



# EMBnet.news

Volume 14 Nr. 4  
December 2008

- **The EMBnet 20th Anniversary Conference**
- **Tutorial on Grid Computing**
- **The EMBRACE Registry**
- **5th RIB Programme and Abstracts  
and more ...**

# Editorial

This is the closing number of the year 2008, the year of EMBnet's 20th anniversary. Naturally, we are all cheerful (hence the fireworks on the cover), in a time where again new challenges present themselves in the next decade of our existence as a network. Fortunately we can say that in spite of several shifts in our *modus operandi* we keep the same spirit and the interaction. Our target, bioinformatics users in the whole world, rewards our effort as is clearly seen by the enormous amount of downloads that are daily taken from our servers, mainly the online collection of this publication. We hope that this rather full issue will please you. On top of the reporting sections that depict several aspects of our life as a community, you will find plenty of technical contributions that we believe will be of great use to bioinformatics practitioners. We are also publishing the abstracts of the 5th RIB Congress, hoping to contribute to the enhancement of the connections between our networks.

In the next volume of EMBnet news you will start seeing the result of our ongoing efforts to better structure this newsletter. With this effort we hope to step-up in maturity, matching the importance of EMBnet's 21st year of existence as a networked community. Enjoy the reading.

The editorial board: Erik Bongcam-Rudloff, Domenica D'Elia, Pedro Fernandes, Andreas Gisel and Lubos Klucar.



Protein Spotlight (ISSN 1424-4721) is a periodical electronic review from the SWISS-PROT group of the Swiss Institute of Bioinformatics (SIB). It is published on a monthly basis and consists of articles focused on particular proteins of interest. Each issue is available, free of charge, in HTML or PDF format at

<http://www.expasy.org/spotlight>

We provide the EMBnet community with a printed version of issue 97. Please let us know if you like this inclusion.

Cover picture: *New Year celebration, 2008* [© Erik Bongcam-Rudloff]

# Contents

Editorial .....	2
The EMBnet 20th Anniversary Conference and Annual General Meeting 2008 .....	3
Tutorial on Grid Computing.....	15
Authenticated Grid access with robot certificate and the GENIUS Grid Portal .....	17
The Grid Problem Solving Environment for Bioinformatics: the LIBI experience.....	20
Querying the LIBI federated database through a data abstraction model .....	30
AntiHunter 3.0 Identification of antisense transcripts: a practical tutorial .....	41
The Job Submission tool (JST) .....	44
Gene Analogue Finder: a GRID solution to find functional analogous genes.....	48
GRID distribution supporting chaotic map clustering on large mixed microarray data sets .....	52
Successful EMBnet-EMBRACE joint Webservices workshop .....	56
The EMBRACE Registry.....	58
Tutorial "Introduction to Bioinformatics" .....	63
Protein spotlight 97 .....	64
5th RIB Programme and Abstracts .....	76
Node information.....	88

## Editorial Board:

Erik Bongcam-Rudloff, The Linnaeus Centre for Bioinformatics, SLU/UU. SE

Email: [erik.bongcam@bmc.uu.se](mailto:erik.bongcam@bmc.uu.se)

Tel: +46-18-4716696

Fax: +46-18-4714525

Domenica D'Elia, Institute for Biomedical Technologies, CNR, Bari, IT

Email: [domenica.delia@ba.itb.cnr.it](mailto:domenica.delia@ba.itb.cnr.it)

Tel: +39-80-5929674

Fax: +39-80-5929690

Pedro Fernandes, Instituto Gulbenkian. PT

Email: [pfern@igc.gulbenkian.pt](mailto:pfern@igc.gulbenkian.pt)

Tel: +315-214407912

Fax: +315-214407970

Lubos Klucar, Institute of Molecular Biology, SAS Bratislava, SK

Email: [klucar@embnet.sk](mailto:klucar@embnet.sk)

Tel: +421-2-59307413

Fax: +421-2-59307416

Andreas Gisel, Institute for Biomedical Technologies, CNR, Bari, IT

Email: [andreas.gisel@ba.itb.cnr.it](mailto:andreas.gisel@ba.itb.cnr.it)

Tel: +39-80-5929662

Fax: +39-80-5929690

## The EMBnet 20th Anniversary Conference and Annual General Meeting 2008



**Laurent Falquet**  
Swiss Institute of  
Bioinformatics, Lausanne

The 20th anniversary conference and EMBnet constituency AGM08 was held in Martina Franca (Italy). The organisers Domenica D'Elia and Andreas Gisel did a great job in preparing the conference and welcoming the participants with a smile.



Andreas & Domenica welcome.

More than 120 participants from 38 countries joined the anniversary conference at the Park Hotel San Michele for 3 days of conference and poster sessions.

### 20th Anniversary Conference report

After a nice introduction and welcome of the participants by Domenica D'Elia, Prof. Cecilia Saccone reminded us about EMBnet history, the first years and how the director of EMBL voted against the creation of the library of sequences. Despite this vote, the EMBL database was created and a few years later EMBnet was born as

an efficient way to distribute it among member states and provide support, training and help desk in the various European countries. She explained that the goals have not much changed since the 80's, apart from an opening to the rest of the world. We should continue to expand the network, touch new fields for teaching (interdisciplinary) and develop research.



Prof. Cecilia Saccone during her talk about the EMBnet history.

Erik Bongcam-Rudloff (chairman) gave a nice view of the future of EMBnet with a shuffled movie to explain the huge amount of data coming. He also did a comparison of network speed and data traffic amounts. Erik also proposed to initiate the discussion of a new name for EMBnet, if possible keeping the acronym. The new name should reflect the links to other collaborating networks: Ibero-American network for Bioinformatics (RIBio), Asia-Pacific Bioinformatics Network (APBioNet) and African Society of Bioinformatics and Computational Biology (ASBCB). This proposal is now a ongoing discussion.



Erik Bongcam-Rudloff talking about the future of EMBnet.



Official "group picture" of EMBnet Conference 2008

## First session: Bioinformatics for Biodiversity

Two keynote speakers, Sarkar Indra Neil and Mehrdad Hajibabaei, nicely introduced the subject of tree of life, biodiversity and barcoding sequences. M. Hajibabaei even showed that using species barcoding two students were able to challenge the identification of commercial fishes in Sushi restaurants of New York City.

Other lectures were given by Monica Santa Maria (Barcode markers in fungi) and by Teresa Regina (Introns as markers in plant mitochondrial DNA).

## Second session: Training and E-Learning

Tin Wee Tan, founding secretary of APBioNet and keynote speaker, presented the APBioNet and its historical links with EMBnet as well as its role

in training bioinformatics. This network is just 10 years old, organizes regular conferences since 2002 and training programmes supported by FAOBMB, IUBMB and UNESCO. They participate in the APBioGrid and produce a LiveCD based on an extension of Bioknoppix, as well as a LAMS learning activity management system (see report of the round table).

Pascal Hingamp described his experience in teaching bioinformatics "by doing" for undergraduate courses of biology Bachelor students in Marseille. They developed an annotation system for the sequences found in the Global Ocean sampling expedition of Craig Venter. Students are asked to report about their findings in public wikiproteins and as a result they are much more motivated than doing standard canned exercises! See Annotation web site.

Other interesting lectures were given by Richard Kamuzinzi (Report on SIMDAT project),



Sarkar Indra Neil - Biodiversity Informatics: Enabling a Macroscopic View of Biology.



Mehrdad Hajibabaei - The Barcode of Life: Bringing Genomics to Biodiversity.



Tin Wee Tan - Policies, Network, Resources, Materials and Curricula for advancing bioinformatics education: 10 years of APBioNet.

Reinhard Schneider (Review of the International Society for Computational Biology, ISCB), Kristian Rother (Programming without a computer science degree), Steve Pettifer (UTOPIA), Patrice Duroux (IMG2 database).

### Third session: "Omics", comparative studies and evolution

The keynote lecturer Alexander E. Kel described the BioBase Inc. company founded in Germany, with offices worldwide. He mentioned the TransFac database and various tools related to transcription factors. Especially he stressed the role of regulation of genes vs the number of genes and the variety of combinations of transcription factors sitting on enhancers and promoters.



Alexander E. Kel - Evolution of gene regulatory code.

Other interesting lectures were given by Jérôme Lane (IMG2/LIGmotif), Ana Conesa (Functional evaluation of time course microarray data), Eija Korpelainen (CHIPSTER use R package and Bioconductor with a graphical interface),

Viviana Piccolo (Analysis of miRNA in *vitis vinifera*), Christian Salgado (Model for retention time prediction in a hydrophobic column), Goran Neshish (Identification of essential 3D structure parameters for maintaining the 2D structure), Vincent Miele (MIXNET software to derive statistics on a model), Alexandru Floares (RODES algorithm for drug dependent gene regulatory networks), Helena Strömbergsson (Exploring the protein-ligand space), Marcella Attimonelli (RHUMTS, reference human mitochondrial sequences), Guy Perrière (Homologue gene database for comparative genomics).

### Fourth session: Advanced Bioinformatics Technologies and Applications

The keynote lecture was given by Vincent Breton, he stressed the importance of GRID for life sciences and the federation and sharing of resources. Apart from some security issues already known, the users face long latency, high failure rate and desperately need a user-friendly interface. BioMed Virtual Organization currently accounts for 7% of the CPU and jobs on the GRID. As an example, he mentioned the EMBRACE curation project to recalculate PDB structures from original X-Ray and NMR data. He pushed everyone to adopt standards and wished the EBI become the centre of gravity for GRID in life sciences.



Vincent Breton - Grids for Life Sciences: status and perspectives.

Other interesting lectures were given by Fabio Polticelli (Non-natural proteins structure prediction on the GRID), Guiseppe La Rocca (Genius grid portal and USB key as certificate), Guillaume Rizk (RNA-RNA interaction by GPU accelerated), Ernesto Picardi (HTC for ASPic), Andreas Gisel

(ENGINEDB, functional analogues by gene ontology), Maria Roubelakis (GOMIR analysis of miRNA and GO clustering), Olivier Lespinet (Finding orphan EC numbers and assigning potential proteins), Alvaro Martinez Barrio (Integrating ERV sequence and structural features with DAS and EBIOX), Domenica D'Elia (UTR annotation via pattern mining), Gianfranco Tarricone (Bioinformatics knowledge discovery via grid computing).

## Round table discussion report

### New tools for Bioinformatics teaching (summarized by V. Ioannidis (CH))

The following tools for Bioinformatics teaching, which are already used by several EMBnet members, have been presented during the session:

1. The Learning Activity Management System (LAMS), presented by Tin Wee Tan: <http://www.lamsinternational.com/>
2. BioManager, presented by Sonia Cattley, <http://www.angis.org.au/>
3. Bioinformatics Training through Metagenomic Sequence Annotation (Annotathon), presented by Pascal Hingamp (Marseille), <http://annotathon.univ-mrs.fr/>
4. USB key based on QEMU, presented by Jose Ramon Valverde
5. BIOSLAX live CD suite of bioinformatics tools, presented by Tan Tin Wee, <http://www.bioslax.com/>
6. Swedish "Kenya" USB key, presented by Erik Langercrantz from Erik Bongcam-Rudloff's group.

LAMS (1) is a convenient tool for designing, managing and delivering online collaborative learning activities. It provides teachers with a highly intuitive visual authoring environment for creating sequences of learning activities. These activities can include a range of individual tasks, small group work and whole class activities based on both content and collaboration.

Although similar to Moodle, its most prominent added value relies in its ability to create structures of learning activities using a graphical interface.

Its main disadvantage is that powerful import/export options are only available in a commercially supported version.

BioManager (2) is an integrated bioinformatics workspace providing a single web interface

to many bioinformatics tools. Users can upload or paste their data into BioManager or use Text Search to extract data from one of the available databases, including GenBank, SWISS-PROT, Blocks, Prosite, Enzyme, Pfam, and StackDB. BioManager records and allows users to view the history of all their analyses, providing a "virtual lab book". Regular analyses can be recorded as a macro for easy automation and standardization. Teachers or supervisors have also several tools to manage and follow users analyses.

Unfortunately, BioManager is very tightly linked to the ANGIS environment (<http://www.angis.org.au/>) and can therefore not easily be installed in another location.

The Annotathon (3) is an internet teaching platform devoted to metagenome annotation. During the discussion, it appeared that it could be also used by EMBnet members for Bioinformatics teaching courses.

This tool was presented during session 2 and also during the poster session (abstract 10). It provides a collaborative environment for practical training and evaluation of students by reiterated use of common bioinformatics tool to annotate sequences from the "Global Ocean Sampling" project.

The tool is hosted at <http://annotathon.univ-mrs.fr/> and is openly available both as OSS and for direct use by the EMBnet community. It has extensive documentation for both teachers and students.

QEMU/KVM (4) based USB key. A work in progress to tackle the difficult problem of teaching and providing bioinformatics tools in areas constrained by privilege and/or resource scarcity (e.g.

developing areas, users with access only to internet cafes). The USB key contains a Linux operating system packed with tools and the QEMU OSS emulator/virtualizer.

The key can be used to boot the contained operating system directly from it. But in addition, when booting is not possible, the inclusion of QEMU allows users to run Linux in parallel with their OS of choice (Linux, Windows or Mac). This virtual instance can be run at near native speed if the user has administrator privileges to install the virtualizing module (much like the free VMware client or server), or at emulation speeds (~1/10) otherwise (unlike any other tool).

In order to run smoothly under emulation or very old PC's (1/10 of current speed) the key uses a lightweight environment (Puppy Linux based) and may require as little as 128 MB of RAM.

The contents of the key provide support for sequence analysis (with EMBOSS/EMBASSY, wEMBOSS and Staden), evolution (Clustal, Phylo\_win, PHYLIP), statistics (R), structural biology (using 'gratis' software: SPDBV, TINKER, TRITON), drug design (AUTDOCK4, 3D-DOCK, GRAMM) and quantum chemistry (mopac, mpqc, psi3, ghemical, openmol). It also provides generic support (office, network, multimedia), automatic network support and secure utilities like SSH and SSHFS.

The major limitations are:

- software is mostly unconfigured pending testing in a training environment to decide if configuration should be taught or not;
- Structural Biology software is free but its distribution requires permission from authors, which is being worked out. Fair use exceptions for academic use may apply in most countries but should be checked out.

Further inquiries should be sent to J. R. Valverde.

*BIOSLAX* (5) is a live CD of Bioinformatics tools. It already contains the most important tools. It is composed of two distinct parts:

- the core system, which contains the OS itself and basic tools;
- the individual modules comprise the individual utilities that users want to have on their system. Since these modules can be put in or removed prior to burning the image disk, the system is fully modular and easily customized.

*BIOSLAX* can also be run using VMware client, but this requires administrator privileges on the machine to install the virtualizer.

Tin Wee circulated a CD version of *BIOSLAX*, which is available for testing purposes on request. Note that a USB key version is oncoming.

*Swedish "Kenya" USB key* (6). This is a 4GB USB key developed by Erik Bongcam and two students (Erik Langercrantz and Alvaro Barrio) for a course in Kenya. It is also going to be used soon at a course in Rosario, Argentina by Oswaldo Trelles (relayed through J. R.).

This USB key is based on Ubuntu Linux and is used to boot the machine into a working environment with many bioinformatics tools; it can also be used to install the system from it.

The tools contained include EMBOSS/wEMBOSS, EMBASSY, PHYLIP, Staden, Clustal and other sequence analysis tools as well as a number of structure visualization tools. All the tools are fully configured to work off the Internet, with sample databases included, or with the eBioX kit developed at Sweden installed on a Mac based server with a full release of relevant databases, software and MRS.

The USB key is available from Erik's group (there is a mirror copy in Spain). Further enquiries about the Mac-based eBioX+databases server should be directed to Erik Bongcam.

## Annual General Meeting of the EMBnet Stichting

### Presentations

Argentina (Oscar Grau – treasurer), Australia (Sonia Cattley), Belgium (Guy Bottu), Brazil (Goran Neshich), Chile (Christian Salgado), Colombia (Emiliano Barreto), Costa Rica (Allan Orozco), Finland (Kimmo Mattila), Greece (Charalampos Moschopoulos), Hungary (Endre Barta), India (Akash Ranjan), Italy (Domenica D'Elia), Mexico (Cesar Bonavides), Norway (George Magklaras), Pakistan (Chohan Shahid), Poland (Piotr Zielenkiewicz), Russia (Sergei Spirin), Slovakia (Lubos Klucar), South Africa (Heikki Lehvälaiho), Spain (Jose Ramon Valverde – EB member), Sri Lanka (Kamani Tennekoon), Sweden (Erik Bongcam-Rudloff – chairman), Switzerland (Laurent Falquet – secretary), IRLI-BECA – Kenya (Etienne de Villiers), UMBER – UK (Teresa Attwood)

### Observers or members of EMBnet committees, with no voting right

Andreas Gisel (Italy), Nils-Einar Eriksson (TMPC chairman, Sweden), Vassilios Ioannidis (Switzerland, ETPC chairman), Guy Perrière (applicant France), Tiziana Castrignano (applicant CASPUR), Alvaro Martinez-Barrio (BMC, Sweden), Matej Stano (ETPC, Slovakia), Eija Korpelainen (CSC, Finland), Ana Tereza Ribeiro de Vasconcelos (BR)

### Represented:

China (proxy to Switzerland), Portugal (proxy to Colombia)

### Absent with apologies:

Austria, Cuba, The Netherlands, EBI, ICGEB

### Absent:

Canada, Israel, ETI, IHCP/BGMO, MIPS/GSF

The Minutes of the AGM 2007 in Malaga Spain were approved without modification.

### Financial report

Due to the 20th anniversary expenses, the accounts are in slight decrease as compared to last year.

We opened a PayPal account in August 2007 to allow for donations via our website. Unfortunately nothing was given to EMBnet since then. Recently we received a request from PayPal to clarify the non-profit status of our account. Work is in progress to solve the issue.



Oscar Grau presenting the EMBnet financial report.

### Re-election of nodes

The following nodes were up for re-election after a 3 years period:

Australia, Belgium, Costa Rica, China, IHCP-BGMO, Italy, Portugal, Russia, Slovakia and Switzerland.

All nodes were re-elected for 3 years, except BGMO that did not reach the two-third limit for positive votes. The specialist node BGMO is noted as suspended until further notice and will be up for re-election in 2009. Costa-Rica was re-elected with apologies from the secretary, because Costa Rica was not up for re-election in 2008 but in 2009. Thus Costa Rica will be up for re-election next time in 2011.

### Nodes discharging

After two consecutive votes (2007 and 2008) Israel is now officially discharged of the EMBnet board.

The EB received a step-down letter from the Austrian ministry of Science, the resignation of Austria was accepted with regrets due to the very good work done by Martin Grabner, its former node manager. Austria is now officially dis-



Eija & Vassilios counting the ballots.

charged of the EMBnet board. JR will try to transfer the data and tools from Austria to an archive server in Spain.

Both countries are welcome to submit a new proposal for candidate node in the future.

### New candidates nodes

#### CASPUR presented by Tiziana Castrignano

Established in 1992, the "Inter-University Consortium for the Application of Super-Computing for Universities and Research – CASPUR" is a non-profit organization, financed by the Ministry of Education, universities and research organisations and by associated universities.

The main scientific activity of CASPUR is to exploit the most advanced computing technologies to accelerate research in scientific computing.

