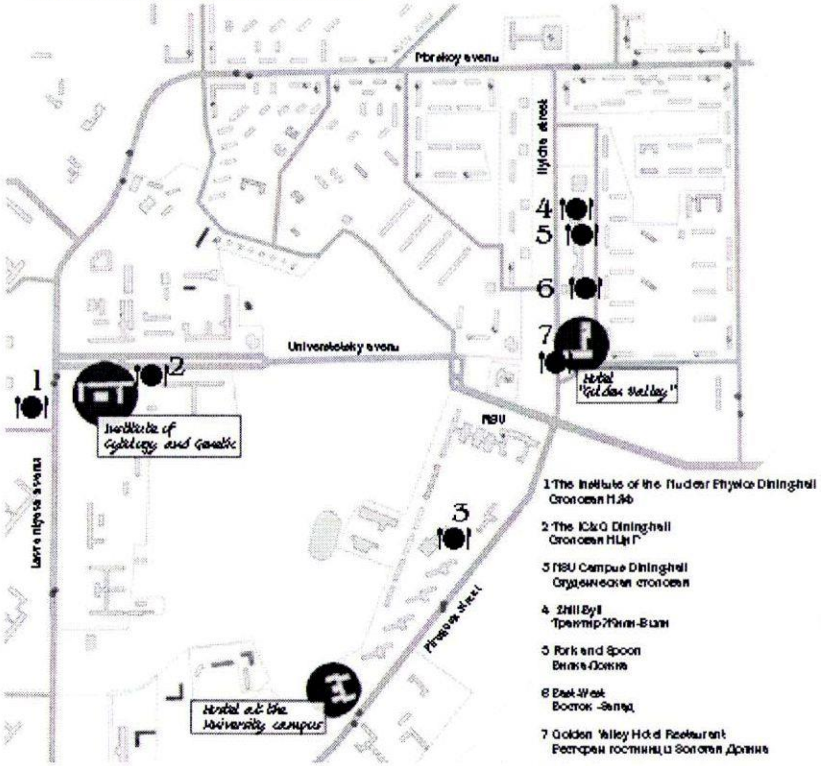
Amino Acid Preferences at the N-terminal Part of Eukaryotic Proteins Correlating with a Specific Contextual Organization of Translation Initiation Signal

Oxana Volkova
Institute of Cytology and Genetics, SB RAS, Novosibisk, Russia

Development of Program for Prediction of Functional Sites in Protein Spatial Patterns

Alexandr Magdysyuk
Institute of Cytology and Genetics, SB RAS, Novosibisk, Russia

AKADEMGORODOK



1. The Nuclear Physics Institute Dining Hall
2. The Institute of Cytology and Genetics Dining Hall
3. NSU Campus Dining Hall
4. Zhili-Byli Restaurant
5. Fork and Spoon Dining Hall
6. East-West Restaurant
7. Hotel Golden Valley Restaurant