

BIOINFORMATICS APPLICATIONS IMPLEMENTED

REPORT ON BI APPLICATION AND IMPLEMENTATION

Document Filename: **BG-DNA3.5-BI-Application.doc**

Activity: **NA3**

Partner(s): **VU, EENet**

Lead Partner: **VU**

Document classification: **PUBLIC**

Abstract: The BalticGRID's deliverable DNA3.5 represents the pilot BI application implementation and the results achieved. The activity describes pilot applications of bioinformatics: sequence pattern discovery and the gene regulatory network reconstruction, as well as modelling of biosensors and other reaction-diffusion processes.

In biology, a DNA sequence motif is a nucleotide pattern (i.e., sequences of letters A, C, G and T) that is spread in the genome and has a biological function. The motif finding procedure aims to identify conserved and overrepresented DNA motifs in a range of species. There are motifs identical in different species, as motifs are conserved in the course of evolution. To identify this an intensive computational procedures have to be produced.

The biosensors are either the catalytical biosensors, based on enzymatic substrates conversion or ether affinity biosensors attributed to specific interaction between antibody and antigen (or hapten), lectin and glycoside, receptor and ligand or complementary interaction of DNA. The implementation of this application is mainly relied on numerical optimization (global optimization) algorithm, specially designed for this bioinformatics task.



Document review and moderation

| | Name | Partner | Date | Signature |
|-----------------------------------|-------------|----------------|-------------|------------------|
| Released for moderation to | | | | |
| Approved for delivery by | | | | |

Document Log

| Version | Date | Summary of changes | Author |
|---------|------------|--|---|
| 0.1 | 13/02/2007 | Draft version | Igor Kuzmitshov, Linas Bukauskas, Julius Zilinskas |
| 0.2 | 22/03/2007 | Applications of EENet added (first draft) | Igor Kuzmitshov, Pavlos Pavlidis, Linas Bukauskas, Agne Brilingaite, Julius Zilinskas |
| 0.3 | 24/04/2007 | Applications of EENet (descriptions improved), the document restructured | Igor Kuzmitshov, Ilja Livenson, Pavlos Pavlidis, Algimantas Juozapavicius |
| 0.4 | 30/04/2007 | Final text improvements | Algimantas Juozapavicius |
| 0.5 | 15/06/2007 | Final version | Algimantas Juozapavicius |

CONTENTS

| | |
|--|-----------|
| 1. INTRODUCTION..... | 5 |
| 1.1. PURPOSE OF THE DOCUMENT | 5 |
| 1.2. GOALS OF APPLICATION SUPPORT | 5 |
| 1.3. ABBREVIATIONS | 5 |
| 2. EXECUTIVE SUMMARY..... | 6 |
| 3. DESCRIPTION OF BI PILOT APPLICATIONS..... | 8 |
| 3.1. ISPEXS | 8 |
| <i>Description of the Task Solved by the Application.....</i> | <i>8</i> |
| <i>Requirements for the Application.....</i> | <i>8</i> |
| <i>Implementation of the Application</i> | <i>9</i> |
| <i>Description of Data Sets Used</i> | <i>10</i> |
| <i>Description of Grid Specifics Applied to the Application</i> | <i>10</i> |
| <i>Job Execution Statistics, Results Achieved, Hours Used, Successful and Failed Jobs.....</i> | <i>10</i> |
| 3.2. SET OF COMMON BI APPLICATIONS | 13 |
| <i>Overview</i> | <i>13</i> |
| 3.3. OPTIMIZATION SCHEME FOR BIOSENSORS AND OTHER REACTION-DIFFUSION PROCESSES..... | 14 |
| 4. CONCLUSIONS..... | 16 |

1. INTRODUCTION

The Baltic Grid project aims i) to develop and integrate the research and education computing and communication infrastructure of the Baltic States into emerging European Grid infrastructure, ii) to bring the knowledge in Grid technologies and use of Grids in the Baltic States to a level comparable to that in EU member states with a longer experience in the development, deployment and operation of Grids, and iii) to further engage the Baltic States in policy and standards setting activities. The integration of The Baltic States into the European Grid infrastructure will primarily focus on extending the EGEE to the Baltic States (with which four partners are already engaged).