- To manage a centre capable of guarantee a high quality and high-powered processing service. Priority is given to associated universities and MIUR, but the centre is open to the whole national scientific community, with particular emphasis on Southern and Central Italian research Institutes;
- to promote the use of the most advanced information processing systems and to support public and private scientific and technological research;
- to be a centre of excellence available to the national Universities, research networks and MIUR, with the aim of spreading the culture of information and communication technology, along with promoting their applications;
- to develop research programmes aiming at more effective and innovative usage of information and communication technology,



in collaboration with other organizations and enterprises.

The "Computational Chemistry and Biology Research Group" of CASPUR involves more than 20 staff members with expertise in computer science, software development, computational chemistry and biology and project management to deliver solid solutions for researchers across Italy and Europe.

In the bioinformatics research area there are agreements with several Institutions located in Central and Southern Italy to support projects in the following areas:

- comparative genomics
- alternative splicing
- biorepository
- microarrays
- structural biology
- systems biology
- advanced technologies for bioinformatics
- chemical properties of macromolecules



Tiziana Castrignanò presenting the CASPUR.

Since 2001 CASPUR has been collaborating with the Institute for Biomedical Technology (the Italian National EMBnet Node) on comparative genomics, prediction of alternative splicing sites, microarray data analysis and bioinformatics web services.

Several agreements have been signed with various institutions in biomedical research.

CASPUR organizes regular courses and a winter school in collaboration with the CNR Institute for Biomedical Technology (Italian National EMBnet Node).

#### **Candidate for national node: France presented by Guy Perrière**

Historically France was represented in EMBnet by Infobiogen a central computing centre provid-

ing access to sequence and tools repository. Infobiogen was stopped in 2005, just at the time when ReNaBi was created.

ReNaBi (Réseau National de Bioinformatique) is the French National Network of Bioinformatics platforms. Presently it gathers resources from 13 different platforms. Very similar to EMBnet, only platforms approved by ReNaBi coordination committee can participate.

Altogether ReNaBi provides access to

- "classical" and advanced software tools;
- sequence (and structure) databases;
- new powerful means of calculation like clusters and grids;
- teaching and training documents;
- <http://www.renabi.fr>.

Financial support for ReNaBi is given by the National Center for Scientific Research (CNRS) funding groups:

- CNRG (National Center for Research in Genomics) up to 2007.
- IBISA (Infrastructures in Biology, Health and Agronomy) since 2007.

In addition each platform is funded individually by various resources.

ReNaBi actions:

- Financial support of scientific projects proposed by the member platforms:
  - 385 k€ for four projects in 2007.
- Financial support of the French national bioinformatics conference (JOBIM).
- Production and distribution of teaching documents:
  - Courses given in the ReNaBi framework.
  - Manuals and tutorials.

Coordination committee:

- One coordinator (elected for two years, non renewable):
  - Presently Claudine Médigue from the Génomoscope (Paris).
- Steering committee with two other persons.
- Two members for each platform.
- Participation of the president of the French Society for Bioinformatics (SFBI):
  - Presently Guy Perrière, from the PRABI platform.
- One annual general meeting.

#### **Expressions of Interest for membership**

Uzbekistan, Azerbaijan, Tajikistan

Shahid Chohan (Pakistan) presented the EMBnet organization in a bioinformatics workshop



Guy Perriere presenting the ReNaBi.

organized in Baku (Azerbaijan). Azerbaijan is almost ready to submit an application. They have verbal approval by their Ministry of Science.

### Report of the committees

#### EB

We moved our official address from Belgium to Sweden. The secretary filled the official forms in Dutch. The EB decided that the official address should follow the chairman:

The EMBnet Stichting  
 Prof. Erik Bongcam-Rudloff  
 Husargatan 3 BMC building  
 Uppsala  
 Sweden



The EMBnet EB at work!

#### VGM report

We held 11 monthly VGMs since last AGM, the usual schedule is the 2nd Tuesday of the month at 4pm CET. The number of participants can vary from 8 to more than 20. VGM minutes are sent by email to the admin list and deposited on the web site in PDF usually within one week after the meet-

ing. On the web site (using your private login):  
<http://www.embnet.org/VGM-reports>

#### ETPC

##### *Activities and Proposals*

The E&T PC maintained contact via several electronic meetings using Marratech.

##### *Changes in the members*

At the AGM 2007, Greece has been formally welcomed to EMBnet and Sofia Kossida, the Greek node manager, joined the E&T PC. Georgina Moulton from UMBER also joined the E&T PC after having expressed interest.

In November 2007, former secretary and chairman of the E&T PC and Belgian node manager, Valérie Ledent stepped down from EMBnet to start a new career. The E&T PC (and EMBnet) will miss her very much. Thank you, Valérie and good luck!

Matej Stano joined the team early 2008 to replace Valérie.

Some contacts have been made with Sonia Cattley, the Australian EMBnet node manager. She may officially join the E&T PC during this AGM.

##### *Quick Guides*

Sofia supervised the creation of two new Quick Guides (QG)

- A Quick Guide to MATLAB Bioinformatics Toolbox;
- A Quick Guide to XCEDE

These QG target commercial tools and the permission to publish them on our website was asked several times with no answer. In a last demand, it was mentioned that the QG would be published very soon unless told not to. Due to the lack of answer, the QG are currently available on the Greek EMBnet website for the moment:

[http://www.bioacademy.gr/bioinformatics/GrEMBnet.htm#QUICK\\_GUIDES](http://www.bioacademy.gr/bioinformatics/GrEMBnet.htm#QUICK_GUIDES)

It was decided that QG should not target commercial tools without the prior agreement of the companies. In any case, preference is given to free tools.

The E&T PC decided to set up a poll to let users of the e-learning website indicate their preference for the next QG to be produced, see:

<http://elearning.embnet.org/mod/choice/view.php?id=2820>

### Courses

Several courses were added to the e-learning website in the archive section, mainly by JR see: <http://elearning.embnet.org/course/index.php>

However, the first course distributed by the E&T PC was added to the active course section! It is entitled "UNIX Fundamentals", see:

<http://elearning.embnet.org/course/view.php?id=53>

This course was edited and designed by Valérie and Vassilios. Matej, Sofia and Jingchu reviewed it and proposed corrections before it was published.

Very few EMBnet members uploaded their courses to the e-learning website. Despite the proposal of the E&T PC's to help for the upload, nobody asked for help. The E&T PC decided to create a page with links to "Other Bioinformatics courses" from the EMBnet community:

<http://elearning.embnet.org/mod/resource/view.php?id=2821>

### Future plans

- Creation of a public email address so that people outside EMBnet can use a single email address to contact the E&T PC.
- Depending on the interest of EMBnet members, the E&T PC will organize virtual EMBnet lab meetings (Webinars) taking advantage of the Marratech system. These meetings will consist of short presentations about 20-30 minutes and the topics may vary depending on suggestions and needs. Members are invited to propose presentations.
- Organization of live sessions where worldwide participants can interact with teachers after they have followed the UNIX fundamentals tutorial on the e-learning website.
- More QG. Not only depending on the results of the poll, but also from spontaneous proposals such a QG on Glite (EGEE Grid) by JR.

### P&PR PC

The main tasks were accomplished last year:

- production of 4 issues of EMBnet.news, published on [embnet.org](http://embnet.org) in 2 PDF resolutions and in an interactive viewer called "issuu". (see: [www.issuu.com](http://www.issuu.com)); the Committee is working on a guideline for authors, but it will be added next year. The future of EMBnet.news was discussed during the AGM;

- management of [embnet.org](http://embnet.org) in collaboration with the TMPC, very few changes, only related to the special pages for the 20th Anniversary Conference;
- EMBnet is being added on Google Maps and Wikipedia;
- the Committee is working on an EMBnet booklet, with eventually an electronic version.

### TMPC

EMBnet's DNS-servers are located in Sweden (primary) with secondary servers in Italy, The Netherlands and Norway. EMBnet's DNS entries are managed by the TMPC. Emil Lundberg, local network administrator for the Swedish node, was added to the TMPC mail-list to reduce the risk for delays in managing the DNS-entries.

All EMBnet mail-lists (emb-adm, emb-eb, emb-et, emb-pr and emb-tech) are based on a LISTSERV system. They are all managed by the TMPC.

EMBnet owns a Marratech based video conference system with two meeting rooms that are free to use for EMBnet related activities such as the VGMs and programme committee meetings.

EMBnet's webserver is an Apple Mac Pro, 2 GB, 2x500 GB, located at the Swedish node. A number of base support tools (MySQL, PostgreSQL, Apache-2, PHP, IMAP/POP, BerkeleyDB...) were added and are maintained by the TMPC. We thus have a solid foundation for additional installations/extra needs.

Last year the TMPC upgraded the old EMBnet web site from a part-static/part-dynamic site into a full dynamic web site, based on Drupal, which is an open source Content Management System (CMS). Drupal is written in PHP, and runs on an Apache web server, and in our case it is working with a MySQL RDBMS. Later César installed a statistics compiling program called awstats on the webserver. The results are presented at: <http://www.embnet.org/stats/awstats.pl?config=embnet>

César is doing a monthly backup of the Drupal Database and a "sync" of the Drupal installation and files used within Drupal (excluding e-learning files). This backup is performed in one of his servers in Mexico. DVD-backups are also done locally at the webserver. Further improvements of the backup situation seem desirable and are being discussed within the TMPC.

Our site offers the ability for all the members to easily log in and be able to add/edit information from all over the site.

### Some facts for the site:

An anonymous user can post Activities (Courses, Meetings and Workshops) in a moderated way, which means the site administrator should "allow or disallow" its publication. An anonymous user can post News, as well as comments for some of the contents of the site, making it a more "live" site.

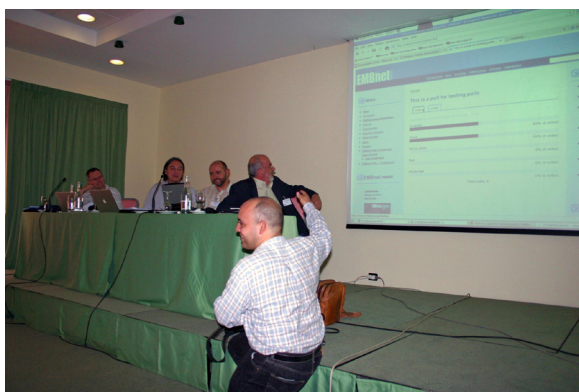
Every registered EMBnet member (Node manager or staff) is able to "allow or disallow" the publication of any "moderated content" (content waiting for publication).

The user registration and uploading of new contents is a matter of "clicks". There is a "shared" repository for images and files used on the site, to make it easy to manage it.

In the near future, visitors to the site will be able to create an account with which, among other things, they will be allowed to rate contents of the site --kind of "five-star-rating"--. Right now the only contents to be rated are those of type "weblinks" (which will become the most "live" part of the site).

The categorizing capabilities of Drupal are something to be exploited, since with the use of "taxonomies" all the contents of a site can be classified and hence can be ordered/viewed/searched in different ways and sorts.

The TMPC team created discussion forums for all the new technologies EMBnet is starting to use: for the EMBnet site (Drupal), for the eLearning site (Moodle) and for the Project Management tool (dotProject).



Cesar Martinez-Bonavides!!!!!! Et voilà.....part of the TM PC report!

In the old site there used to be a separate database of Node Managers. The new site is no exception and has this particular functionality directly embedded into the Content Management System; To accomplish this, the TMPC created a specific "content type" called "EMBnet node info" (using the CCK Module) in which all the fields that are relevant to information about Nodes are declared. Drupal manages it by physically creating a table with fields, and it is populated by records that are generated when a user creates a content of type "EMBnet node info". In that way we can still query the database to get all the names of the EMBnet Node managers, or the staff members with institutes and so on (see <http://www.embnet.org/?q=view/EMBnetNodes>) --in fact whichever kind of query-- this is something that users should have to explore/explode, by creating particular views (or queries) to get specific information.

Drupal is a very capable CMS, and options (modules, themes, configurations, etc.) are limitless, so if you find or hear about a module that could be useful for the EMBnet community, please do not hesitate to contact the TMPC to have a look at it and have it installed on the site.

In an emergency event that occurred in December 2007, the Belgian EMBnet node lost all communications with the Internet. The TMPC assisted the Belgian EMBnet node by taking over its vital domains and posting relevant status messages to its user base, until all communications were restored by the Belgian system administrators.

The TMPC participates at practically all VGMs for information sharing.



Lubos Klucar reporting about P&PR PC 2008 activities.

## Other reports

### Node report

Endre Barta from Hungarian node announced that the Hungarian node was closed in August 2008 in Budapest due to a change in the former Institute policy, but fortunately he could find another position in a different academic institution at Gödöllő where he revived the node. The EMBnet community is happy that this issue was rapidly solved and that the Hungarian node is continuing its activities.

### EU-FP7 grant proposal task force

Unfortunately even if the call ITN-PEOPLE seemed to fit well with our organization (non-EU countries allowed, no defined theme, money for salaries, travel and management, mandatory move from one country to the other, etc...), only a few participants did show some interest. Our idea was to create an EMBnet PhD School in Bioinformatics. One of the missing points was a clear common project, we evaluated several possibilities and choose to select a sister project in Health or Medical Informatics. The obvious choice was the EuroKup project where Erik and Terri are partners. Alas we learned during the writing phase that another group (all of EuroKup) was putting up a similar proposal. We decided to stop our project since it did not make sense to compete with them. We would like to thank the members of our task force: Sophia, Terri, JR, Andreas, Georgina and Erik.

### COST proposal twice rejected

Unfortunately both our COST proposals have were by 2 reviewers and rejected by a single reviewer and missed the cut for a fraction of a point. The main criticism being both times the lack of a clear project. Erik commented that Bioinformatics alone is not appealing enough. We should focus on a problem and help with bioinformatics. Thanks to all the participants of the task forces for these proposals: Erik, JR, Terri, Domenica, Nils-Einar, Andreas, Lubos, Federico, Luciano, Cesar, Laurent, Irena and Sofia.

### Pakistan scholarships

The ministry of Science in Pakistan will fund 50 scholarships to study overseas. Erik helped to evaluate applications and interviewed the students. 16 will study in Sweden. There will be more grants in the future. This is a good example of "pairwise" EMBnet collaboration.

## EMBRACE

Invitation from EMBRACE to the EMBnet community to organise a workshop on how EMBnet can provide EMBRACE compliant Web Services. The workshop was organised by Erik in Uppsala in November 2008. A report on the workshop is in the present issue.

## Future directions

Proposal summary after a long discussion:

- Starter kit for EMBnet node (Erik)
- E-learning (JR)
- Distributed database and storage e.g. MRS federation (George)
- Funding opportunities (FP7 group, Laurent will gather information)
- EMBnet.news (peer-reviewed or not?) to be discussed during next VGM

## Elections

### EB

Oscar Grau (AR) and Laurent Falquet (CH) were at the end of a three years mandate and up for re-election.

Both were re-elected, the EB is composed of Erik Bongcam-Rudloff, Oscar Grau, Laurent Falquet and Jose Ramon Valverde.

### ETPC

No member was up for re-election, however several changes happened during the year. Georgina Moulton (UMBER) stepped down for personal reasons. She will be welcome to join again the committee. Matej Stano (SK) was appointed during the year to replace Georgina, but he needed to be formally elected. Sonia Cattley (AU) stepped forward to join the committee.

Both candidates were elected, the committee is composed of Vassilios Ioannidis, Sophia Kossida, Matej Stano, Sonia Cattley and Jingchu Luo.

### P&PRPC

Kimmo Mattila (FI) and Lubos Klucar (SK) were at the end of a three years mandate and up for re-election. Kimmo wanted to step down. Andreas Gisel(IT) stepped forward to replace Kimmo.

Both were elected, the committee is composed of Pedro Fernandes, Domenica D'Elia, Lubos Klucar and Andreas Gisel.

### TMPC

One position was open since David Coornaert (BE) was leaving the Belgian EMBnet node



The 2008 Ethnic party: just an official snapshot!!!!

and thus was stepping down from his position. George Magklaras (NO) was at the end of a three years mandate and up for re-election. Emil Lundberg(SE) was proposed by Nils-Einar. Guy Bottu(BE) was proposed by Laurent.

All 3 were elected, the committee is composed of Nils-Einar Erikson, Cesar Bonavides, George Magklaras, Guy Bottu and Emil Lundberg.

### Next meetings

#### VGMs

The monthly schedule of "Virtual General Meetings" (VGM) using the Marratech e-conferencing software proved adequate over the last year.

Then we will continue with the usual schedule (every second Tuesday of the month at 4pm CET).

#### AGM2009

2 candidates emerged: Mexico and Portugal

Both candidate were asked to bring a detailed proposal with budget for the regular VGM of October 14, decision will be taken at the November VGM<sup>1</sup>. The details should be sent one week in advance to the EB.

#### Conclusion

The chairman thanks everybody for the good work and closes the meeting at 19:45.

Participants dispersed for the Ethnic party at 21:00 in the Park Hotel San Michele.

Report by the secretary of the EB, Laurent Falquet

Pictures provided by "paparazzi": Cesar, Domenica and Erik (we are sorry for missing photos of Vassilios, Nils and George during their committee reports but the main paparazzo, Cesar Martinez-Bonavides, was distracted during their presentation!)

*EMBnet is grateful to sponsors and supporters for their kind contribution to the success of 2008 EMBnet Conference!*

#### Platinum Sponsors:



#### Gold Sponsors:



#### Silver Sponsors:



#### Student travel fellowships:



#### Supporters Sponsors:



<sup>1</sup> Editors comment: was changed to the February VGM on November VGM.

# Tutorial on Grid Computing

Satellite event of the EMBnet Conference 2008



**Giorgio P. Maggi**<sup>1</sup>,  
**Domenica D'Elia**<sup>2</sup>,  
**Andreas Gisel**<sup>2</sup>,  
**Giacinto Donvito**<sup>1</sup>,  
**Guido Cuscela**<sup>1</sup>,  
**Giuseppe La Rocca**<sup>3</sup>

<sup>1</sup> INFN, Sezione di Bari, Bari (IT)

<sup>2</sup> CNR, Institute for Biomedical Technologies, Bari (IT)

<sup>3</sup> INFN, Sezione di Catania, Catania (IT)

Recent successes in solving computer and data intensive bioinformatics analysis by using the GRID technologies were reached in the framework of the Bioinfogrid European project ([www.bioinfogrid.eu](http://www.bioinfogrid.eu)) and the LIBI Italian FIRB project ([www.libi.it](http://www.libi.it)). These successes were strong motivations to disseminate this knowledge by organizing a tutorial on GRID computing for life science applications as a satellite event of the EMBnet Conference 2008 on September 17th.

The tutorial was organized thanks to a joint effort of EMBnet and the LIBI. It was aimed at research students, post-docs, and senior researchers with an interest in using or developing applications for distributed computing environments.

Efficient GRID computing is nowadays possible thanks to the Enabling Grids for E-science (EGEE) project, which has assembled the largest multi-disciplinary GRID infrastructure in the world. It brings together more than 140 institutions to produce a reliable and scalable computing resource available to the European and global research community. At present, the EGEE GRID infrastructure consists of approximately 300 sites in 50 countries and gives access to 80,000 CPU cores around-the-clock to about 10,000 users.

The EGEE Grid infrastructure, with its huge amount of storage and CPU power, has proved to be very effective in handling situations where a complex bioinformatics analysis can be subdivided in a number of independent elementary tasks. The number of independent tasks can be quite large and it may require hundreds of thousands of jobs to be submitted. Several bioinformatics applications are of this type, especially when large scale or genome-wide analysis are required. By assigning each elementary task to a Worker Node (WN), one of the 80,000 CPU cores available, many tasks can be executed in parallel reducing significantly the time needed to reach the solution.