The BalticGrid e-infrastructures are used to design and implement most suitable applications from Bioinformatics, as researched by universities and academic institutions of the Baltic States. The goal of this deliverable is to identify pilot BioInformatics applications developed and implemented in the BalticGrid project, to provide analysis of design and technical aspects of such applications. During the design and implementation of such applications the project emphasis is laid also to the guidelines for extension of such applications to wider publicity and to other grid infrastructures, especially to possible cooperation with other bioinformatics application in grid infrastructures of EGEE, and similar ones.

1.1. PURPOSE OF THE DOCUMENT

The purpose of this document is to present the results of conceptual and technical analysis done for Bioinformatics application identification, design and implementation, while reaching milestones MNA3.1 (Technical analysis of applications in HEP, MS, BI). This document helps to identify the Bioinformatics most suitable or most advanced applications for the users of Baltic Grid, adjusting them to computing infrastructure in the best possible way. It presents also a conceptual and technical design for applications mentioned for BalticGrid infrastructure.

1.2. GOALS OF APPLICATION SUPPORT

The main objective of “BI Application implementation” in Baltic Grid project is to serve scientists from various universities of Baltic states involved in research and development of advanced biotechnological or biological research and related matters including methods and tools for the most advanced genome research, enzyme and drug design.

1.3. ABBREVIATIONS

BG – Baltic Grid
BI – Bioinformatics
JDL – Job Definition Language
SUP – Application Support
SE – Storage Element
VO – Virtual Organisation

2. EXECUTIVE SUMMARY

The activity NA3 (Application Identification and Support) identifies pilot applications and their communities for the BalticGrid and develops support for them. The recent report of the NA3 activity describes the pilot application of bioinformatics (BI):

- sequence pattern discovery and the gene regulatory network reconstruction.

These pilot applications running in BalticGrid are being developed by BIIT group lead by Jaak Vilo, University of Tartu, with assistance of EENet. These applications belong to the sequence pattern discovery area.

In biology, a DNA sequence motif is a nucleotide pattern (i.e., sequences of letters A, C, G and T) that is spread in the genome and has a biological function. The motif finding procedure aims to identify conserved and overrepresented DNA motifs in a range of species. There are motifs identical in different species, this is happening because these motifs are conserved in the course of evolution.

Researchers from different areas would like to search for motifs that are overrepresented in a set of genes and conserved among different species. Molecular biologists are interested in BI discoveries because wet-lab experiments are very time- and money-consuming. Doctors and molecular biologists are interested in finding motifs for studying of serious diseases. Computational biologists and bioinformaticians are also interested in motifs regulating genes to study gene regulatory networks.

The task of iSPEXS is to find such overrepresented motifs using homologous information, i.e., information about similarity of gene sequences (more precise: using orthologous information, i.e., information about similarity of gene sequences due to their shared ancestry).

iSPEXS is a set of bash and Perl scripts, added as archive to grid jobs' input sandbox. It uses SPEXS program (compiled binary), that can be either pre-deployed as a VO-specific application, or included in the archive with the scripts.

In order to simplify the process of running simple BI jobs on grid a set of tools was deployed gridwise (installed in the BalticGrid cluster VO-specific application area). These tools currently include the following:

- AlignACE (Aligns Nucleic Acid Conserved Elements) is a software used for the application to find sequence elements conserved in a set of DNA sequences.
- The MEME system is a tool for discovering motifs in a group of related DNA or protein sequences. A motif is a sequence pattern that occurs repeatedly in a group of related protein or DNA sequences. MEME represents motifs as position-dependent letter-probability matrices which describe the probability of each possible letter at each position in the pattern. Individual MEME motifs do not contain gaps. Patterns with variable-length gaps are split by MEME into two or more separate motifs. MEME takes a group of DNA or protein sequences (the training set) as input and outputs as many motifs as requested. MEME uses statistical modelling techniques to automatically choose the best width, number of occurrences, and description for each motif.
- Pratt is a tool that allows the user to search for patterns conserved in a set of protein sequences. The user can specify what kind of patterns should be searched for and how many sequences should match a pattern to be reported.
- SPEXS is able to exhaustively enumerate all patterns present in input sequences according to the pattern language definitions. Statistics on pattern occurrences and the output are calculated (e.g. how frequent they are in each of the given input data set). By setting thresholds

“uninteresting” patterns can avoid being reported. Others, for example motifs overrepresented in one of the data sets, can be output.

- Trie*agrep family. Exhaustive pattern matching, all against all.

Usage of BalticGrid resources allowed to dramatically reduce overall time of computation. For example, one run of iSPEXS (see Section 3.1) needs about 650 hours of computation on a single computer (with SPECint of about 1000), but it takes about 40 hours to run on the grid.

<http://www.spec.org/cgi-bin/osgresults?conf=cpu2000>

The other topic foreseen in the BalticGrid proposal:

- modelling of biosensors and other reaction-diffusion processes

is under research, development and testing by the group of researchers at Vilnius University, Vytautas Magnus University and the Institute of Mathematics and Informatics (Lithuania). The implementation of this application is mainly relied on numerical optimization (global optimization) algorithm, specially designed for bioinformatics task mentioned above.

The global optimization algorithm was modified mainly to cluster (MPI) environment, and the confidence of the algorithm depending on values of parameter as well as of different distances used were tested.

3. DESCRIPTION OF BI PILOT APPLICATIONS

This section lists BI applications run on BalticGrid:

- iSPEXS – a program for finding DNA segments that are conserved in a range of species;
- Set of common BI applications (can be used as subroutines or utilities),
- Optimization scheme and its implementation for reaction-diffusion processes.

3.1. ISPEXS

Description of the Task Solved by the Application

In biology, a DNA sequence motif is a nucleotide pattern (i.e., sequences of letters A, C, G and T) that is spread in the genome and has a biological function. The motif finding procedure aims to identify conserved and overrepresented DNA motifs in a range of species. There are motifs identical in different species; this is happening because these motifs are conserved in the course of evolution.

The overrepresentation of a motif in a cluster of genes A (sets of genes are usually called clusters) means that the motif is found in the genes of A with a statistically greater frequency than in the other genes of the organism. It also means that these motifs may be important for the particular cluster of genes. For example, if the motif ACTTCCG is overrepresented in the cluster “Glycolysis” (set of genes participating of glycolysis, i.e., decomposition of glucose and generation of energy) then it means that the regulation of many of the genes of the glycolysis pathway is governed by the regulatory element ACTTCCG; hence, this motif is important for the glycolysis pathway.

On the other hand, conservation means that natural selection has prevented mutations in this motif. This means that there is a need for the organisms this motif to stay as it is, not to change. This may mean that this motif is biologically functional (i.e., it participates in some biological procedure) and that is why natural selection prevents the establishment of any mutations in it, since mutations have the catastrophic effect usually.

Researchers from different areas would like to search for motifs that are overrepresented in a set of genes and conserved among different species. Molecular biologists are interested in BI discoveries because wet-lab experiments are very time- and money-consuming. Doctors and molecular biologists are interested in finding motifs for studying of serious diseases. Computational biologists and bioinformaticians are also interested in motifs regulating genes to study gene regulatory networks.

The task of iSPEXS is to find such overrepresented motifs using homologous information, i.e., information about similarity of gene sequences (more precise: using orthologous information, i.e., information about similarity of gene sequences due to their shared ancestry).

Requirements for the Application

iSPEXS is a set of bash and Perl scripts, added as archive to grid jobs' input sandbox. It uses SPEXS program (compiled binary), that can be either pre-deployed as a VO-specific application, or included in the archive with the scripts. To run iSPEXS jobs, the following software environment is needed:

- bash shell;
- Perl programming language;
- tar, gzip and bzip2 archiving tools;
- wget (to access BIIT's g:Orth web tool for getting orthologous information);

- gnuplot (for saving results as a picture).

Hardware requirements are the following:

- about 600 MB of file storage for downloaded databases and temporary data.

Implementation of the Application

iSPEXS uses a rational assumption that the probability for a conserved motif to be present in closely related species is higher than to be present in the most distant ones. Thus, using homologous clusters from closely related species, the motif discovery procedure is enhanced. On the other hand, homologous clusters from distant species may add noise.

This application is based on SPEXS (Vilo 1998, 2002), an algorithm that discovers motifs that are overrepresented in a cluster of genes. iSPEXS (incremental SPEXS) uses homologous information incrementally (adding species step-by-step), and SPEXS finds the overrepresented motifs in every step.