The tutorial was opened by Domenica D'Elia, Italian EMBnet Node Manager, and chaired by Prof. Giorgio Maggi of the National Institute of Nuclear Physics (INFN) and Politecnico di Bari (IT).

Josè R. Valverde of the Centro Nacional de Biotecnología (CSIC) in Madrid (E), made the introductory talk describing the pros and cons about GRID computing from the point of view of a bioinformatician. After this presentation the tutorial entered in the core of its content with presentations, hands-on and demos on some of the tools recently developed inside the LIBI project.

Particular emphasis was given to the GRID Problem Solving Environment developed and set up for the LIBI project, the bioinformatics grid portals enabled with robot certificates, the LIBI federated databases approach and the tools used for accessing it from a GRID environment. Finally the AntiHunter and the Job Submission Tool (JST) were presented. The JST utility was demonstrated in several use cases. The tool is particularly useful to manage submission of a large number of jobs to the EGEE Grid infrastructure, their monitoring and bookkeeping. Once initialized for a given application, JST takes care of executing on the grid all of the elementary tasks, eventually resubmitting the failed ones, hiding to the user the complexity of operating in a grid environment.

The list of the presentations given during the tutorial together with all the material used can be found in the agenda page at the following link: <https://agenda.cnaf.infn.it/conferenceDisplay.py?confId=164>.

More than 30 people and 12 teachers attended the tutorial, which was particularly useful to spot bioinformatics problems where GRID can



The tutorial participants and teachers at work (Park Hotel San Michele, September 17, 2008).

give a real added value. Another important result was that teachers were glad to contribute to this EMBnet.news issue with their articles on tools presented during the tutorial thus allowing to extend its benefit to the whole EMBnet's users community. In parallel, during the three days of the EMBnet conference, demos of the tools object of the tutorial were provided by Guido Cuscela to any interested participant of the conference.

A further important contribution to the EMBnet Conference 2008 related to GRID Computing for bioinformatics applications came from the EGEE and EMBRACE partner in France, the research group led by Vincent Breton. Jean Salzemann and Vincent Bloch from the CNRS, Laboratoire de Physique Corpusculaire, Clermont-Ferrand (FR) held permanent demos on the bioinformatics platform integrating bioinformatics tools and databases they contribute to develop on the EGEE GRID infrastructure. Demonstrations showed how the platform works by submission of simple jobs through a web portal which automates the submission of jobs, their monitoring and retrieval of results. These demonstrations aimed to show how the platform mechanisms can handle large-

scale deployments or punctual submissions of short tasks. Moreover, they also showed some examples of bioinformatics workflow executions on their platform through high-level workflow management tools like Taverna.



## Authenticated Grid access with robot certificate and the GENIUS Grid Portal



**Giacinto DONVITO**<sup>1</sup>



**Giuseppe LA ROCCA**<sup>2</sup>

<sup>1</sup> INFN, Sezione di Bari, Bari (IT)

<sup>2</sup> INFN, Sezione di Catania, Catania (IT)

### Introduction

In the last few years, the scenario of international collaborations in Research and beyond has swiftly evolved with the gradual but impressive deployment of large bandwidth networks. A number of advanced services and applications have been using these networks, enabling new ways of remote collaboration. The environment resulting from the integration of networking and other resources, such as computing, storage, instruments and related systems is also known as "e-Infrastructure" (Fig. 1). This term, which is mainly used in the research and development context, is used to identify new generation of integrated ICT-based infrastructures. E-Infrastructures exploit several separate components and layers, such as networks, supercomputers and other computing resources, storage, remote resources and instrumentation i.e. sensors.

The use of e-Infrastructures is rapidly changing the landscape of science. Remote access to computing services, instrumentation and resources in general, creates new opportunities for researchers to bring existing applications to higher levels of usability and performance. Additionally, it enables researchers to deploy new strategies in approaching scientific problems with simulation tools and intensive applications. Another benefit of e-Infrastructures is that they stimulate the creation of new scientific communities, joining researchers who are working on similar challenges and are willing to share resources and reach new levels of collaboration. Researchers can gain access to scientific data and unique instruments located in top level laboratories around the world

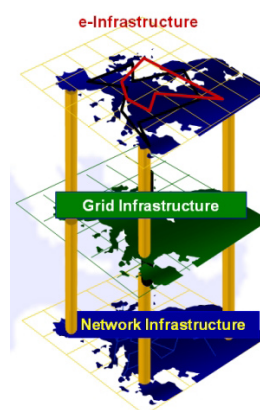


Figure 1. The European e-Infrastructure.

without the need to travel, speeding up the creation of results of scientific relevance and opening new unthinkable scenarios up to few years ago. This is why these modern ICT-based infrastructure are going to be spread to different scientific domains such as : Astronomy, Computational Chemistry, Earth Science, Financial Simulation, High Energy Physics and Biomedicine as well. An important component of a e-Infrastructure is represented by the Grid infrastructure, a revolutionary distributed environment for sharing heterogeneous computing resources and mass storage systems distributed world wide and interconnected by large bandwidth networks.

Unfortunately up to now, the basic know-how requested to access these modern e-Infrastructures is not so trivial, especially for not ITC expert users. There is a sort of "scientific gap", represented by complex computing protocols and rather complex Command Line Interface, that modern scientists have to overcome before to start to exploit all the advantages introduced by these e-Infrastructures. Moreover, the high security policy requested to access the distributed computing resources is a rather big limiting factor to increase the usage of Grids to a wider community of users. Grid security is indeed based on the public key infrastructure of X.509 certificates (each user has to visit one of the Certification Authorities and apply for a personal certificate) and the procedure to get and manage those certificates is unfortunately not straightforward.

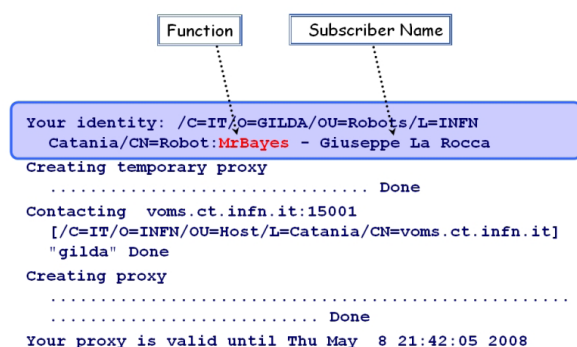
### The GENIUS Grid Portal and the robot certificates

Using the Web technology and its new recent developments, Grids details can be hidden to the

end users, giving access to the infrastructures in a very easy way as such as the common sense in web usage. This can be done with a Computing Web Portal. GENIUS Grid portal [1-3], powered by EnginFrame [4], is an increasingly popular mechanism for creating customizable, Web-based interfaces to Grid services and resources.

In order to improve the exploitation of Grid technologies and foster the adoption of this new paradigm in new community of users the EnginFrame Java framework, on which GENIUS is built, has recently been extended in order to support robot certificates. Basically robot certificates can be used to identify a person responsible for an unattended service or process acting as client and/or server. These certificates may be used to authenticate the service to another Grid entity, possibly by signing proxy certificates. The fundamental characteristic of a robot certificate is that each one is related to a specific application shared by all members of a given Virtual Organization. A generic robot certificate includes the full name of the subscriber as determined by the Virtual Organization/Site's Registration Authority as well as a label reflecting the purpose of the robot entity, i.e. the application that can be executed with the certificate.

Certificates must apply to unique individuals. Private keys associated with Robot certificates may not be shared between people. These new certificates can be installed on a smart card (e.g. Aladdin eToken PRO 32K [5]) and used behind the portal by everyone interested in running a portal specific application in a Grid environment using a user-friendly graphic interface. From now on when the smart card is inserted in the server where GENIUS is running, the portal will start generating a new user's proxy signed by the robot certificate, otherwise the normal authentication



```

Function
Subscriber Name

Your identity: /C=IT/O=GILDA/OU=Robots/L=INFN
Catania/CN=Robot:MrBayes - Giuseppe La Rocca
Creating temporary proxy
..... Done
Contacting voms.ct.infn.it:15001
[/C=IT/O=INFN/OU=Host/L=Catania/CN=voms.ct.infn.it]
"gilda" Done
Creating proxy
..... Done
Your proxy is valid until Thu May 8 21:42:05 2008
  
```

Figure 2. A glance at robot certificate.

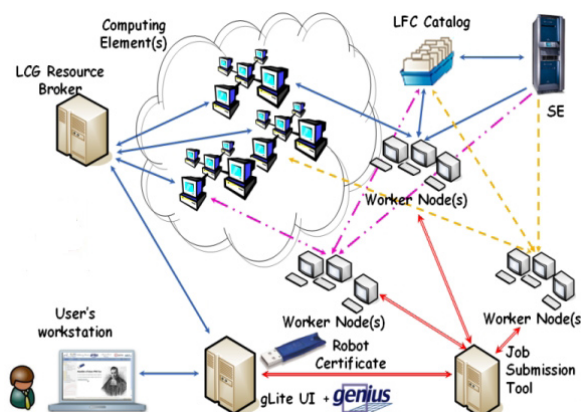


Figure 3. MRBAYES: A simple bioinformatics use case.

based on a dedicated Java applet will be performed. Once the proxy is generated the user is automatically redirected to the home page of the application related with the certificate. Any other attempt to access to unauthorized applications will be blocked by the portal. Moreover, in order to enhance the security of the system a User Tracking System has also been introduced to register and monitor the most relevant actions performed by users.

### An use case : Bayesian Phylogenetic Inference on a large scale.

The GENIUS Grid portal with a transparent support to the robot certificate has been successfully used by non-grid users, involved in the context of the LIBI Italian Laboratory for Bioinformatics to run a bioinformatics application on a Grid Infrastructure. In this section some details about the application's workflow which has been set up in order to run this application on the Grid Infrastructure is shown.

The client side is represented by a user's workstation running a web browser. The server side is represented by a gLite User Interface machine, equipped with the middleware services to submit jobs and manage data on Grid, the Apache Web Server, the Java/XML portal framework EnginFrame developed by NICE Srl and the GENIUS Grid Portal itself. After user's login, a proxy certificate is requested by the portal to access the distributed resources of a Grid Infrastructure according to the Globus Security Infrastructure standard. If no user's proxy is available and an Aladdin eToken PRO 32Kb with on board a robot certificate is plugged into the server, it will be used by the portal to generate the necessary

proxy. This operation is completely transparent for the end-user. In a few minutes the robot certificate stored on the USB token is read to generate the proxy certificate. Once the proxy certificate has been successfully created, the user is automatically redirected to the home page of the application related to the robot certificate. Thanks to the services developed with the portal user can provide input settings for the application before to submit its parallel version to the Grid Infrastructure.

Moreover, in order to improve the reliability of the workflow and dealt with possible job failures a Job Submission Tool (JST) [6], developed by INFN Bari, has been also introduced in the architecture. This tool has been adopted for the submission of large number of jobs in an almost unattended way. It is based on the concept of "task" to be executed. The entire problem is first subdivided into elementary tasks; then all the tasks are inserted into a DB server. In the submission phase all the jobs are identical, in fact when the job is submitted it does not know which task has to execute. Only when the jobs land and start executing on a WN, it requests to the central DB a task to execute. Information on the execution of each task is logged in the central DB. Only if all steps are correctly executed by the job, the status of that particular task on the central DB is updated to "Done". In this way the central DB provides a monitoring of the task execution and no manual intervention is required to manage the resubmission of the failed tasks: tasks which are found in a "running" state after a given time interval, are considered failed and automatically reassigned to new jobs. Figure 4 shows the monitoring and the visualization system developed for the bioinformatics application.

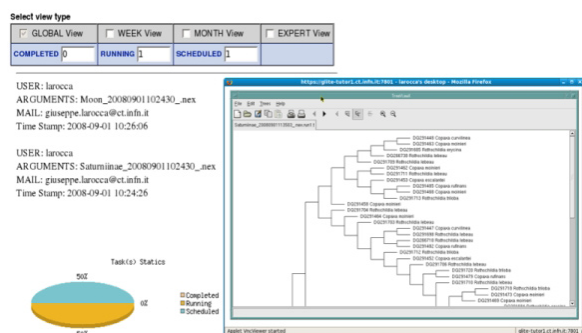


Figure 4. Monitoring and Visualization services.

## Conclusions

The valuable benefits introduced by robot certificates can really contribute to make Grid technology more appealing, providing an asset in raising Grid awareness to a larger number of potential users.

## Acknowledgements

This work was supported in part by the MUR FIRB Italian projects LIBI (Italian Laboratory for Bioinformatics) and by the European Specific Support Action BIOINFOGRID (contract number: 026808).

Web Site: <https://glite-tutor1.ct.infn.it>  
Information: giuseppe.larocca@ct.infn.it,  
giacinto.donvito@ba.infn.it

External material:

[https://gilda.ct.infn.it/Bari/LAROCCA\\_MrBayes\\_AVI.avi](https://gilda.ct.infn.it/Bari/LAROCCA_MrBayes_AVI.avi)  
[https://gilda.ct.infn.it/Bari/LAROCCA\\_MrBayes\\_MOV.mov](https://gilda.ct.infn.it/Bari/LAROCCA_MrBayes_MOV.mov)

## References

1. Andronico G, Barbera R, Falzone A, Lo Re G, Pulvirenti A, Rodolico A – GENIUS: a web portal for the grid. Nucl. Instrument and Methods in Phy. Res. A 2003. Visit also the official GENIUS web site: <https://genius.ct.infn.it/>
2. Barbera R, Falzone A, Ardizzone V, Scardaci D – The GENIUS Grid Portal : Its Architecture, Improvements of Features ,and New Implementations about Authentication and Authorization. Enabling Technologies: Infrastructure for Collaborative Enterprises, 2007. WETICE 2007. 16th IEEE International Workshops.
3. "The GENIUS Grid Portal and the robot certificates: a new tool for e-Science"- XII International Workshop on Advanced Computing and Analysis Techniques in Physics Research – ACAT 2008, Erice, 3-7 Nov. 2008 (see <http://indico.cern.ch/materialDisplay.py?contribId=217&materialId=poster&confId=34666>)
4. EnginFrame – <http://www.enginframe.com/>
5. Aladdin eToken PRO - <http://www.aladdin.com/etoken/devices/pro-usb.aspx>
6. JST – Job Submission tool - Giulia De Sario, Andreas Gisel, Angelica Tulipano, Giacinto Donvito, Giorgio Maggi, High-throughput GRID computing for Life Sciences, in Mario Cannataro (Ed.), Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine and Healthcare, IGI Global (to appear)

## The Grid Problem Solving Environment for Bioinformatics: the LIBI experience



**Maria Mirto<sup>1</sup>, Italo Epicoco<sup>1,2</sup>, Massimo Cafaro<sup>1,2</sup>, Daniele Tartarini<sup>1</sup>, Alessandro D'Anca<sup>1</sup>, Alessandro Negro<sup>2</sup>, Marco Passante<sup>1</sup>, Giovanni Aloisio<sup>1,2</sup>**

<sup>1</sup> University of Salento, Lecce & SPACI Consortium, (IT)

<sup>2</sup> Euro Mediterranean Centre for Climate Change, Lecce (IT)

### Introduction

Problem Solving Environments (PSEs) are a very useful tool which combines simulation and visualization into a single package [1, 2]. The consequential benefit of such a system is that it facilitates experimentation with minimal additional effort from the user.

The integration of Grid technology with PSEs is a natural step in this evolution [3].

Often, several biological applications need powerful computational resources for supporting large-scale experiments. Genomics and proteomics data are produced in huge quantities and analysing them also requires investigating their correlation.

The Grid [4] is the ideal environment to provide the computational resources needed for several studies and an integrated environment, a Grid PSE, is needed for hiding the underlying complexity, so that the users can benefit from user-friendly interfaces.

Finally, the idea of "workflow" [5] in Grid terminology correspond very closely with how PSEs are generally constructed within e-Science environments.

Indeed, workflows are needed for the simulation of complex experiments because they integrate several applications and data that often

are distributed and that need careful orchestration.

These were the main reasons for adopting the Grid PSE technology inside the LIBI virtual laboratory.

The goal of this project is setting up an advanced Bioinformatics and Computational Biology Laboratory, focusing on basic and applied research in modern Biology and Biotechnologies [6].

In the startup phase of the project (2005), all of the involved technological partners, which are the University of Salento, Lecce, the INFN sections of Bari, Padova, Catania and CNAF in Bologna, IBM Italy of Bari and Cineca in Bologna, have studied general and ad hoc solutions for resource and data management in order to embed them into the Grid PSE solution.

In the EMBnet Tutorial on Grid Computing we presented our technological solution for a Grid PSE composed by a Grid Portal that, hiding the complexity of the Grid, offers web interfaces for submitting and monitoring jobs. Through the web, a workflow editor is also offered for the composition of the available applications in the Grid. Finally, a Meta Scheduler embedding the functionalities of a workflow engine, is responsible for the management of gLite [7], Unicore [8] and Globus [9] services. gLite, which is the EGEE middleware, represents the best solution for parameter sweep jobs, that usually are independent elementary tasks. The EGEE Grid infrastructure [10], with its huge amount of storage and CPU power, has proved to be very effective in handling this kind of situation. Instead, the Unicore middleware, used in the DEISA Grid [11], represents a viable solution for submitting parallel jobs requiring a huge amount of CPU at a glance, through the CINECA nodes. Finally, Globus can be used in both cases, providing a unified solution for parameter sweep and parallel jobs.

In the following Sections, after introducing a few details about a Grid Problem Solving Environment, the technological solution proposed by the LIBI project will be described along with several case studies.

### The Grid PSE meets the Bioinformatics requirements

Managing the deluge of biological data produced by large-scale experiments such as genome projects is one of the major global chal-

lenges of bioinformatics. Biological datasets present challenges from many different perspectives. The scale of data to be processed requires access to large collections of widely distributed resources such as those found in computational grids.

The analysis of these data needs to be farmed out to hugely provisioned computing resources with large and efficient storage devices. Indeed, bioinformaticians need for their daily analyses of genomes efficient access to these biological data and bioinformatics programs. Moreover, these datasets are not static: as discoveries are made, new entries are added to the database and existing ones are updated.

Complicating the problem of data management is the fact that most bioinformatics applications are not designed with grid computing in mind. Indeed, many valuable applications were designed, tested, and validated long before grid computing arose. As a result, such applications are designed to perform simple local Input/Output and have no facility for attaching to grid data systems. Nevertheless, these applications need both high-throughput computing and huge data storage [12, 13].

Should we simply rewrite such applications to take advantage of the grid? Although this might be possible for a small number of applications, re-writing would be an enormous amount of work to address all of the legacy bioinformatics codes in use today. Many are commercial codes for which source is not available. In some cases, the construction and validation of codes is tightly integrated into an audited scientific process; changing the code for grid deployment would invalidate the application, or at least require a new validation. To make grid computing easy, we must find a way to access data through familiar interfaces without changing the applications.

The Problem Solving Environment (PSE) is an approach and a technology that can fulfil such bioinformatics requirements. The PSE can be used for the definition and composition of complex applications, hiding programming and configuration details to the user that can therefore concentrate only on his/her specific problem.

PSEs have been investigated over the past years. Culler and Fried [1] initiated investigating automatic software systems for solving mathematical problems with computers focusing primarily on applications issues instead of program-

ming issues. At that time, the term "application" indicated scientific and engineering applications that were generally solved using mathematical solvers or scientific algorithms managing vectors and matrices.

Despite the time passed from that early research work, still there is not a precise definition of what a PSE is. The following well-known definition was given by Gallopoulos, Houstis, and Rice: "A PSE is a computer system that provides all the computational features necessary to solve a target class of problems. . . . PSEs use the language of the target class of problems." [2].