Incremental discovery of motifs is depicted in Figure 1: In each step (denoted by a Roman number) a species is added in the analysis. In the first step, the sequences from human are analysed. In the second step, the sequences from chimp have been added experimental data set. The motif is enhanced while more species are being added (the middle part, other mammals have been added). Enhancement means that the signal of the motif (i.e., a function of its p-value, a measurement of its statistical significance) becomes stronger. After the point of the maximum enhancement the signal starts to become weaker (the last part, the motif is detracted by adding data from some distant species).

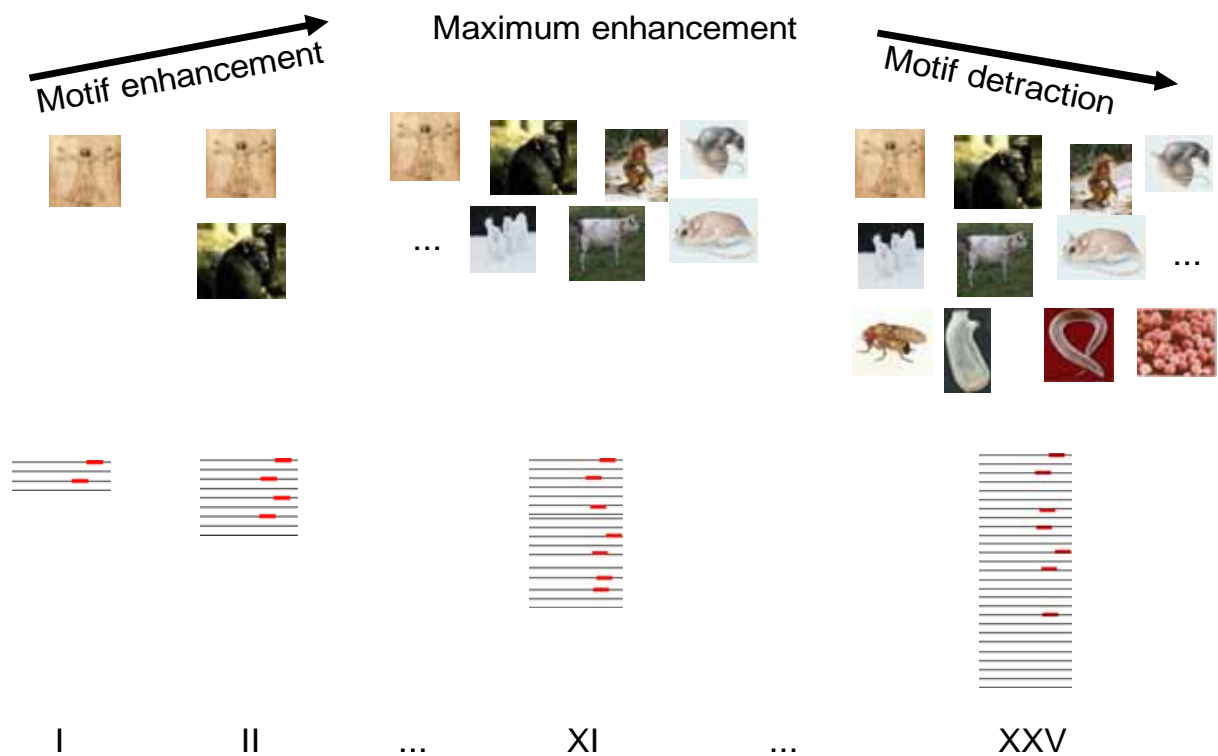


Figure 1: Motif enhancement

Lines in the bottom part of the image represent sets of promoter sequences for the species in each step. The red marks depict a match of a motif in the sequence.

Description of Data Sets Used

iSPEXS uses Ensemble database to get gene sequences and homologues. The Ensemble project (<http://www.ensembl.org/>) provides a comprehensive and integrated source of annotation of chordate genome sequences.

Description of Grid Specifics Applied to the Application

Additions to the application:

- Application scripts modified to be run on an arbitrary node;
- Pre-processing script added to generate a set of JDL files for an experiment;
- Running script added to download the data sets from SE, prepare data and program environment, and run application, saving results to SE.

Data management:

- Storing data sets (gene databases) in SE;
- Retrieving resulting output from SE.