PSEs can benefit from advancements in hardware/software solutions achieved in parallel and distributed systems and tools. One of the most interesting models in the area of distributed computing, is the Grid paradigm.

Grid computing [4] represents an opportunity for PSE designers and users. It can provide an high-performance infrastructure for running PSEs and, at the same time, a valuable source of resources that can be integrated in PSEs. Grids can be used for building geographically distributed collaborative problem solving environments, and Grid-aware PSEs (GPSE) can search and use dispersed high performance computing, networking, and data resources.

Grid-aware Problem Solving Environment [3] can offer a solution for handling and analysing so much disparate data connecting many computers within and among institutions through middleware software.

Interaction between bioinformatics tools and biological databases should be simplified and each component (i.e. biological data banks, experimental data, bioinformatics tools) should be seen as an atomic service (Web Service) [14, 15] that can be easily integrated in different systems, through standard interfaces and protocols. In the next section, we present the Grid PSE solution for the LIBI virtual laboratory.

### Proposed technological solution

The architecture of the system (Figure 1) is Service-Oriented (SO) [16], and it is built on a network infrastructure for the exchange of information.

On the top layer there are the application services that can run separately or can be inserted into a graph representing a workflow.

Several applications commonly used in the Laboratory are:

- *PSI-BLAST* (Position-Specific Iterated Blast) [17]: it is a sequence multi alignment;
- *PatSearch* [18]: it searches sequence patterns;
- *Gromacs* [19]: it is a molecular dynamics software;
- *AntiHunter* [20]: it is a tool for the identification of expressed sequence tag (EST) antisense transcripts from BLAST output;
- *MrBayes* [21]: it performs Bayesian inference of phylogeny.

The applications access the high level services namely *Resource Management and Data Management*. The *Resource Management* contains the execution logic of the applications and exploits a Meta Scheduler, based on the GRB technology [22], which implements the functionalities for job submission, job monitoring, file transfer and management of user's credentials.

The Meta Scheduler is built on top of basic services offered by different grid middleware such as gLite for parameter sweep applications, Unicore for parallel jobs and Globus for both types.

Additional functionalities provided by the Resource Management are the scheduling and monitoring of the user requests, the optimal allocation of resources in a distributed environment, etc.

The Resource Management involves also the Job Submission Tool (JST), a tool developed by the INFN in Bari, for parameter sweep submissions.

The *Data Management* embeds the grid logic for advanced data management, allowing a transparent and dynamic access to heterogeneous and geographically distributed data.

The LIBI Data Management provides bioinformatics data federation services for managing and accessing the LIBI Federated DBs. The LIBI federator server [23], developed by the IBM in Bari, performs the data federation task enabling federated databases to access several heterogeneous and distributed data sources such as UTRef, MitoRes, UTRSite, GenBank, OMIM, Pubmed, EMBL, HmtDB, Uniprot, EMBL\_CDS. It provides a unified data-management interface (both for query and insert statements) based on SQL.

The GRelC DAS [24] exposes the federated DB in grid making it available to all of the LIBI grid

users. It currently provides a uniform grid access interface to a wide range of relational (Mysql, Oracle, Postgresql, SQLite, Microsoft SQL Server, etc.) and non relational data sources (XML DB engines such as eXist, XIndice, etc.).

This grid service also includes secure mechanisms for data transfer on the Grid through advanced protocols to guarantee a high level of efficiency. The security layer concerns managing and accessing both services and resources within the LIBI distributed environment.

It provides mechanisms for authentication, authorization, data protection, etc.

In the following Section, we describe the architecture of the Meta Scheduler, which is the core component of the system.

## An overview on the architecture

The architecture of the Meta Scheduler is shown in Figure 2.

It is composed by:

1. an editor, that allows the composition of the applications and input data through an arbitrary graph that can contain cycles and saves the workflow in two XML files (abstract workflow), by using our own format. These files contain respectively the information about the graphical position of each component (view file) and the JSDL (Job Submission Description Language [10]) file with the information related to both control and data flow of the application component and input data. JSDL is an extensible XML specification from the Open Grid Forum for the description of simple or complex tasks to non-interactive computer execution systems. It has been extended at the University of Salento to support workflow jobs. The editor has been implemented as a signed Java applet so that the view file can be saved onto the local machine by the user whereas the JSDL file is saved on the web server from which the applet is downloaded, through the Application Manager service;
2. an Application Manager (AM), which is an intermediate service between the editor and the engine, allowing resource discovery (the hostname of the worker nodes, the logical name of the available applications etc). Moreover, it handles user account and application profile, checks data consistency and saves JSDL files;

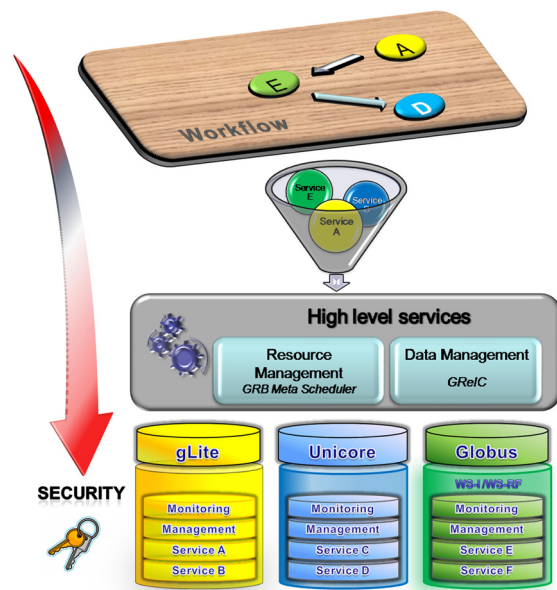


Figure 1. The LIBI Grid Problem Solving Environment Architecture.

- an engine (WF engine), based on a Meta Scheduler that completes the JSDL files with the references to actual executables and data files as well as specific resources (concrete workflow), chooses the opportune resource broker on the basis of the selected application in the workflow and translates the JSDL in the proper language depending on the specific middleware (JDL/DAG for gLite and AJO for Unicore). For the Globus Toolkit, the GRB scheduler takes as input JSDL files and translates them in the Resource Specification Language (RSL). Finally, for applications with Web Service interface the WSDL is parsed by the editor to import both available methods and types of the input data, whereas the engine parses the WSDL to create a stub and a service's client. The JSDL description can also involve data retrieval and access. The user can specify data sources, filters and data related operations useful for the computation. This kind of tasks are processed by the WF engine through the GReIC DAS.

As shown in Figure 2, the Meta Scheduler uses the GRBLib for the submission and monitoring of jobs on Globus grids by using the GRB scheduler. It implements a mechanism that taking as input a JSDL file translates it in RSL and automatically determines the grid resources needed for the submission.

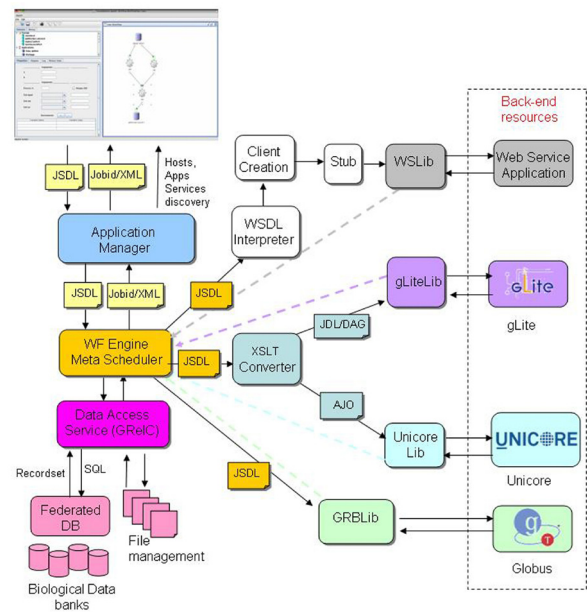


Figure 2. The Meta Scheduler Architecture.

It is based on Web Service methods, implemented in C using the gSOAP toolkit [33]. For job monitoring, a service is responsible for the retrieval of the overall workflow status and statuses of the workflow sub-tasks. It saves the status in an XML file which is returned to the editor, by means of the AM component. The editor parses the file and prints on the editor canvas the status.

Regarding the gLite middleware, a library implements the Web Service client of the gLite Workload Management System (WMS) for the submission of a batch or workflow job. The WMS is responsible for the distribution and management of tasks across the gLite Grid.

Moreover, this library also contains the clients for the Logging & Bookkeeping (L&B) Service which is related to job monitoring. Regarding data transfer, the GridFTP protocol is used. The TLS (Transport Layer Security) protocol is used to secure the communications between the client and the WMS, whereas the GSI (Grid Security Infrastructure) plug-in is used between the client and the L&B services.

To access the Unicore grid middleware a library has been developed. It includes several functions in Java for the submission and monitoring of a job. These functions include the client of the Unicore NJS (Network Job Supervisor) service. A wrapper in C, through the Java Native Interface (JNI) functions, has been added and will be in-

cluded in the Meta Scheduler (implemented in the C programming language).

The TLS protocol is used to provide security and the GridFTP protocol is used for data transfer.

Finally the WS Lib includes the clients of the Web Services that wrap the applications. The GSI plug-in is used to secure the communications.

## How to access the system

The system is available at the following url: <http://www.libi.it/libi/biotools/libi-biotools>. The user is invited to register using an opportune interface; once registered, the user is provided by the system with credentials in the form of login and password, to be sent to the remote server by means of TLS.

This represents a first level of authentication. Indeed, resources on Grids often need to be accessed with varying degree of security. Usually, a virtual organization (VO) has a subset of its resources, typically data on web pages, which is public; another subset could be accessed by people using a weak form of authentication mechanism, while for other resources access control will be strictly enforced based on the strongly verified identity and/or attributes of a requestor.

The Grid Security Infrastructure (GSI) [25] provides a mechanism for secure single sign-on and access control to resources. GSI is based on Public Key Infrastructure (PKI), with credentials issued by a Certificate Authority. GSI has proven itself to be a viable mechanism for resources that need strong access control, however it can be overly burdensome to access resources that may instead just require weak authentication of a requestor.

Before using the LIBI services, the user must acquire an end user certificate, issued by a Certification Authority. The LIBI project currently recognizes the following CAs: INFN (<http://security.fi.infn.it/CA/>) and SPACI. Users will be issued an end user certificate after being properly identified by a Registration Authority.

However, the end user certificate issued will never be used directly to contact remote services: instead, for security reasons, a proxy certificate derived from the end user one will be actually used.

Therefore, the user must create on his/her machine a proxy certificate by issuing the `voms-proxy-init` command and specifying the VO to be

used (in the LIBI project, the libi VO is used) as follows, entering his/her password protecting the certificate's private key when prompted to:

```
$> voms-proxy-init -voms=libi
```

the output is similar to:

```
Your identity: /C=IT/O=INFN/OU=Personal
Certificate/L=HPCC University of Lecce/
CN=Maria Mirto
Enter GRID pass phrase:
Creating temporary proxy
..... Done
Contacting voms.cnaf.infn.it:15015
[/C=IT/O=INFN/OU=Host/L=CNAF/CN=voms.
cnaf.infn.it] "libi" Done
Creating proxy
..... Done
Your proxy is valid until Sat Dec 20
07:26:32 2008
```

The next step is to create a proxy on the MyProxy server (e.g. the MyProxy server of the CNAF site) by using the `myproxy-init` command:

```
>$ myproxy-init -s myproxy.cnaf.infn.it
-l userlibi
Your identity: /C=IT/O=INFN/OU=Personal
Certificate/L=HPCC University of Lecce/
CN=Maria Mirto
Enter GRID pass phrase for this
identity:
Creating proxy..... Done
Proxy Verify OK
Your proxy is valid until: Fri Dec 26
19:29:22 2008
Enter MyProxy pass phrase:
Verifying password - Enter MyProxy pass
phrase:
A proxy valid for 168 hours (7.0 days)
for user uselibi now exists on myproxy.
cnaf.infn.it.
```

The user can specify a temporary pass phase and a logical name (in the example `userlibi`). A temporary password will be requested; it is worth noting here that this password is completely independent of the one protecting the certificate's private key.

By using the LIBI Grid Portal, the user will specify in a web interface the MyProxy server (`myproxy.cnaf.infn.it`), the user's logical name (`userlibi`) and chosen password. The system will then download the proxy from the specified MyProxy server, and will use it for submitting the jobs, on behalf of the users (Figure 3).



## Credential manager

The screenshot shows a web-based interface for adding user credentials. The main form is titled "Add User Credential" and contains the following elements:

- Proxy Server:** A dropdown menu currently showing "myproxy.cnaif.infn.it".
- Add a new server (hostname and port):** An input field.
- Login:** An input field.
- PEM:** An input field.
- Buttons:** "Ok" and "Reset" buttons.

Below the form is a table titled "User Credentials":

DN	Login	Expire	Delete
/C=IT/O=INFN/OU=Personal Certificate/L=HPCC University of Lecce/CN=Maria Mirto useribiti	Sat, 20-Dec-2008 07:45:01 GMT		

Figure 3. Interface to acquire a proxy from a MyProxy server.

Moreover, a "robot" certificate will be also provided for several services but this will automatically limit and enforce the use of a specific set of resources.

By using the Grid portal the user can access the previous application services; moreover, he/she can access the JST.

## Solving bioinformatics problems: two case studies

In this section, two case studies regarding Multiple Sequence Alignment (MSA) of several human proteins [26] and inference phylogenetic computing of several barcode sequences [27] are presented. The former requires a parameter sweep job, whereas the latter is handled as a parallel job.

### Multiple Sequence Alignment

A major result in the prediction of protein structures (secondary or tertiary) was obtained by adopting evolutionary information, usually in the form of protein profiles.

This is achieved by aligning to a query sequence all of the retrieved similar chains detected using a similarity search algorithm. Routinely, the most widely used program is PSI-BLAST, because of its speed and accuracy. In practice, any modern "state of the art" tool used to predict protein structures and features (such as secondary structures, membrane protein topology, protein solvent accessibility, protein-protein interactions protein stability changes etc.) takes as input some form of evolutionary information to achieve an accuracy compatible with real-world applications. However, similarity search and compilation of the sequence profiles are the most time-consuming steps for the prediction, but are necessary intermediate steps in the majority of prediction tools.

For this reason, a system that can speed up the similarity search step can be profitable both for accelerating the prediction phase and for testing more ideas to improve current state of the art methods.

In this optic an experiment related to the MSA of about seventy thousand human proteins against the data bank of UniProt NREF Uniref90 has been done.

- The data flow (see Figure 4) consists in the extraction of the sequences by several files, for each sequence a run of a MSA tool is carried out and then an optimization of the results is made.

The Involved tools are:

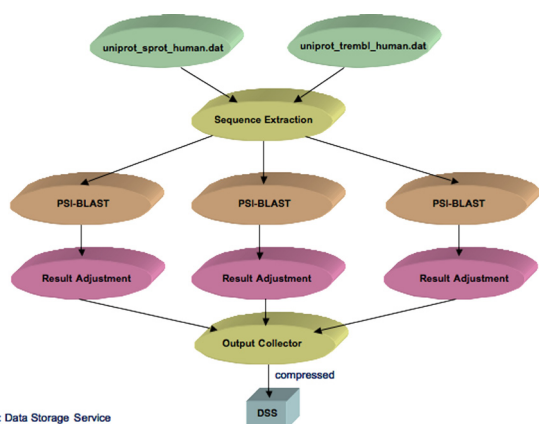
- a library for sequence extraction;
- PSI-Blast of the NCBI for multiple sequence alignment;
- a tool to optimize/adjust the results.

A library has been developed to extract the sequences by annotated input files, i.e. files containing both sequences and other information about species, organism, bibliographic references, etc., in EMBL format.

Indeed, dynamic libraries for accessing biological flat files are available within the library in order to simplify the access to flat files and to provide seamless access. Some features of this library are:

1. connection to flat files;
2. data manipulation;
3. information extraction;
4. printing the result in various formats such as Fasta and XML;
5. creation of an XML dump of flat files.

PSI-BLAST is a sensitive sequence similarity search tool that uses an iterative searching method and unique scoring scheme to detect weakly related homologues.



✓ DSS: Data Storage Service

Figure 4. Multiple Sequence Alignment Data Flow.

Finally, in order to reduce the redundancy of the results, taking into account that each result file contains several iterations and the last one is the most important, a module for reducing the dimension of output files has been developed. It parses the output files and deletes the results of intermediate iterations.

In order to support this experiment, several requirements have been met:

- access to a flat file data bank (UniProt NREF Uniref90) whose dimension is about 800MB;
- extraction of 70,845 sequences by annotated input files (human proteins - UniProtKB database);
- for each run, the application produces the result of several iterations, specified by the user (three iterations in this experiment); the last one is the most important, so that output files must be resized deleting the results of all of the iterations but the last one;
- management of produced results;
- need to reduce the total computing time.

In order to accelerate the access, the data bank has been locally installed on grid nodes, where the application is run, and thereafter indexed. Indeed, PSI-Blast only runs on indexed data banks.

The sequences are extracted by using the above cited library; parsing the results allows reducing the redundancy.

Regarding the management of produced results, taking into account that each output file has a dimension that ranges from 200 KB to 2 MB and that these results are on grid nodes, it is important to use efficient mechanisms to optimize the file transfer time. To this end, the GridFTP

protocol has been used so all of the produced files are retrieved on a storage grid node. Finally, in order to reduce the total computing time, dynamic scheduling algorithms have been used to allow load balancing in a distributed environment.

The experiment has been designed using the workflow editor (Figure 5).

By using this tool, the user can assemble the experiment through drag and drop of the available applications, available on the left window.

By clicking on each application it is possible to insert the required arguments, libraries, environment variables and so on.

The graphical representation is mapped to an XML file (Abstract Workflow) shown in Figure 6.

When the user submit the request to the Meta Scheduler, a JSDL file (concrete workflow) is produced. In this file a mapping from logical to physical names is made (Figure 7).

## Estimation of phylogenetic structure of barcode reference database for phylogenetic queries

We now describe a test case regarding the use of bayesian phylogenetic inference.

This experiment uses MrBayes software [21] to perform the inference, applied to the general problem of Barcode.

MrBayes is a program for Bayesian inference of evolutionary parameters and topologies from a set of nucleotide sequences. Evolution is generally studied only by inference: a pattern (in this case a set of biological sequences) is observed and different possible processes are evaluated to infer what process could have produced this pattern.

The inference regarding parameters' values (i.e. substitution matrix, topology, branch length) in the program is done within a Bayesian framework. The program uses a Metropolis-coupled Markov Chains Monte Carlo protocol (MCMCMC) [28, 29] as markovian integration to solve numerically the Bayesian formula and obtain a sample from the posterior distribution of the parameters. The posterior distribution is the probability of the parameters' values given the data and assuming that the true process is included in the models of evolution taken under consideration.

The MCMCMC implementation of the markovian integration allows MrBayes to be naturally

ready to work in a parallel environment. The availability of large number of CPUs allows dealing with biological problems that require a great number of chains to adequately explore the parameters space and arrive to convergence.

The test case for this kind of implementation was simply a series of large nucleotide sequences data sets that represent the nascent barcode reference database for Lepidoptera. DNA barcodes consist of a standardized short sequence of DNA (about 700 bp in the case of our *Lepidoptera* data sets) that in principle should be generated and characterized for all of the species on the planet. A massive on-line digital library of barcodes will serve as a standard to which the DNA barcode sequence of an unidentified sample from the forest, garden, or market can be matched [30].

The Consortium of Barcode of Life (CBOL; <http://barcoding.si.edu/>), established in 2004, is an international consortium that tries to establish standard (both for data production and data analysis) for automated molecular species diagnosis and is encouraging the production of DNA sequences relative to taxonomic lineages for which a standard locus for species recognition is identified (the barcode sequence) that follows a given criteria of quality.

The goal of our work is build a pipeline to infer the phylogenetic structure of several large data sets in order to facilitate a tool being developed, that would query a given barcode database, for species diagnosis of an unknown sequence. Given the large number of sequences that typically compose these databases, a general query system uses simplified phylogenetic inference

```
<?xml version="1.0" encoding="UTF-8"?>
<workflow diagram xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="/home/simone/workspace/Workflow/bin/Workflow/workflow.xsd">
<storage_nodes>
<storage x_position="330.0" y_position="53.0">
<hostname>sigma2.unile.it</hostname>
<ports>
<connector is_input="false">
<id_connector>1</id_connector>
<type>Simple (File)</type>
<filename>prova2</filename>
<dir_path />
<cond_operator />
<value />
</connector>
</ports>
<storage x_position="195.0" y_position="47.0">
<hostname>sara.unile.it</hostname>
<ports>
<connector is_input="false">
<id_connector>2</id_connector>
<type>Simple (File)</type>
<filename>prova1</filename>
<dir_path />
<cond_operator />
<value />
</connector>
</ports>
</storage>
</storage_nodes>
```

Figure 6. Abstract Workflow.

approaches such as overall similarity approach, neighbor-joining, or parsimony. The LIBI Grid PSE is a software platform allowing more realistic inferences that, cyclically updated, would inform the query system.

Moreover, we have implemented an enhanced service that wraps the parallel MrBayes application, providing added features and the execution in a Grid environment.

In particular, offered features are:

1. submission of single or multiple Nexus files by providing a compressed directory in various formats (gzip, tar, zip, etc.) or a single file from the user machine or even from a remote machine. Alternatively, submission of a FASTA file of nucleotide or amino acid aligned sequences, as input. In this case the commands and the parameters are collected through a dynamic web interface and, exploiting a python script, are wrapped into a Nexus file.
2. Checking each Nexus file in order to retrieve the values provided for two MrBayes parameters: *nchains* and *nruns*. The former is the number of chains used to explore parameters space, while the latter is the number of independent analysis to be performed. The product of these values is important for determining the maximum number of processes to use for the simulation. The parallelism level is indeed strictly related to the input files with an upper bound given by the *nchains* by *nruns* product. The GUI constrains the user to specify only the upper limit in terms of number of processes or its half, owing to the fact that the performance analysis carried out on several sequences [27] showed that if the number of CPUs is chosen

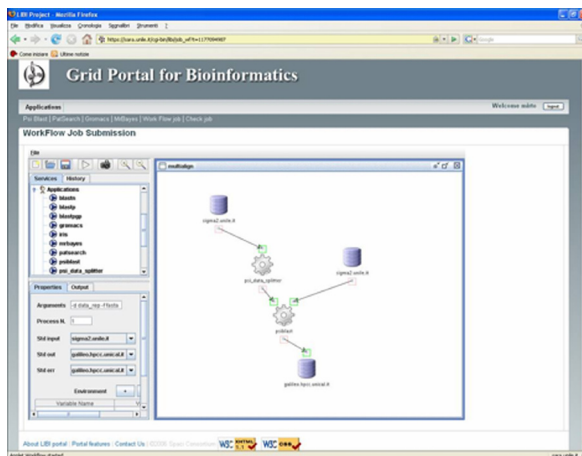


Figure 5. The LIBI Grid Portal: the Workflow Editor interface.

```

<?xml version="1.0" encoding="UTF-8"?>
<jds:JobDefinition xmlns:xsi="http://www.example.org" xmlns:jds="http://schemas.ggf.org/jds/2005/06/jds" xmlns:jds-
posix="http://schemas.ggf.org/jds/2005/06/jds-posix" xmlns:jds-extension="http://schemas.unile.it/jds/2005/06/jds-extension" xmlns:jds-
rpi="http://schemas.unile.it/jds/2005/06/jds-rpi" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:rd="http://www.w3.org/2001/XMLSchema-instance" xmlns:jobref="http://www.w3.org/2001/XMLSchema-instance">
  <jds:JobIdentification>
    <jds:JobName>PDBL-ST</jds:JobName>
    <jds:Description>Workflow Job</jds:Description>
    <jds:JobProject>Project 1</jds:JobProject>
  </jds:JobIdentification>
  <jds-extension:Vertex id="2">
    <jds:Application>
      <jds:ApplicationName>psl_data_splitter</jds:ApplicationName>
      <jds-extension:JobType count="1">single</jds-extension:JobType>
      <jds-extension:Task>
        <jds-posix:Arguments>-o 3 -d data_rep -b uniprot -p "uniprot_sprot_human.dat_10" </jds-posix:Arguments>
        <jds:DataStaging type="data">
          <jds:FileName>uniprot_sprot_human.dat_10</jds:FileName>
          <jds:CreationFlag>overwrite</jds:CreationFlag>
        </jds:DataStaging>
        <jds:Source>
          <jds:URL>psl://sigma2.unile.it/home/bioinformput_human_protein/uniprot_sprot_human.dat_10</jds:URL>
        </jds:Source>
        <jds:DataStaging>
          <jds:DataStaging type="data">
            <jds:FileName>bioinfo</jds:FileName>
            <jds:CreationFlag>overwrite</jds:CreationFlag>
            <jds:Target type="multiple">
              <jds-extension:Vertex_Ref vertex_id="3" input_data_ref="3{filename}">
            </jds:Target>
          </jds:DataStaging>
        </jds-extension:Task>
      </jds:Application>
    </jds-extension:Vertex>
  </jds-extension:Vertex id="3">

```

Figure 7. Concrete Workflow.

between half the product and the upper limit, the performances are good; specifying half the product will allow the job to be scheduled earlier w.r.t. the other possible choice.

3. Checking Nexus file. Before submitting the job to the Meta Scheduler, the generated Nexus file or the input Nexus files can be visualized and modified without reloading the web page, exploiting AJAX capabilities.

We tested the system on all of the 3523 sequences of cytochrome oxydase I from Lepidoptera having the "BARCODE" keyword (that guarantees that the record's authors used the data standard promoted by the Consortium of Barcode of Life). The sequences were downloaded from EMBL, then reduced to 2080 non redundant sequences subdivided in 11 groups based on a priori phylogenetic information. The 11 groups ranged from 27 to 635 sequences, with 4 groups with more than 200 taxa. All of the data sets had approximately the same length. We chose a realistic but generic nucleotide model composed of a GTR substitution matrix [31] and a gamma distribution to model site variability [32], branch lengths without molecular clock constraint and unconstrained topology.

## Summary

The University of Salento group is responsible along with other partners for the creation of the LIBI Grid PSE. In particular, our action is oriented towards providing re-engineered bioinformatics applications and their composition into a workflow. Therefore, the applications use computational resources managed by different grid middlewares such as gLite, Unicore and Globus. Our goal is not only to provide the submission on a single machine, but also the interoperability between different resources.

The provided services are offered through a Grid Portal that also allows managing the resources, applications and databases available on the Grid.

Each service features an on line help to simplify its usage and to clearly understand the kind of input data required.

During this tutorial two case studies aiming at Multiple Sequence Alignment and Bayesian inference have been discussed. The users, once provided with a test account, could actually test the features of the system, submitting jobs and monitoring them through a web interface.

Future work is oriented toward a stress test of the system with other case studies in order to provide access to our virtual laboratory to external LIBI users for production purposes.

## Acknowledgements

We thank professor Rita Casadio and her collaborators for planning and developing the MSA service, and professor Cecilia Saccone, professor Graziano Pesole and Dr. Saverio Vicario for supporting us in the development of the MrBayes service.

## References

1. Culler G.J. and Fried B.D. An On-Line Computing Center for Scientific Problems. In Proc. 1963 Pacific Computer Conf., IEEE, Piscataway, N.J., pages 221-242, 1963.
2. Houstis E., Gallopoulos E., Bramley R., and Rice J. Problem-Solving Environments for Computational Science. IEEE Comput. Sci. Eng., 4(3), (1997), 18-21.
3. Von Laszewski G., Foster I., Gawor J., Lane P., Rehn N., and Russell M. Designing Grid-based Problem Solving Environments and Portals. In Proceedings of International Conference on System Sciences (HICSS-34), 2001.
4. Berman F., Hey A.J.G., and Fox G. Grid Computing: Making The Global Infrastructure a Reality. Wiley & Sons, (2003).
5. Workflow management coalition reference model. <http://www.wfmc.org/>.
6. Mirto M et al., (2008) The LIBI Grid Platform for Bioinformatics. In "Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine and Healthcare", IGI Global (to appear).
7. gLite Project. <http://glite.web.cern.ch/glite/documentation/>.
8. Erwin D.W. and Snelling D.F. UNICORE: A Grid

- Computing Environment. Lecture Notes in Computer Science, 2150, (2001), 825.
9. Foster I. and Kesselman C. Globus Toolkit Version 4: Software for Service-Oriented Systems. In Springer-Verlag LNCS 3779, editor, IFIP International Conference on Network and Parallel Computing, (2005), 2-13.
  10. EGEE Grid infrastructure from <http://www.eu-egee.org/>.
  11. DEISA Grid from <http://www.deisa.eu/>.
  12. Breton, V., Blanchet, C., Legré, Y., Maigne, L. and Montagnat, J. : Grid Technology for Biomedical Applications. M. Daydé et al. (Eds.): VECPAR 2004, Lecture Notes in Computer Science 3402, pp. 204–218, 2005.
  13. Jacq, N., Blanchet, C., Combet, C., Cornillot, E., Duret, L., Kurata, K., Nakamura, H., Silvestre, T., Breton, V.: Grid as a bioinformatics tool., *Parallel Computing, special issue: High-performance parallel biocomputing*, Vol. 30, (2004).
  14. Burbera F., Duffler M., Khalaf R., Mukhi N., Nagy W., and Weerawarana S. Unraveling the Web Services Web - An Introduction to SOAP, WSDL, and UDDI. *IEEE Internet Computing*, 6(2):86–93, 2002.
  15. Roy J. and Ramanujan A. Understanding Web services. *IT Professional*, 3(6):69–73, 2001.
  16. Krafzig D., Banke K., and Slama, D. Enterprise SOA: Service-Oriented Architecture Best Practices (The Coad Series). Prentice Hall PTR, November 2004.
  17. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lip-man, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17), 3389-402.
  18. Grillo, G., Licciulli, F., Liuni S., Sbisà, E. & Pesole G. (2003). PatSearch: a program for the detection of patterns and structural motifs in nucleotide sequences. *Nucleic Acids Res.*, 31, 3608–3612.
  19. Lindahl, E., Hess, B. & van der Spoel, D. (2001). Gromacs 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Mod.*, 7, 306-317.
  20. Lavorgna, G., Triunfo, R., Santoni, F., Orfanelli, U., Noci, S., Bulfone, A., Zanetti, G. & Casari, G. (2005). AntiHunter 2.0: increased speed and sensitivity in searching BLAST output for EST antisense transcripts. *Nucleic Acids Res.*, 1,33(Web Server issue),W665-8.
  21. Ronquist, F. & Huelsenbeck, J. P. (2003). MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19, 1572-1574.
  22. Aloisio, G., Cafaro, M., Carteni, G., Epicoco, I., Fiore, S., Lezzi, D., Mirto, M., & Mocavero, S. (2007). The Grid Resource Broker Portal. *Concurrency and Computation: Practice and Experience*, 19(12), 1663-1670.
  23. Alur N, et al. (2005) Patterns: Information Aggregation and Data Integration with DB2 Information Integrator. IBM Redbook publication: IBM Press. 334 p.
  24. Aloisio, G., Cafaro, M., Fiore, S. & Mirto, M. (2005a). The Grid Relational Catalog Project. In L. Grandinetti (Ed.), *Advances in Parallel Computing, "Grid Computing: The New Frontiers of High Performance Computing"* (pp.129-155). Elsevier, PA.
  25. Tuecke, S. Grid Security Infrastructure (GSI) Roadmap, Internet Draft, 2001, URL: [[www.gridforum.org/security/ggf1\\_-200103/drafts/draft-ggf-gsi-roadmap-02.pdf](http://www.gridforum.org/security/ggf1_-200103/drafts/draft-ggf-gsi-roadmap-02.pdf)].
  26. Mirto M., Rossi I., Epicoco I., Fiore S., Fariselli P., Casadio R., Aloisio G. "High Throughput Protein Similarity Searches in the LIBI Grid Problem Solving Environment" *Proceedings of the 5th International Symposium on Parallel and Distributed Processing and Applications (ISPA07)*, Niagara Falls (Canada), pp. 414-423.
  27. Mirto, M., Vicario, S., Tartarini, D., Epicoco, I. Saccone, C., Aloisio, G. "Bayesian Phylogenetic Inference in the LIBI Grid platform: a tool to explore large data sets". *IEEE Proceedings of the International Symposium on Parallel and Distributed Processing and Applications (ISPA 2008)* - December 10-12, 2008 - Sydney, Australia, pp. 855-860.
  28. Geyer, C.J. "Markov chain Monte Carlo maximum likelihood". In Keramidas (ed.), *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*. Interface Foundation, Fairfax Station, 1991, pp. 156-163.
  29. Altekar, G., Dwarkadas, S., Huelsenbeck, J. P., and Ronquist, F. "Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference". *Bioinformatics*, 20 (3), 2004, pp. 407-415.
  30. Kress, W.J., Erickson, D.L. "DNA barcodes: genes, genomics, and bioinformatics", *Proc. Natl. Acad. Sci. USA*. 2008 Feb 26; 105(8), pp. 2761-2.
  31. Lanave, C., Preparata, G., Saccone, C. Serio, G. "A new method for calculating evolutionary substitution rates". *Journal of Molecular Evolution*. 1984, 20:86-93.
  32. Yang, Z. "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods". *Journal of Molecular Evolution*, 1994, 39:306-314.
  33. Aloisio, G., Cafaro, M., Lezzi, D., Van Engelen, R. (2003). Secure Web Services with Globus GSI and gSOAP, *Proceedings of Euro-Par 2003*, 26th - 29th August 2003, Klagenfurt, Austria, Lecture Notes in Computer Science, Springer-Verlag, N. 2790, pp. 421-426.

## Querying the LIBI federated database through a data abstraction model



Luigi Doronzo<sup>1</sup>, Pietro Leo<sup>1</sup>, Graziano Pappadà<sup>2</sup>, Vincenzo Quinto<sup>2</sup>, Gaetano Scioscia<sup>1</sup>

<sup>1</sup> IBM Italy S.p.A. - Innovation Lab, Bari, (IT)

<sup>2</sup> Exhicon srl, Bari, (IT)

### Introduction

One of the most challenging task, probably exciting and disappointing at the same time, biologists would deal with during their activities is on one hand the huge and increasing volume of collected information publicly available, on the other the fact that information is distributed (literally *scattered*) across the web at a high level of untidiness. This fragmentation implies that a large number of different tools and methods have to be learned, sometimes in a time-consuming manner, in order to be able to extract knowledge from data.

The main consequence of such a scenario is under everybody's eyes: an invaluable richness of information as much as a frustrating difficulty to extract knowledge from it. Really this has been the feeling when in 2005 we began to work to the data and knowledge management subsystem of the LIBI project [1].

In the EMBet Tutorial on Grid Computing we presented how this kind of issue is being faced within the LIBI platform and some of the main tools researchers can find in such a platform to extract data and knowledge from the data sources containing relevant pieces of information for the research activities currently carried out in our project. In order to accomplish this task we focused our tutorial on two main topics: (i) data federation and the LIBI federated database we set-up to solve the problem of integrating data originated from several, dislocated, heterogeneous sources; (ii) data abstraction modelling as the approach to design a rationalized, data-decoupled view on the federated amount of informa-

tion and an appropriate web tool by means of querying the data abstraction model to valuably gathering new knowledge. The strong point of such a solution is that with just one tool researchers could perform queries against a number of different data sources simultaneously.

Both these issues have been exhaustively explored in the tutorial, and the full potentiality of the adopted approach has been enhanced through a simple case study, whose solution has been assumed as guideline during the hands-on session of the tutorial.

In the following sections the issues introduced before will be considered in details. First of all questions concerning data integration and the LIBI federated database will be examined; then the case study will be described in details, and finally the web tool for exploring and querying the abstract data model will be firstly described, and then used to address the case study.

### The LIBI federated database

At present, in the Bioinformatics' domain, an increasing number of grid applications manage data at very large scales of both size and distribution. The complexity of data management on a grid arises from the scale, dynamism, autonomy, heterogeneity and distribution of data sources [2]. Among these issues, heterogeneity is the most crucial factor to be considered in order to integrate biological data sources. It is a quite common experience finding in our laboratories or in a large number of life science projects a variety of data source formats, such as relational databases (e.g. Oracle, DB2, MySQL) mixed with non-relational sources. Such non-relational sources include XML documents, spread sheets, and other disparate file formats.

Therefore, the ability to manage, integrate and analyse related structured and unstructured information, possibly hosted in heterogeneous data source repositories, is fundamental for accelerating bioinformatics research.

Taking in count that two main classes of actors may need to access data in the bioinformatics domain, namely researchers (to simply examine them) and *algorithms* (to read data for analysis purposes), the key elements that should be addressed by the data management framework concern data reliability and uniformity of the access way.

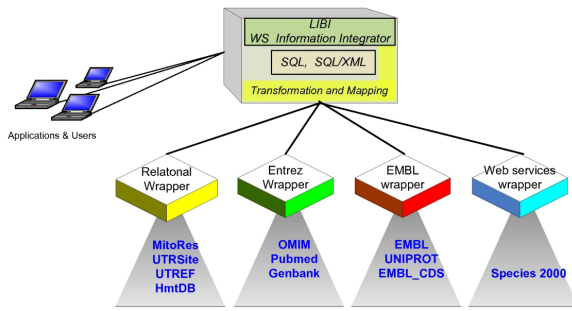


Figure 1. High-level component model of the data federation architecture implemented in the LIBI platform.

In the following, evaluations carried out for the LIBI's data management and integration approaches, and the finally adopted solution are reported in detail.

### Federation vs Data Warehouse

According to the LIBI strategy and the architecture of its platform, both data federation [3] and data consolidation (data warehousing) [4] have been evaluated. Although there are no definitive considerations in favour of just one of these two viable approaches, being the solution strictly connected to the research scenario and goals, also in the LIBI project we leaned towards a hybrid approach, but with a slightly favour for the fed-

eration mechanism. In fact federation has some pros with respect to data warehousing for what concerns (i) the preservation of eventual special-search capabilities a data source would be able to expose (definitively ruled out in the case data were aggregated in a data warehouse), (ii) a federated approach to both data and information integration provides the ability to synchronize distributed data without requiring they are moved to a central repository, since data remain where they are, (iii) data federation allows users and algorithms to transparently access current data from multiple, heterogeneous, dislocated data sources simultaneously, through a single standard interface. So, for bioinformatics problems, in particular for those managed within LIBI, data federation seems the most promising solution.