Job Execution Statistics, Results Achieved, Hours Used, Successful and Failed Jobs

One run of the program (let us call it one query) takes a list of size s of human genes and finds its most significant motifs (in 25 steps, by the number of different species). The results are saved in two formats: as text (not shown here) and as image (see sample image on Figure 2). The image is good for human to quickly interpret the results (as described below); the text file is good for further processing by analysis programs.

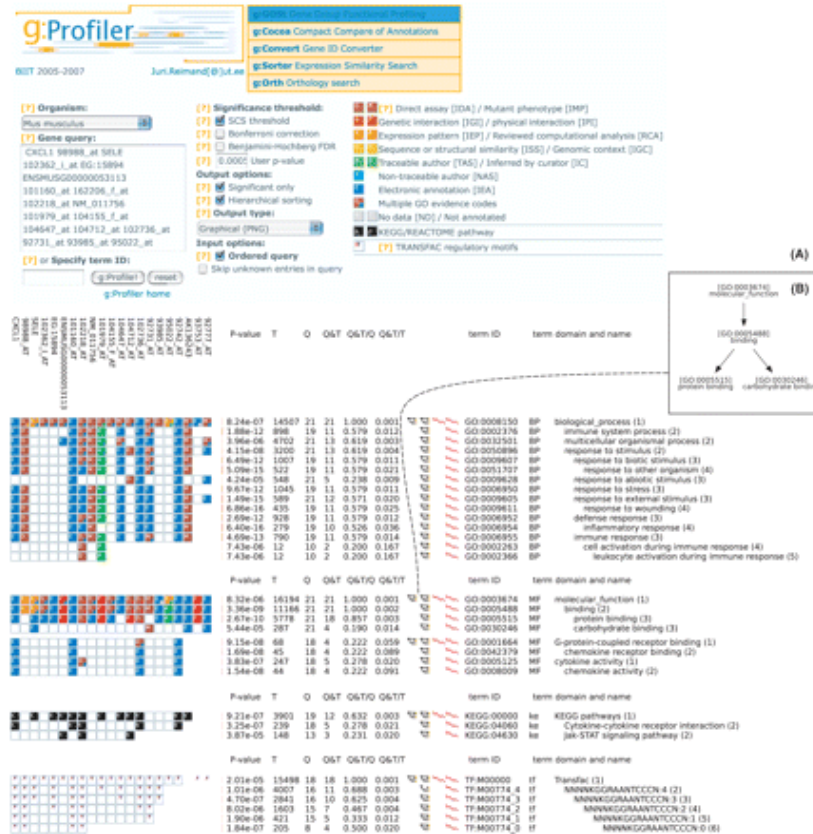


Figure 2. The significance of motifs.

Figure 2 presents typical user input and output scenario of g:Profiler. User inserts a set of genes in the main text window and optionally adjusts query parameters. Results are provided either graphically or in textual format. Genes are presented in columns, and significant functional categories in rows. The analysis of an ordered list shows the length of the most significant query head. GO annotation evidence codes are coloured like a heat map, showing the strength of evidence between a gene and GO term. The legend is provided at the top of the page. It is displayed when the user clicks on the tree icon on the results page. The g:Orth, g:Convert and G:Sorter tools are directly linked to relevant genes from the current query. Additional examples are available in Supplementary Data. (B) Hierarchical relations between the resulting GO categories can be browsed by clicking on corresponding icons. Positions denote the significance of different motifs. The y-axis measures the significance, in p-value terms, and the x-axis denotes how many species has been used in the analysis. Each line represents the p-values of one motif.

So, quick visual interpretation is that we look for lines having minimum points with:

- the smallest y-value (more significant compared to others), because the lower that the lines go, the lower the p-value is, thus the probability that this motif is important is higher;
- the largest x-value (significant for many species), because the motif is conserved for more species.

For example, if the minimum value of a motif is represented by the pair $(3, 4 \cdot 10^{-55})$ then, for three species used, the p-value for the specific motif is $4 \cdot 10^{-55}$. However, if it would be $(10, 4 \cdot 10^{-55})$ then the same motif would be conserved not only for three species, but for ten and this would be more important biologically.

It is necessary to compare results for random gene data and real gene data. Motifs in real data have much smaller p-value, hence, are more significant.