The implementation of the data federation in the LIBI platform [1] is schematically depicted in Figure 1, where the high-level component model of the federation architecture is shown. The engine of this solution is the component that plays the role of data federation server, which is based on IBM WebSphere Information Integrator [5]. It allows users and applications to access all federated data sources by means of a standardized interface: SQL. In Figure 1, the wrappers' layer re-

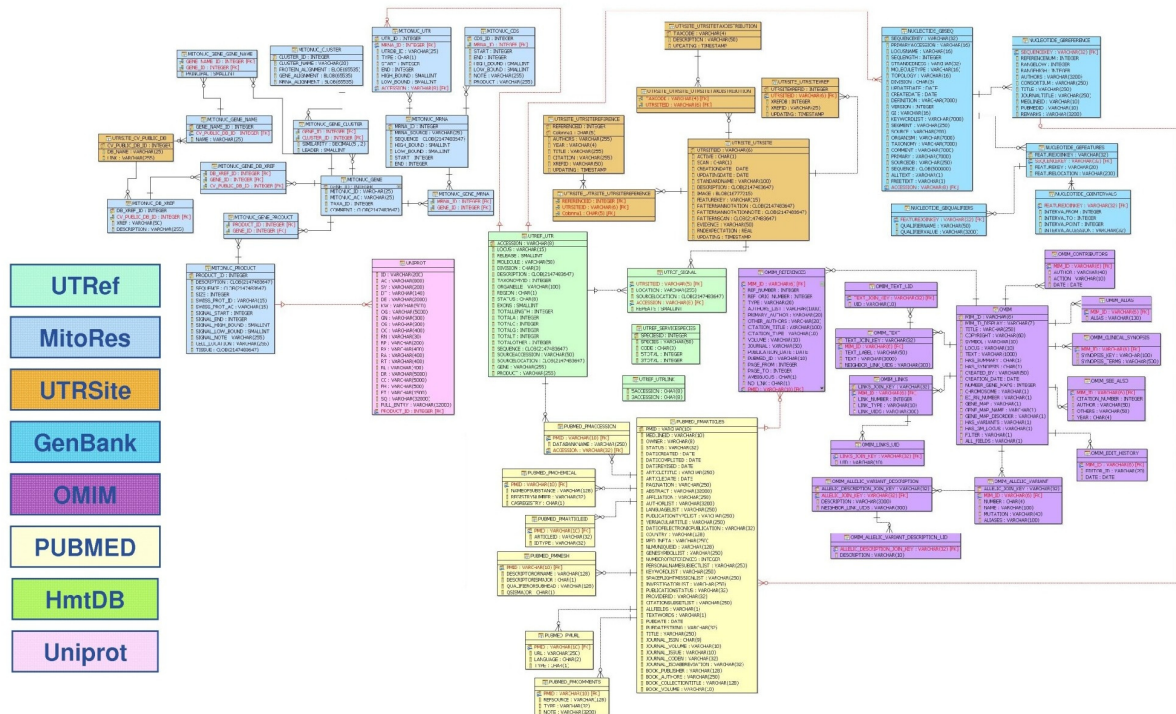


Figure 2. Simplified schema of the LIBI federated database. Different colours of the tables refer to the specialized source from which data are coming; lines among tables show discovered inter-source relations.

ports the four components that, plugged within the engine, allows the federator to communicate with the data sources. Since each of these components is specialized with respect to the source it connects to, the wrappers act as the channels by means of which the federation server builds a fully-effective interaction with the heterogeneous data origins. A number of wrappers useful for typical bioinformatics data source formats are already available, and other could be designed and implemented by means of the SDK provided by the WebSphere Information Integrator, as the case of EMBL Wrapper [6]. Figure 1 also shows main bioinformatics databases already integrated in the federated schema. Some of them are well-known public specialized databases such as GenBank [7], PubMed [8], and OMIM [9] provided by NCBI, the mitochondrial resource HmtfDB [10], UNIPROT [11], Species 2000 [12], MitoRes [13], UTRef and UTRSite [14].

In such a manner the federation server allows to setup a relational schema that is the sum of the contributions coming from the federated sources. A simplified schema containing the most important pieces of information of the LIBI federated database is shown in Figure 2, where a colour legend identifies the different contributions, and inter-tables lines refer to discovered inter-relations among data belonging to different sources.

### Solving bioinformatics problems: a case study

A typical need of a large number of biologists consists in retrieving pieces of information from data sources that are dislocated and also heterogeneous for what concerns the access interface. Once they individuate the data sources useful for their research activities, they have to query them and then organize the retrieved data in a suitable manner to answer to the scientific question they have in mind. In order to perform these tasks, the most commonly retrieval systems used by biologists are SRS [15,16] available at EBI and Entrez [17] available at NCBI. However, if biologists would have to search data resident in databases available only in their institution, and if these databases are not indexed in one of the retrieval systems cited above (namely SRS), they must find a different way to retrieve their data. For example, in such a case they could use query systems designed specifically for those inter-

nal databases, but they would use other tools if their queries would involve databases available outside their institution only. Here we present a simple case study that highlights the difficulties a researcher must deal with when making queries against specific data sources.

#### A simple case study

In our case study we suppose that a biologist is studying the regulation of the expression, at the mRNA level, of the mitochondrial topoisomerase I. Since the user is interested in the regulation at the mRNA level, he decides to investigate if there are annotations about UTR sequences that regulate the expression of topoisomerase I; if these sequences are present, he is interested also in knowing which regulatory motif is present in the UTR sequences. The biologist knows that the pieces of information to retrieve are dislocated in different databases and so he has to jump from one database to another. The three databases involved in her query are MitoRes [13], a resource of nuclear-encoded mitochondrial genes and their products in Metazoa, UTRef [14,18], containing 5' and 3' mRNA untranslated regions from RefSeq [19], and UTRSite [14,18], a collection of functional motifs located in 5' and 3' UTR sequences.

In this case study, our biologist is interested in the gene coding for the mitochondrial topoisomerase I; this information is available in the database MitoRes. From this database, using "topoisomerase I" as search criterion, the biologist retrieves the MitoRes entry accession number, the product description and the name of the gene coding for the product. Then, if the gene under analysis contains UTR sequences, the biologist retrieves the accession number and the type (if 5' or 3') of the UTR from the UTRef database. Finally, to retrieve the accession number and the standard name of the regulatory motif contained in the UTR sequence, the biologist needs to query and retrieve data from the UTRSite database.

Once the user obtained all the pieces of information of interest, he has to organize them, for example writing a summary table.

#### Case study solution through the federated DB

The case study introduced previously is a suitable test bed for the capabilities offered by the LIBI federated database, because it integrates the three data sources from which information has to be extracted.



```

SELECT DISTINCT
  "t1"."DESCRIPTION" AS "Description", "t2"."MITONUC_ID" AS "ID",
  "t3"."UTRDB_ID" AS "UTRRef ID", "t3"."TYPE" AS "UTR type",
  "t4"."NAME" AS "Gene name", "t5"."UTRSITEID" AS "UTRSite ID",
  "t5"."STANDARDNAME" AS "Standard name"
FROM
  "LIBI"."MITONUC_GENE" "t2" LEFT JOIN "LIBI"."MITONUC_GENE_PRODUCT" "t6"
ON "t2"."GENE_ID" = "t6"."GENE_ID" RIGHT JOIN "DDQB"."MITONUC_PRODUCT_DDQB" "t1"
ON "t6"."PRODUCT_ID" = "t1"."PRODUCT_ID" LEFT JOIN "LIBI"."MITONUC_GENE_MRNA" "t7"
ON "t2"."GENE_ID" = "t7"."GENE_ID" RIGHT JOIN "DDQB"."MITONUC_MRNA_DDQB" "t8"
ON "t7"."MRNA_ID" = "t8"."MRNA_ID" LEFT JOIN "DDQB"."MITONUC_UTR_VIEW" "t3"
ON "t8"."MRNA_ID" = "t3"."MRNA_ID" LEFT JOIN "LIBI"."UTREF_UTR" "t9"
ON "t3"."UTRDB_ID" = "t9"."ACCESSION" LEFT JOIN "LIBI"."UTREF_SIGNAL" "t10"
ON "t9"."ACCESSION" = "t10"."ACCESSION" LEFT JOIN "DDQB"."UTRSITE_UTRSITE_DDQB" "t5"
ON "t10"."UTRSITEID" = "t5"."UTRSITEID" LEFT JOIN "LIBI"."MITONUC_GENE_GENE_NAME" "t11"
ON "t2"."GENE_ID" = "t11"."GENE_ID" RIGHT JOIN "LIBI"."MITONUC_GENE_NAME" "t4"
ON "t11"."GENE_NAME_ID" = "t4"."GENE_NAME_ID"
WHERE
  UPPER("t1"."DESCRIPTION") LIKE '%TOPOISOMERASE I%'

```

Figure 3. The federated query that solves the case study when it is sent against the LIBI federated database.

In this section we plan to solve the case study by designing a SQL query that will be sent against the federated database using an appropriate environment so that the result set may be retrieved for evaluation and further analyses.

In this case, a researcher with robust SQL skills carefully examines the federated database schema reported in Figure 2 in order to identify all the tables containing the entities relevant for her query. In such an evaluation he will take in count not only the fields containing the data useful to solve the test case, but also those fields that allow to build relationships both within the same data source and across different data sources. By putting all this together, an SQL-expert user should be able to write down a query as that reported in Figure 3, that allows to retrieve the entire information he needs from the federated database.

In order to send this query against the LIBI federated database a SQL-client must be used to establish a connection to the database manager and execute the query.

Figure 4 shows an example of results produced by the execution of query reported in Figure 3. This result set consists of six records reporting the solution to our case study. In the case of the LIBI federated database the retrieval process elapsed 11 seconds, 10 of which have been spent by the federated database in order

to resolve the query and retrieve results from the original sources, and the rest have been spent by the SQL-client to visualize the result set.

This kind of solution for the proposed case study seems quite short to implement, but a crucial point is that the work to write down a SQL query such that in Figure 3 may be a non-trivial task, also for SQL-skilled people. Certainly we cannot expect this kind of competencies from biologists or from a large number of bioinformatics scientists.

## Modelling data abstraction in Bioinformatics

The interface provided by the LIBI federator server is SQL, so it can be used directly by specialized bioinformatics applications or by SQL-skilled people. In order to allow also non-technical people access the federated database layer, a specific tool to retrieve information as well as to allow data mash-up is also required. To this end a new, specialized tool has been introduced in the LIBI platform: the IBM Data Discovery and Query Builder for Healthcare and Life Sciences (DDQB) [20]. It is a powerful search tool with a graphical interface that enables users with various levels of expertise to easily configure queries and leverage the full spectrum of information assets.

Figure 5 shows the high-level operation flow of DDQB. The key concepts of this system are the

ID	Description	UTRRef ID	UTR type	Gene name	UTRSite ID
HSAPTOP1M	DNA topoisomerase I, mitochondrial precursor ...	BR035994	5	TOP1MT	U0011
HSAPTOP1M	DNA topoisomerase I, mitochondrial precursor ...	CR038312	3	TOP1MT	<null>
HSAPQ86V82	Mitochondrial topoisomerase I.	BR035994	5	TOP1MT	U0011
HSAPQ86V82	Mitochondrial topoisomerase I.	CR038312	3	TOP1MT	<null>
DRERQ6T721	Mitochondrial topoisomerase I.	CR166253	3	TOP1MT	<null>
DRERQ6T721	Mitochondrial topoisomerase I.	BR163315	5	TOP1MT	<null>

Figure 4. Federated query results.

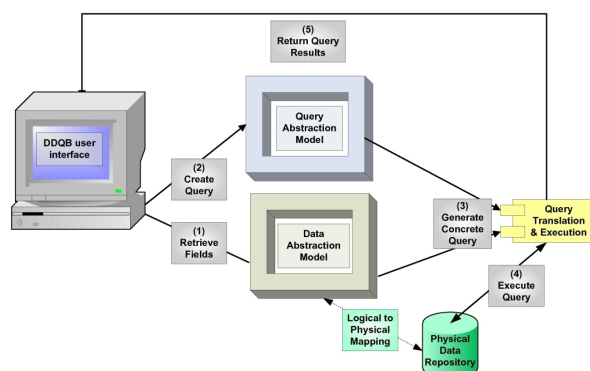


Figure 5. High-level Data Discovery & Query Builder operation flow.

*Data Abstraction Model* (DAM) and the *Query Abstraction Model* (QAM). The former consists of an XML repository that stores all the information concerning the DB physical model, the abstract entities the end-user can handle, and their mapping (this logical to physical mapping is designed just once by the DDQB administrator). Most DDQB behaviour is driven by this component. On the other hand the Query Abstraction Model provides a XML representation of an abstract query in terms of logical fields, and is formulated by considering both the authorization level of the user (with a granularity down to the field level) and the selection logic imposed by the query criteria. The component *Query Translation & Execution* is responsible for composing a concrete, well format SQL query on the basis of the user inputs, and sends it against the federated database (step (4) in Figure 5). End-users construct queries by operating on DAM and QAM in a graphical way by navigating and acting on a generic taxonomy pointing to the LIBI federated database biological concepts. These aspects of DDQB and whatever concerns the end-user perspective of working with this tool will be discussed in the following sections.

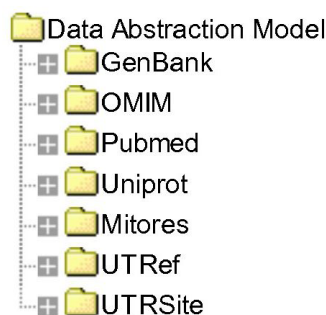


Figure 6. The LIBI Data Abstraction Model.

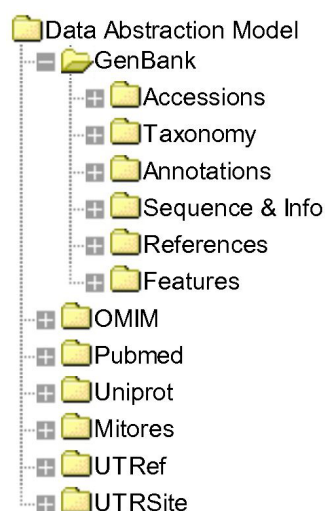


Figure 7. The LIBI Data Abstraction Model: sub-categories for GenBank.

## The LIBI Data Abstraction Model

In order to satisfy the research needs of the LIBI users, we designed a specific Data Abstraction Model (DAM). The DAM is the crucial part the *DB administrator* has to face in setting up a DDQB application. In fact, designing it requires a considerable amount of work since in the DAM both the knowledge of the physical structure of the underlying database and the acquaintance of the query exigencies of the bioinformatics users should be condensed. This means that the *DB administrator* has to be not only a subject matter expert in Bioinformatics, but also a professional with valuable technical project knowledge. For these reasons, the task to design a complex DAM is generally delegated to a multidisciplinary team.

The DAM set up for the LIBI project is, as far as we know, the first attempt to apply DDQB solely to the bioinformatics domain, being DDQB conceived principally for the clinical world [20].

In the LIBI DAM we organized data in categories corresponding to the databases integrated within the LIBI federated DB. In each category there are sub-categories that collect data of similar content or significance in order to simplify the user task of finding the field he is searching for during the composition of the query. In particular, the LIBI data abstraction model is composed by seven categories corresponding to seven of the databases we integrated in the LIBI federated DB. Figure 6 shows the tree of the DAM as

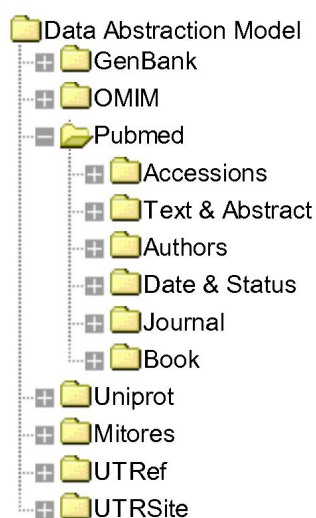


Figure 8. The LIBI Data Abstraction Model: sub-categories for PubMed.

viewed by a user working with the tool DDQB we will describe in the following.

First level categories of the LIBI DAM are GenBank, OMIM, PubMed, Uniprot, MitoRes, UTRef and UTRSite.

Each of these categories has sub-categories. For example, the entity GenBank (Figure 7) has six sub-categories: *Accessions*, *Taxonomy*, *Annotations*, *Sequence & Info*, *References* and *Features*. Inside the sub-category *Accession* we have catalogued all pieces of information regarding unique IDs identifying the entry such as accession number, GI number, etc. and information like the version of the entry and the creation and update date of the entry. In the same way, in the sub-category *Taxonomy* we put together all data concerning taxonomic information, such as the organism name, the taxonomic division and the taxonomic classification.

In the sub-category *Annotations* we grouped all fields related to annotations associated to an entry, such as descriptions, keywords or comments.

Sub-categories with the same name may or may not be present in different parent categories, as can be seen comparing Figure 7 with Figure 8. From this, the user can infer that categories with the same names refer to entities analogous either from the logical or semantic points of view.

It is important to note that this data abstraction model is one of the possible data abstraction models that can be designed for the LIBI users; it is also possible to define more abstract DAMs



Figure 9. DDQB home page.

that do not refer explicitly to technical terms like the database name, but instead refer to more abstract concepts strictly related to the biological domain. In such a DAM abstract entities and physical entities are loosely coupled.

In conclusion: an administrator of the federated database may design her appropriate DAM targeted to the users that access federated data.

### Exploring the LIBI DDQB

One of the goals of the LIBI project is to allow LIBI users, and in general researchers without programming skills, to perform complex queries against biological databases in order to make easier the research tasks. With a simple access to the knowledge stored in the federated database, the LIBI users can gain their research results in quite a short time.

IBM Data Discovery and Query Builder (DDQB) [20] is a system that provides a web-based interface that helps researchers to formulate and execute database queries to identify and correlate information stored in a relational database. DDQB helps users with different skills to perform queries both on structured and unstructured data. The user can:

- aggregate, query and save data;
- mark a query as public to allow other users to execute the same query or modify it;
- mark a query as private to allow the access to that query only to a selected team of collaborators.

Some peculiarities of DDQB that aid to simplify, speed up and optimize the work of the researchers are:

1. web-based customizable interface;
2. suitable for user with different information technology skills level, with no need to know specific query language like SQL;
3. possibility to build simple or complex queries;

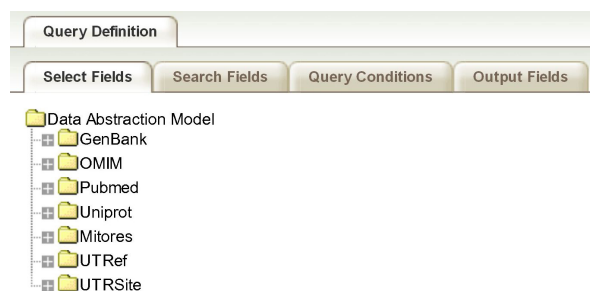


Figure 10. DDQB Query Definition.

4. share query and results with other researchers.

DDQB offers to the users both simple and advanced functionalities to simplify the definition of complex queries.

Once the user has logged in, he has four possible tasks to select (Figure 9); he can:

1. create a new query;
2. work with saved queries;
3. work with saved query results;
4. work with saved analysis results.

If the user decides to create a new query, DDQB will present the "Query Definition" section in which he has the possibility to compile a new query (Figure 10).

In this section there are four tabs:

1. "Select Fields" - here the user can browse the DAM in order to select the entities useful to compose the query.
2. "Search Fields" - helps to search for a particular field in the DAM instead of browsing it;
3. "Query Conditions" - here the user can take a look at the query he is composing. The query is expressed in a "natural language" representation, instead of more technical SQL or other query languages. In this section the user can edit the query conditions grouping them by means of boolean operators, deleting conditions, modifying search criteria;
4. "Output Fields" - here all the fields that will be returned in the query output are summarized. The user can delete some output fields or reorder them;
5. "New Query Condition" - this tab appears when a user is imposing new search criteria. In this tab the user can choose how to set search criteria, as can be seen in Figure 11.

After the user has executed his query, a new tab appears: this is the "Query Results" tab (Figure 12) in which the results are displayed. The user can

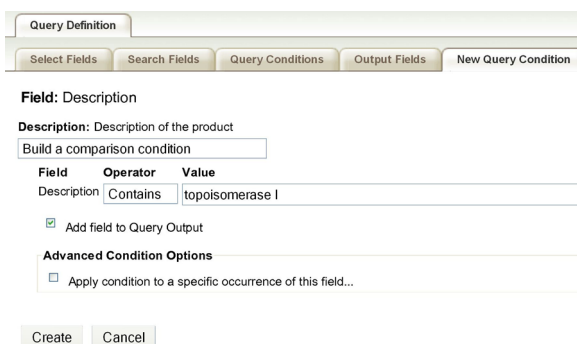


Figure 11. Defining a new query condition in DDQB.

save them in different formats (CSV, CSV for Excel, Tab delimited and XML).

Looking at the top of the DDQB interface, there are four high-level menus:

1. *File* - the user can select the Entity Model and manage saved queries and results;
2. *Run* - the user can select if she wants to run her query with an on-line interaction or in background;
3. *Query* - the user can select some options about the query (e.g. showing the SQL query corresponding to what the user composed, setting advanced options to view advanced output, etc.);
4. *Session* - here the user can logout from the session or set session preferences;
5. *Help* - to view a guide for DDQB.

As described above, the user composes the query by browsing the data abstraction model and using the fields of interest involved in the query. In order to correctly use the DAM, it is important to notice that for each field there is a link and a checkbox (Figure 13). By clicking on the link the user can add search criteria, while checking the checkbox the user will simply add that field to the output returned by the query, without imposing search criteria on it.

## Case study solution through the Data Abstraction Model

In this section, the case study we introduced previously will be solved by using DDQB. To this end, the following instructions have to be carefully carried out.

The user, after authentication, clicks on the first link "Create a new query" (Figure 9).

In the following sections detailed steps concerning the DDQB tutorial are reported.

Query Definition*		Query Results*				
Results						
1 - 6 of 6						
Alternate Output Formats						
<a href="#">CSV</a> - <a href="#">CSV for Excel</a> - <a href="#">Tab Delimited</a> - <a href="#">XML</a>						
Description	ID	UTRef ID	UTR type	Gene name	UTRSite ID	Standard name
DNA topoisomerase I, mitochondrial precursor (TOP1mt)	HSAPTOP1M	BR035994	5	TOP1MT	U0011	Terminal Oligopyrimidine Tract (TOP)
DNA topoisomerase I, mitochondrial precursor (TOP1mt)	HSAPTOP1M	CR038312	3	TOP1MT	null	null
Mitochondrial topoisomerase I.	DRERQ6T721	BR163315	5	TOP1MT	null	null
Mitochondrial topoisomerase I.	DRERQ6T721	CR166253	3	TOP1MT	null	null
Mitochondrial topoisomerase I.	HSAPQ86V82	BR035994	5	TOP1MT	U0011	Terminal Oligopyrimidine Tract (TOP)
Mitochondrial topoisomerase I.	HSAPQ86V82	CR038312	3	TOP1MT	null	null

Figure 12. DDQB Query results.

### Compose query

Follow these steps to compose the query that solves the case study:

1. go to the "File" menu and select "New Query: Mitores". In this way you select the entity model, i.e. the entity that is the starting point of our query (Figure 14);
2. expand the tree following these steps: *MitoRes* → Product, and click on the link Description;
3. from the dropdown menu select "Build a comparison condition";
4. from the dropdown menu "Operator" select "Contains"; in the "Value" field write "topoisomerase I" (the textual criterion of our query) and then click on the "Create" button (Figure 11);
5. take a look at the condition just created and then click on the "Select Fields" tab to return to the Data Abstraction Model tree page;
6. expand the tree following these steps: *MitoRes* → Accessions and check the checkbox for ID;
7. expand the tree following these steps: *MitoRes* → mRNA → UTR and check first the checkbox for UTRRef ID; then check the checkbox for UTR type;

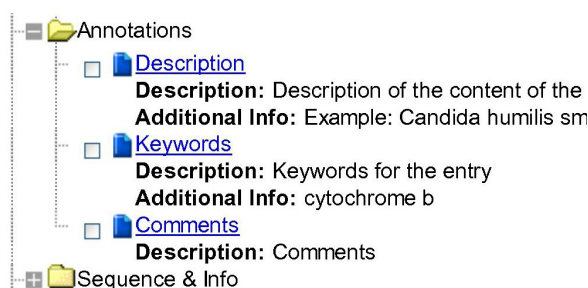


Figure 13. Composing query in DDQB: the link is used to impose search criteria on the field, the checkbox to add it to the output.

8. expand the tree following these steps: *MitoRes* → Gene, and check the checkbox for Gene name;
9. expand the tree starting from the root of the DAM: UTRRef → Signal and check the checkbox for UTRSite ID;
10. expand the tree starting from the root of the DAM: UTRSite → Annotations and check the checkbox for Standard name;
11. click on the "Output Fields" tab to view a summary of the fields that will be returned in the output.

### Execute query

Now we have to execute the query we just composed. Follow the steps:

1. go to the "Run" menu and select "Run";
2. wait until the query is executed;
3. see the results.

The results are organized in a tabular way. There is a column for each field selected during the query composition; for instance, in the first column there is the Description field, in the second column the MitoRes ID field, etc. (Figure 12). The results show that the gene for mitochondrial topoisomerase I,

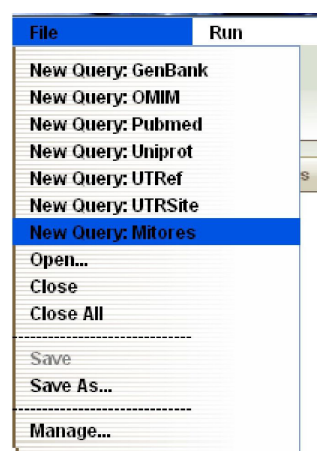


Figure 14. DDQB: Select entity model.

Query Definition\* Query Results\*

Select Fields Search Fields Query Conditions Output Fields

Advanced Output Options

Query Settings:

Fetch data as:  Rows  Entities

Limit output to first:

Randomize the output order

Duplicate output rows should be removed

OK Cancel

Figure 15. DDQB: Fetch data as entities.

whose gene name is TOP1MT, contains a regulatory motif called "Terminal Oligopyrimidine Tract (TOP)" in the 5' UTR, while there is no information in the database about a regulatory motif in the 3' UTR.

These results answer to the research problem formulated in our case study.

### Export results

Once we have the results, we can export them:

1. in the output display page, above the results table, there is a link named "CSV"; click on it to export results in a Comma-Separated Value format, and save the file on the desktop of your workstation;
2. open the file to view the exported results.

### Use advanced output

Now we are going to view the output results in a different, more readable way. In order to do this, follow these steps:

1. click on the "Query Definition" tab to return to the "Output Fields" page;

2. go to the "Query" menu and select "Advanced Output...";
3. change the selection from "Fetch data as:" "Rows" to "Entities" and click on the "OK" button (Figure 15);
4. go to the "Run" menu and select "Run";
5. wait until the query is executed;
6. view how different is the table reporting the results (Figure 16).

In Figure 16 the results of our query are shown in a different format: here they are aggregated by entities instead by rows, instead of eight columns we have only four columns; the most interesting one is the third one. For each MitoRes ID (first column), that represents our entity, the third column summarizes all the pieces of information regarding the UTR sequences and the standard name of the regulatory motif, if present. Comparing this advanced view with the simpler view we saw previously (view as rows), we notice that now information is presented to the user in a more organic way, and knowledge can be inferred easier. In fact, by viewing the results as entities we can individuate immediately which MitoRes entries do not have an notations of UTR sequence.

### Run parameterized queries

Although in the previous section the case study has been solved, now we can explore an interesting feature of this query system. By using the parameterized queries feature, a user can write a query leaving the search criterion blank. He can add the search criterion later. This feature enables the user to reuse the same query with different search criteria. We can apply this feature to the query we just executed. Follow the steps:

1. click on the "Query Definition" tab to return to the "Output Fields" page;

ID	Description	UTRRef ID	UTR type	UTRSite ID	Standard name	Gene name
HSAFPTOP1M	DNA topoisomerase I, mitochondrial precursor (TOP1mt).	CR038312	3			TOP1MT
		BR035994	5	U0011	Terminal Oligopyrimidine Tract (TOP)	
CELETP2M	Putative DNA topoisomerase II, mitochondrial precursor.					
HSAPQ86V82	Mitochondrial topoisomerase I.	CR038312	3			TOP1MT
		BR035994	5	U0011	Terminal Oligopyrimidine Tract (TOP)	
RNORQ6IM78	Mitochondrial DNA topoisomerase I.					TOP1MT
MMUSQ9D6H0	Mus musculus adult male hippocampus cDNA, RIKEN full-length enriched library, clone:2900052H09 produ					TOP1MT
DRERQ6T721	Mitochondrial topoisomerase I.	CR166253	3			TOP1MT
		BR163315	5			
GGALQ6T722	Mitochondrial topoisomerase I.	BR181563	5			
		CR188559	3			

Figure 16. DDQB results aggregated by entities instead by rows.

**Field:** Description

**Description:** Description of the product

Defer condition details (parameter)

Add field to Query Output

Figure 17. DDQB parameterized query: defer condition details.

- click then on the "Query Conditions" tab, select the condition displayed and then click the "Edit" button;
- change "Build a comparison condition" in "Defer condition details (parameter)" and click the "Update" button (Figure 17);
- the query condition has now been modified. Instead of the search criteria there is a question mark (?), which means that the search criterion has to be defined later, at the execution time (Figure 18);
- go to the "Run" menu and select "Run";
- from the dropdown menu select "Build a comparison condition" (Figure 19), choose the operator "Contains", write as value "acyl-CoA dehydrogenase" and finally click the "Finished" button;
- wait until the query is executed;
- view the results

### Save Queries & Results

The results and the query can be saved for future reuse. Follow these steps:

- go to the "File" menu and select "Save As...";
- insert a name for the results to be saved, for example "UTR regulatory motif for acyl-CoA dehydrogenase";
- insert a description for the query;
- click the "Save" button;
- click on the "Query Definition" tab to return to the Query Conditions page;

Query Definition\* Query Results\*

Select Fields Search Fields Query Conditions Output Fields

Select  Condition

Description equals ?

Negate Edit Copy Paste Delete Label

Group AND Group OR Ungroup

Figure 18. DDQB parameterized query.

Query Definition\* Query Results\*

Select Fields Search Fields Query Conditions Output Fields Run Query

Name:

Wizard Steps

1) Description

**Field:** Description

**Description:** Description of the product

Build a comparison condition

Field	Operator	Value
Description	Contains	acyl-coa dehydrogenase

<< Previous Next >> Finished Cancel

Figure 19. DDQB: filling in the search criterion for parameterized queries.

- go to the "File" menu and select "Save As...";
- insert the name of the query to be saved, for example "Find UTR regulatory motif";
- insert a description for the query, e.g. "Parameterized query to find regulatory motif in UTRs from MitoRes database";
- click the "Save" button;
- now you can logout.

As we saw above, by means of DDQB we were able to solve rapidly our case study without writing down any SQL statement. Moreover, the DDQB system offers interesting features (of which we explored just few) that can facilitate the reuse of the query and results, can improve information visualization for a better knowledge extraction, and so on.

### Summary

The IBM group is acting within the LIBI project as responsible for the information and knowledge management layer of the laboratory platform [1]. It is focused in designing and making available within the LIBI platform on one hand a set of systems able to introduce standards for addressing issues as accessing, managing and integrating data; on the other hand a set of tools to endow the researches using the platform to boost and facilitate what concerns data and knowledge mining and retrieval.

Our studies leaded us to prefer, in a grid environment and in the bioinformatics domain, solutions based on a federated approach. In fact, it fits better the needs of biologists, which are asking for always updated, uniform and integrated data. From such a research path the LIBI Federated Database arose. It plays the role of a virtual layer aiming to hide the heavy data untidiness and fragmentation existing in the present biomedical world.

The federated database exposes a SQL interface that allows to retrieve data from federated sources, solving the fragmentation problem. Although such an interface can be suitable for application developers, it is not addressed to biologists, since valuable technical skill should be required. To this end, another abstraction level has been introduced upon the federated layer. This new layer defines how information concerning conceptual entities (domain-specific entities like genes, products and so on), contained inside the database, map to real data. This task is carried out using IBM DDQB, which provides both a decoupling layer between physical and abstract data (the Data Abstraction Model) and a web tool for composing complex queries using a graphical interface.

During this tutorial a case study aiming at discovering how the expression of mitochondrial topoisomerase I is regulated at mRNA level has been presented and solved both using the SQL-based and the DDQB (DAM-based) approaches.

The latter solution provided us with the opportunity to explore some of the DDQB's interesting features, and to appreciate the valuable potential of the combination of a Data Abstraction Model built upon a federated database. Such a solution allows people without strong IT skills to use the federated database to solve quite complex tasks of information extraction. The DDQB system earned a considerable, positive interest by several attendants to the tutorial. It proved, one more time, to be a useful tool to discover and retrieve information in the biomedical domain.

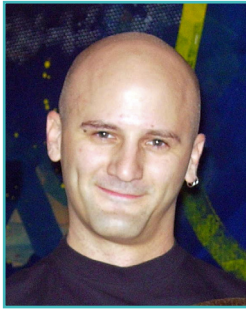
## References

1. Mirto M et al., (2008) The LIBI Grid Platform for Bioinformatics. In "Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine and Healthcare", IGI Global (to appear).
2. Bourbonnais S, et al. (2004) Towards an information infrastructure for the grid. IBM Systems Journal 43(4) 2004: 665-688.
3. Haas L, et al. (2002) Data Integration through Database Federation. IBM Systems Journal 41(4) 2002: 578-96.
4. Shah SP, et al. (2005) Atlas – a data warehouse for integrative bioinformatics. BMC Bioinformatics 2005 Feb 21: 6-34.
5. Alur N, et al. (2005) Patterns: Information Aggregation and Data Integration with DB2 Information Integrator. IBM Redbook publication: IBM Press. 334 p.
6. Leo P, et al. (2008) EMBL/FASTA Wrapper for WebSphere Information Integrator. IBM Alphaworks. Available: <http://www.alphaworks.ibm.com/tech/em-blwrapper>. Accessed 11 November 2008.
7. Benson DA, et al. (2008) GenBank. Nucleic Acids Res. In press.
8. Delwiche FA (2008) Searching MEDLINE via PubMed. Clin. Lab. Sci. 21(1): 35-41.
9. Amberger J, et al. (2008) McKusick's Online Mendelian Inheritance in Man (OMIM(R)). Nucleic Acids Res. In press.
10. Attimonelli M, et al. (2005) HmtDB, a human mitochondrial genomic resource based on variability studies supporting population genetics and biomedical research. BMC Bioinformatics 2005 6 Suppl 4:S4.
11. The UniProt Consortium (2009) The Universal Protein Resource (UniProt). Nucleic Acids Res. In press.
12. Edwards JL, Lane MA, Nielsen ES (2000) Interoperability of biodiversity databases: biodiversity information on every desktop. Science 289(5488): 2312-4
13. Catalano D, et al. (2006) MitoRes: a resource of nuclear-encoded mitochondrial genes and their products in Metazoa. BMC Bioinformatics 2006 24: 7-36.
14. Mignone F, et al. (2005) UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res. 33(Database issue): D141-6.
15. Zdobnov EM, et al. (2002) The EBI SRS server – recent developments. Bioinformatics 18(2): 368-73.
16. Zdobnov EM, et al. (2002) The EBI SRS server – new features. Bioinformatics 18(8): 1149-50.
17. Baxevanis AD (2006) Searching the NCBI databases using Entrez. Curr. Protoc. Bioinformatics. Chapter 1: Unit 1.3
18. Pesole G, et al. (2002). UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. Nucleic Acids Res. 30(1):335-40.
19. Pruitt KD, Tatusova, T, Maglott DR (2007). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35(Database issue): D61-5
20. Adan M, Dettinger R (2003) IBM Data Discovery and Query Builder: Plug-ins by Example. IBM Redpaper publication: IBM Press. 126 p.



## AntiHunter 3.0

### Identification of antisense transcripts: a practical tutorial



Francesco Falciano<sup>1</sup>,  
Giovanni Lavorgna<sup>2</sup>,  
Andrew Emerson<sup>1</sup>,  
Elda Rossi<sup>1</sup>,  
Andrea Vanni<sup>1</sup>

<sup>1</sup> CINECA, Casalecchio di Reno, BO (IT)  
<sup>2</sup> DIBIT, HSR, Milano (IT)

The functional role for mRNA antisense in prokaryotes and viruses has been documented in several works (1). Further studies in eukaryotes suggest that antisense RNAs are involved in gene regulation (2) including transcriptional interference (3), genomic imprinting (4,5), RNA interference (6), translational regulation (7), alternative splicing (8), X-inactivation (9), RNA editing (10) and promoter methylation (11).

Through the LIBI portal, it is possible to submit a Unicore job, on a remote machine where the program called Antihunter is installed. Its main goal is to identify potential antisense transcripts within a genomic region of interest (12).

(<http://www.libi.it/libi/biotools/libi-biotools/>)

To obtain an account send an email to maria.mirto@unile.it

### Introduction

In the biological process called *transcription*, the information stored by the DNA in the nucleus is written using a molecule called mRNA that is carried in the cytosol where there are specialized structures that read this information and translate it into a protein. One way to "interfere" with the translation process is to introduce transcripts complementary to endogenous mRNA. Such molecules are called "antisense RNA" and inhibit translation of a complementary mRNA by base pairing to it and physically obstructing the translation machinery. These antisense molecules are produced by the same organism or can be

introduced artificially. Several publications have shown how antisense RNAs are utilized by organisms to regulate the translation process and how they could be potentially used for performing this regulation artificially.

### How to use the AntiHunter application on the LIBI portal

AntiHunter takes in input a sequence and performs a BLAST search on the dbEST of the same organism (or even of other organisms). The ESTs detected are compared with the list of annotated transcripts given in input to check if they can be potential antisense transcripts and positive results are sent to the user.

You have two options to submit your query in the web form. The first option is to type (better copy and paste) the query sequence in the first window, and annotations about transcripts encoded by the query sequence in the second window.

Another option is to submit the "coordinates" of the query sequence (species, chromosome, nucleotide starting position and nucleotide ending position). Several genomes of a number of organisms are available. In this case it is not necessary to provide the annotations because the db already store this information.

The second part of the form provides options for performing a more advanced search.

In the first field you can choose the organism for which you want to perform the search of the EST. Notice that you can look for the antisense RNA of a different organism of the sequence given in input. This option can be useful if for example, you are looking for some potential candidates to be investigated in gene therapy experiments.

The second field allows searching for only the EST that has in the description field a specific keyword. For example, you can use the keyword "heart" to look for only the EST expressed in heart, or "tumor" to look for only the EST expressed in tumors.

"Start from EST number" allows to start the search in a specific order of the databases. This is a trick introduced due to a limitation: for technical reasons the search is performed only in the first 200,000 (two hundred thousand) ESTs, so, you can perform the complete search in 2 or more steps. In this field you can put the number 200,001 and the search will start from here.