We need to run application on large amount of random data to determine the threshold of significance (to be used as a baseline for discoveries in real data). For a given size s of the gene list, an experiment of 1000 queries is enough to get an accurate estimation. We used experiments of 2000 queries of a given size expecting some jobs to fail. Then we have to run experiments for different sizes s to get the threshold estimation we need. For running on the grid, queries of one experiment are grouped into 40 jobs of 50 queries. Computation of one experiment takes approximately:

- 650 hours on a single computer;
- 40 hours on the grid.

In just a few days we could get enough data to use in the draft of an article. For the moment, about ten experiments have been run.

In the beginning of the pilot, BalticGrid VO was used. Later a separate VO, BIIT, has been taken into use (one cluster in Estonia first, two clusters in Estonia and one in Latvia currently).

Current problems running jobs on the grid include:

- The main problem is that myproxy cannot be used and many jobs are aborted due to the proxy expiration, resulting in loss of 30% of the results (problem does not occur when using balticgrid VO; reason remains unknown);
- Results of every query are saved on SE using LFC commands. Due to occasional LFC timeouts, saving results can fail.

These restrictions and failures were discussed at the BalticGrid events with operational people, supporting middleware. The reasonable conclusions are not reached yet.

3.2. SET OF COMMON BI APPLICATIONS

Overview

In order to simplify the process of running simple BI jobs on grid a set of tools was deployed gridwise (installed in the BalticGrid cluster VO-specific application area). These tools currently include the following:

- AlignACE (Aligns Nucleic Acid Conserved Elements) is a software used for the application to find sequence elements conserved in a set of DNA sequences.
- The MEME system is a tool for discovering motifs in a group of related DNA or protein sequences. A motif is a sequence pattern that occurs repeatedly in a group of related protein or DNA sequences. MEME represents motifs as position-dependent letter-probability matrices which describe the probability of each possible letter at each position in the pattern. Individual MEME motifs do not contain gaps. Patterns with variable-length gaps are split by MEME into two or more separate motifs. MEME takes a group of DNA or protein sequences (the training set) as input and outputs as many motifs as requested. MEME uses statistical modelling techniques to automatically choose the best width, number of occurrences, and description for each motif.
- Pratt is a tool that allows the user to search for patterns conserved in a set of protein sequences. The user can specify what kind of patterns should be searched for and how many sequences should match a pattern to be reported.
- SPEXS is able to exhaustively enumerate all patterns present in input sequences according to the pattern language definitions. Statistics on pattern occurrences and the output are calculated (e.g. how frequent they are in each of the given input data set). By setting thresholds “uninteresting” patterns can avoid being reported. Others, for example motifs overrepresented in one of the data sets, can be output.
- Trie*agrep family. Exhaustive pattern matching, all against all.

Addition of new tools is a simple process:

- The user sends a request to the support staff (*lcgadmin* of the user’s VO), specifying the location of the program, installation procedure (in case it's not obvious) and the legal information about the usage.
- In case the request passes the evaluation, it is installed and the new tag is published in the information system, indicating that the software is tested and ready for use.

3.3. OPTIMIZATION SCHEME FOR BIOSENSORS AND OTHER REACTION-DIFFUSION PROCESSES

The biosensors can be separated into two groups according to a principle of their action. The biosensors of the first group, i.e. the catalytical biosensors are based on enzymatic substrates conversion. The specific interaction between antibody and antigen (or haptén), lectin and glycoside, receptor and ligand or complementary interaction of DNA is the basis of affinity biosensors attributed to the second group.

The implementation of algorithms based on numerical search for the optimal solution permitted to model the behaviour of biosensors at complex boundary conditions, at external and internal diffusion limitations and complicated enzyme processes. The phenomena of biosensor signal amplification by cyclic substrates conversion has been described. The next step of the investigations is to build the model describing a multilayer biosensors with perforated and porous membranes.

Global optimization method, used in the case of biosensors has however much more general context. It may be used to solve practical problems across many branches of engineering, many problems in applied sciences, and even in basic ones. Lots of new global optimization methods are being proposed. A method, developed recently and implemented in Balticgrid environment gives advances in global optimization practice focusing on novel theoretical and algorithmic achievements.

Application of algorithms to solve practical problems crucially depends on efficiency and reliability of algorithms implementing global optimization methods. However development of such algorithms is not trivial and a mathematician proposing new methods for global optimization is not necessary a good programmer.

The method suggested for BG implementation presents a C++ package for development of algorithms implementing covering global optimization methods. The package includes vector templates to define feasible region and subregions, heap and queue templates to define lists of tasks and solutions, implementation of branch and bound algorithm, implementation of timer for measuring speed of algorithms, definitions of mathematical test functions and example problems for global optimization. Only the evaluation of bounds for the values of the objective function over the sub-region should be implemented by the user. Global optimization algorithms are based on interval arithmetic and balanced random interval arithmetic and have been implemented using the proposed package.

When computing power of usual computers is not sufficient to solve a practical global optimization problem, the high performance parallel computers may be helpful. An algorithm is more applicable in case its parallel implementation is available, because larger practical problems may be solved by means of parallel computers. Because of that tools for parallelization of global optimization algorithms have been included in the proposed package. A standardized message-passing communication protocol MPI is used for communication between processors.

The optimization method is using two dimensional scales, id est there are n objects, which dissimilarity is given by a dissimilarity matrix δ_{ij} . Algorithm searches for points in 2D space $x_i = (x_{i1}, x_{i2})$, $i = 1, \dots, n$, that their distances d_{ij} are the nearest to the dissimilarity of objects. These points form an 2D image. The optimal criteria usually used is known as a STRESS function:

$$S(X) = \sum_{i < j}^n w_{ij} (d_{ij}(X) - \delta_{ij})^2,$$

here $X = ((x_{11}, \dots, x_{n1}), (x_{12}, \dots, x_{n2}))$; weights have positive values: $w_{ij} > 0$, $i, j = 1, \dots, n$; distances between points \mathbf{x}_i and \mathbf{x}_j are denoted by $d_{ij}(X)$ and stress the dependence on points coordinates. Distances are expressed by Minkowski metrics:

$$d_{ij}(X) = \left(\sum_{k=1}^2 |x_{ik} - x_{jk}|^r \right)^{\frac{1}{r}}$$

Tables below present results of time used for the modelling procedures when Euclidean distance ($r = 2$) and city block distance ($r = 1$) are used:

| | $p = 20$ | $p = 40$ | $p = 60$ | $p = 80$ | $p = 100$ |
|--------------------|-----------|-----------|-----------|-----------|-----------|
| $N_{init} = 1000$ | 46 | 46 | 46 | 46 | 46 |
| $N_{init} = 2000$ | 53 | 53 | 53 | 53 | 53 |
| $N_{init} = 4000$ | 50 | 50 | 50 | 50 | 50 |
| $N_{init} = 6000$ | 45 | 44 | 45 | 45 | 45 |
| $N_{init} = 8000$ | 42 | 42 | 42 | 42 | 42 |
| $N_{init} = 10000$ | 43 | 43 | 43 | 43 | 43 |
| $N_{init} = 12000$ | 39 | 39 | 36 | 39 | 40 |

Table. 1. The dependance of algorithm's efficiency from parameter p values in the case of Euclidean distance

| | $p = 20$ | $p = 40$ | $p = 60$ | $p = 80$ | $p = 100$ |
|-------------------|----------|----------|-----------|-----------|-----------|
| $N_{init} = 2000$ | 58 | 68 | 93 | 83 | 87 |
| $N_{init} = 4000$ | 57 | 83 | 86 | 89 | 85 |
| $N_{init} = 6000$ | 55 | 85 | 93 | 89 | 85 |
| $N_{init} = 8000$ | 52 | 80 | 88 | 92 | 87 |

Table. 2. The dependance of algorithm's efficiency from parameter p values in the case of city-block distance

4. CONCLUSIONS

For the conclusion, the usage of Grid infrastructure for BI applications (both for motifs computing as well as for biosensors modelling) proven to be very useful and practical. Usage of BalticGrid resources allowed users to dramatically reduce overall time of computation. In particular, one run of iSPEXS requires about 650 hours of computation on a single computer (with SPECint of about 1000), but it takes about 40 hours to run on the grid. Global optimization for biosensors is mainly designed for the cluster environment and the grid access to computing procedures also helps to find oput the best suitable and fastest cluster to compute.