**CUMULATIVE and SUBMATCH E-VALUE:** these fields are introduced for setting the specificity of the BLAST program.

BLAST is a program that performs an alignment provided between two sequences trying to find the best match. The lower is the E-value, the more accurate the match looked for. The cumulative is the match of the whole sequence, the SubMatch is for the sub-sequences that the input sequence is split into. So if you don't get many results you can increase these two values to lower the stringency of the sequence alignment.

**BASESTOLERATEDUPSTREAMandDOWNSTREAM:** sometimes in the database the 5' or 3' portion of a gene (or EST) is not perfectly characterized so these parameters allow the extension of the search at the requested number of bases 3' and 5' of the results of the query respect to the user supplied list.

Furthermore, there are articles showing that there are "regulatory sequences" that stand 5' or 3' of the selected gene, so they are technically NOT antisense RNA, but practically are involved in the regulation of a particular gene. Enlarging the search with this parameter allows searching for these sequences too.

**MINIMUM FRACTION OF EST INVOLVED:** represents a ratio - number of bases of the EST involved in the search vs the number of total bases of the EST.

The default value represents a good compromise between the need to discard those matches with only few bases involved from those having only few bases but that could carry useful information.

**OFFSET TO ADD TO RESULTS:** when BLAST performs a search the numbering starts invariantly from 1. Setting this offset allows having a different numbering in the results.

**WORD SIZE:** is another BLAST algorithm setting. Higher values will speed up the search but will decrease the sensitivity.

**BASES SEARCHED TO FIND SPLICING CONSENSI:** splicing is a biological processing of an immature RNA where some specific bases are recognized and cut to produce a mature molecule of RNA which is then carried to the cytosol and translated.

This parameter allows the definition of the number of bases to extend the search in order to find the sites of splicing.

**TREAT ALL KEYWORDS AS SINGLE KEY:** all keywords must be found in the same order as specified.

**USE OR COMBINE KEYWORDS:** all keywords are combined by the logical operator OR

**REPORT ALSO NON DOUBLE CHECKED EST ENTRIES:** normally, AntiHunter reports as result the EST that has both, splicing sites and a "POLY A" tail at 3'. By checking this button the results will include all the ESTs found which do not satisfy this condition.

**REVERSE ANTISENSE SEARCH:** remember that AntiHunter will look for the EST COMPLEMENTARY to the sequence in input (AntiSense mRNA). Selecting this button AntiHunter will report all the REVERSE of the ANTISENSE, so the SENSE EST.

**DON' USE SPLICING CONSENSI IF THE EST STRAND IS MISSING:** sometimes, you don't know the directionality of the strand stored in the EST DB. AntiHunter uses the splicing consensi of the main sequence (that are "asymmetric") to get this information. Here you can disable this function.

**DON'T FILTER THE QUERY SEQUENCE:** AntiHunter applies a program called REPEAT MASKER to the input sequence. This will filter the sequence for "repeated" bases which almost certainly do not belong to a gene (or EST) and that can "confuse" BLAST. Here we can disable this filter and make the search faster. Of course this could mean that you should enter an already filtered sequence. If you chose option 1 (search for a sequence by using the coordinates) the sequences are already filtered.

**EMAIL ADDRESS:** this is optional, you can see the results "live" but if the job you submit is very long, you can receive via email a notification that the job is finished, then login, check and subsequently download the results. The sequence name is the name you assign to give to the session.

Click on submit.

Now the job has been successfully submitted on the CINECA cluster.

You can now click on "check job" and as you can see the job named "... " is executing.

You can refresh and see the LIVE situation or the "View Log" to see a more specific log of what's happening to your job.

When the job finishes, if you provided your email address, you will receive an email with the results.

The resulting output is the list of the anti-sense EST that can interact with the genomic region given in input.

The false positive rate, deriving from erroneous EST strand annotation (5' instead of 3' and *vice versa*), is reduced by incorporating a test for the presence of canonical splicing *consensi* in the reported ESTs and by considering only unspliced 3' ESTs that possess a polyA tail. Other information (such as the length of the spanned genomic region, the length of the EST, the actual splicing sites plus some flanking sequences, etc.) and a quality control check (if there were overlapping genes in the user-annotated genes) are added to the output as well.

## References

1. Wagner E.G., Altuvia S., Romby P. Antisense RNAs in bacteria and their genetic elements. *Adv. Genet.* 2002;46:361–398.
2. Lavorgna G., Dahary D., Lehner B., Sorek R., Sanderson C.M., Casari G. In search of antisense. *Trends Biochem. Sci.* 2004;29:88–94
3. Prescott E.M., Proudfoot N.J. Transcriptional collision between convergent genes in budding yeast. *Proc. Natl Acad. Sci. USA.* 2002;99:8796–8801.
4. Moore T., Constanica M., Zubair M., Bailleul B., Feil R., Sasaki H., Reik W. Multiple imprinted sense and antisense transcripts, differential methylation and tandem repeats in a putative imprinting control region upstream of mouse *Igf2*. *Proc. Natl Acad. Sci. USA.* 1997;9:12509–12514.
5. Sleutels F., Zwart R., Barlow D.P. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature.* 2002;415:810–813.
6. Billy E., Brondani V., Zhang H., Muller U., Filipowicz W. Specific interference with gene expression induced by long, double-stranded RNA in mouse embryonal teratocarcinoma cell lines. *Proc. Natl Acad. Sci. USA.* 2001;98:14428–14433.
7. Li A.W., Murphy P.R. Expression of alternatively spliced FGF-2 antisense RNA transcripts in the central nervous system: regulation of FGF-2 mRNA translation. *Mol. Cell. Endocrinol.* 2000;170:233–242.
8. Munroe S.H., Lazar M.A. Inhibition of c-erbA mRNA splicing by a naturally occurring antisense RNA. *J. Biol. Chem.* 1991;266:22083–22086.
9. Lee J.T., Davidow L.S., Warshawsky D. Tsix, a gene antisense to *Xist* at the X-inactivation centre. *Nature Genet.* 1999;21:400–404.
10. Kumar M., Carmichael G.G. Nuclear antisense RNA induces extensive adenosine modifications and nuclear retention of target transcripts. *Proc. Natl Acad. Sci. USA.* 1997;94:3542–3547.
11. Tufarelli C., Stanley J.A., Garrick D., Sharpe J.A., Ayyub H., Wood W.G., Higgs D.R. Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nat Genet.* 2003 Jun;34(2):157–65.
12. Lavorgna G., Triunfo R., Santoni F., Orfanelli U., Noci S., Bulfone A., Zanetti G., Casari G. AntiHunter 2.0: increased speed and sensitivity in searching BLAST output for EST antisense transcripts. *Nucleic Acids Res.* 2005 Jul 1;33(Web Server issue):W665–8.

## The Job Submission tool (JST)



**Giacinto Donvito, Giorgio P. Maggi, Guido Cuscela**  
INFN, Sezione di Bari, Bari (IT)

### Introduction

Recent developments in bioinformatics have brought to a large increase of the computational power requirements both in terms of CPU time and storage capabilities. The possibility to sequence genomes in a really short time, offers to researchers the opportunity to compare different genomes between themselves or to analyse them with new algorithms. All this activity requires a huge amount of CPU time.

In most cases, however, problems can be solved through the execution of a large number of different and independent small jobs. This is the ideal case for exploiting the potentiality of a grid environment, such as the EGEE/gLite European production infrastructure.

The management of this large number of jobs in a complete distributed environment represents a non trivial problem, since the failure of some of them for whatever reason will require the researcher intervention in order to check their results and to resubmit the failed ones till all the executions needed are correctly terminated.

### The Job Submission Tool in a nutshell

A job submission tool (JST) [1] has specifically been developed to allow the submission of large number of jobs and keep track of all of them in an almost unattended way.

The first step of the JST workflow is to store into a DB (the TaskListDB) server the task list, i.e. all the atomic "task" in which the full problem can be subdivided. The TaskListDB is then used to control the assignment of tasks to the jobs and to monitor the jobs execution (Figure 1).

For the proper monitoring of the task assignment and job termination, each task is described by several parameters:

- the task status - initially it is set to "Free" (not assigned), "Running" (assigned) and "Done" (terminated). If the status is "Free" the task can be assigned to a job. Also if the status "Running" was there for more then a fixed time interval, meaning that it has been assigned to a job which probably has failed, the task can be re-assigned to a new job. If the status is "Done" or is "Running" for less then the fixed time interval, the task is ignored during the tasks assignment process;
- the task dependencies - a tasks may require the execution of a different task before it can be executed (dependency) and only tasks with no dependencies or with all the tasks, from which they depend on, in the "Done" status, can be executed;
- priority - it is possible to assign an arbitrary priority to each task. Priority is used to select the task that has to be executed first;
- job provenance - it is possible to know which grid job has actually performed which task;
- the task description - a link to a specific script to execute on the WN. In this way is possible to change the executable (in case a bug is discovered, or there is a need for a new optimization) also after the submission of the jobs;
- number of failures - if a task fails, the system logs the event in order to avoid the resubmission of always failing tasks. The task then gets resubmitted up to a settable maximum number of resubmissions;
- date and time of the execution.

In the submission phase all the jobs are identical: when a job is submitted it does not know which task(s) it has to execute. A background daemon is used to submit at a given rate, which can be tuned, always the same JDL to the Grid. It is strongly suggested to use more then one daemon in parallel, each one pointing to one gLite WMS, in order to avoid that a failure on a single WMS could stop the submission procedure. The daemons automatically stop the jobs submission when no more unassigned tasks are found in the TaskListDB.

The jobs submitted to the grid are enclosed in a job wrapper. The job wrapper and the script that contains the executable can send to the DB

server any information useful for the monitoring of the running jobs and of the tasks execution. This Monitor DB server (MonDB) can be separated from the one hosting the task list, the TaskListDB, if the scalability of the MonDB itself becomes an issue.

Any kind of information can be stored in the MonDB, its schema looks as follows:

- the date and time on the WN that is sending the information;
- the name of the host that is sending the information;
- the name of the variable to be monitored;
- the value of the variable to be monitored;
- the JOBID of the job that is sending the information;
- the date (and time), on the server, where the information is stored.

This general schema provides the user with the possibility to monitor any job operation by adding new variables as needed without changing the DB schema. If required, this feature can provide information on how many jobs are running at a given time. In this case the job-wrapper should be configured to send regularly monitoring information at fixed time intervals.

Only when a job lands and starts running on a WN, it requests to the central TaskListDB the assignment of a task for its execution. Information on the execution of each task is logged in the central TaskListDB according to the parameters mentioned above. Only if all steps are correctly executed, the status of that particular task is updated to "Done". In this way the TaskListDB provides a complete monitoring of the task assignment and job execution and no manual intervention is required to follow each step and to manage the eventual resubmission in case of failures. Usually we see two kind of failed jobs: the first kind of failures includes cases of jobs that were killed by the queue manager, while the second one is related to internal problem of the job (some operation failed). In the first case the wrapper can not update the TaskListDB, while in the second case the wrapper script updates the TaskListDB increasing the number in the "FAILURE counter".

In order to deal with the first kind of failure, tasks which are found in a "RUNNING" state by more then a fixed amount of time are considered failed and automatically reassigned to a new job.

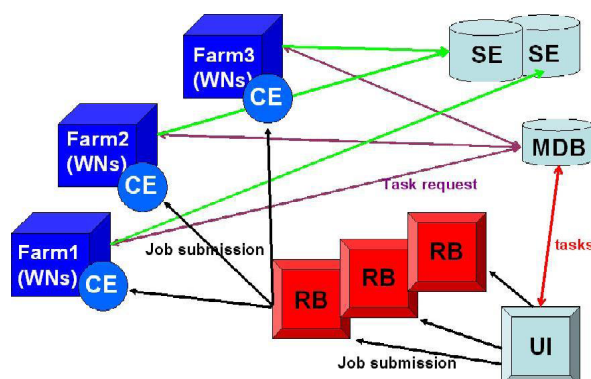


Figure 1. Diagram of the JST.

To optimize the input and output operations and to avoid bottlenecks and failures, two procedures were set up to randomly choose the Storage Element (SE) source of the job input file as well as the SE where to store the job output. In this way if one SE fails temporarily one can continue running the task and store the output on the other SE.

### The JST web interface

JST was used with success for running on the grid several different bioinformatics applications. This encouraged us in providing a simpler way to submit new applications or re-submit the old ones in order to make the intervention of a JST/grid expert unnecessary.

A JST web interface was developed to provide bioinformatics users with a simple way and a guidance to prepare and submit their applications to grid. This interface also exploits the characteristic of the new robot certificates [2] avoiding to users the necessity to hold a valid personal certificate in order to submit their jobs to grid.

Users are required to register to the web interface in order to avoid a misuse of the grid. So far three applications are available on the web interface, namely:

- Gene Analogous Finder (GAF)
- CSTminer
- Clustering

Even if the main goal of the interface is to simplify the use of JST to access the grid infrastructure, other important benefits are provided such as the customization of the application, the monitoring of the tasks and the retrieving of the output. Furthermore, with the adoption of the XSLT mechanism, the inclusion of new applications

requires a very little effort and few changes to the standard interface .

The web interface allows the user to upload the input files (figure 2) and the executable (and eventually the libraries required). The input files are used by JST to split the complete problem in a set of smaller tasks. Normally, at least one of the input files consists in a zipped directory containing a certain number of non overlapping partial input files. For example if one has to investigate a full set of sequences, each file in the input directory should contain a subset of the full list of sequences. JST assumes that it has to execute one task for each file in the directory (or a task for each combination of the files in two different directories). In other words, the user, by properly organizing the application input files, can determine how the entire problem will be split in a number of independent tasks which will be executed, under the control of JST, in parallel over the grid. To have an efficient submission mechanism, the user has to choose the right number of elementary tasks, and in turn the right number of the partial input files in the input directory (directories). The grid jobs, in fact, should not be too long (longer than 24 hours) or too short (shorter than one hour). Long job can fail for any kind of reasons thus reducing the advantage of parallelization. Short jobs (few seconds or few minutes jobs) are not convenient due to the big overheads introduced by the grid submission (WMS latency and the queuing time in the batch queue).

The user can also decide to upload its own version of the application in case he has introduced modification to the standard copy to adapt the application to his specific problems. Furthermore the user is requested to input a criterion (figure 3) that will be used by a post processing script to decide if the run was successful. This parameter is fundamental for JST, it determines if a specific task can be classified as done or not completed yet, influencing all the JST workflow. A particular attention should be put in defining such a criterion, which has to be very stringent: only the tasks with all the steps correctly and successfully executed and that have produced the correct output files should fulfil the criterion. If any of the steps is missing or was not perfectly executed, than the criterion should not be met.

When the web form, with the full description of the application the user intends to execute,

Figure 2. The first page "buid\_task.php" allows users to upload the application input files.

gets submitted, JST creates all the files, which are necessary to submit the job to the grid:

- the JDL file (the file that describes the requirements of the jobs);
- the XML file (containing all the info related to application description);
- the "Job wrapper" generated from xml (is the executable that will run on the WNs).

Another important feature is that the web tool provides users with a complete monitoring of the application submitted with a set of charts (figure 4) and detailed info about failed jobs. In this way either users or developers could easily be aware of what is happening (i.e. if something went wrong in the tasks creation and execution).

The last phase of the JST submission procedure is the retrieval of the output files produced during the full challenge execution. When the execution of all the tasks is completed, the user will be notified by JST with a mail providing the link to the location where he will find all the output files generated by the application.

## CSTminer application

During the tutorial were also presented some practical examples of applications executed with the help of JST.

One of them, named CSTminer, compares

Figure 3. Selection of the type of check to execute on the output files in order to establish if the task has been executed correctly.

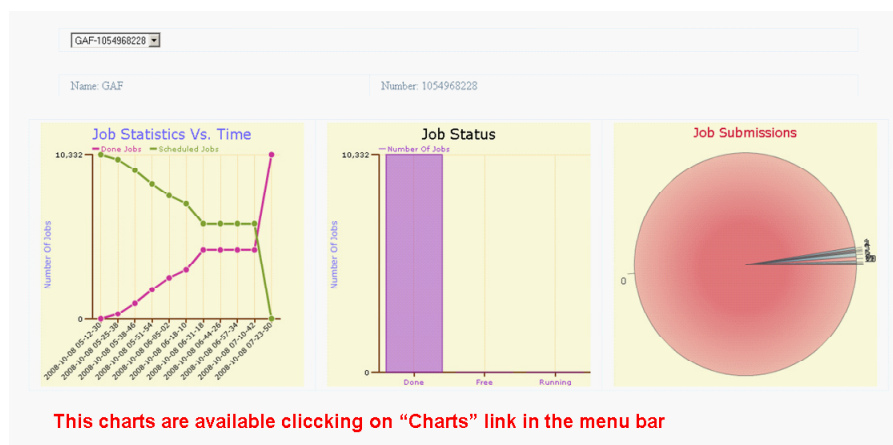


Figure 4. The charts page where it is possible to view how many tasks are executed, how many have left and if there are jobs re-submitted.

two entire genomes to identify conserved tracts and classifies them as coding (likely belonging to protein coding genes) or non-coding (potential regulatory regions).

In order to do the full genome comparison, CSTminer compares the sequences contained in two files, coming from the two different genomes, each containing a slice of the entire genomes. It is therefore possible to split the full genome comparison in a huge number of tasks each one comparing a slice of a given genome with all the slices coming from the other genome.

This is a good application to test the possibility of gLite grid infrastructure, since it is a CPU efficient application with small input and fairly small output.

In order to submit it using the JST web interface, it is enough to submit the two input files, each one consisting in a zipped directory and containing respectively files with the slices of the two genomes. It is up to the user to decide how many slice to build and how big they should be.

Obviously since we have thousand of CPU in a grid infrastructure it is better to provide several slices, but keeping in mind that it is not useful to have a run (a task) that takes less than one our of processing in order to avoid waste resource time during submission and preparation phase.

## Conclusions

Thanks to this new web interface, JST represents a very simple way for a biologist (but easily the interface can be adapted to the needs of other user communities) to get access to the grid and exploit its potentiality in the solution of his research challenges. In fact the user is allowed

to submit a complex application that only the grid could solve, without the need of having a deep knowledge about the grid technicalities. Furthermore the user could re-submit the same application with different data any time he wants or with different customizations (new executable, new command line etc.).

Finally, it is very important to notice that the implementation done with the XSLT and XML allows this tool to be included in every portal or high-level services developed in order to speed up the integration process of a new application.

## Acknowledgements

We would like to thank Rohit M. Dhamane for providing the JST monitoring tool.

### Web Site:

<http://webcms.ba.infn.it/cms-software/index.html/index.php/Main/JobSubmissionTool>

### Information:

giacinto.donvito@ba.infn.it, guido.cuscela@ba.infn.it, giorgio.maggi@ba.infn.it

## References

1. JST – Job Submission tool - Giulia De Sario, Andreas Gisel, Angelica Tulipano, Giacinto Donvito, Giorgio Maggi, High-throughput GRID computing for Life Sciences, in Mario Cannataro (Ed.), Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine and Healthcare, IGI Global (to appear)
2. Donvito G. and La Rocca G. - Authenticated Grid access with robot certificate and the GENIUS Grid Portal - EMBnet.news, 14 (4), 18-20.

## Gene Analogue Finder: a GRID solution to find functional analogous genes



Giulia De Sario, Angelica Tulipano, Andreas Gisel

Institute for Biomedical Technologies, CNR, Bari (IT)

### Introduction

**Engine** is an algorithm for finding, within the same or different species, functional analogous gene products [1]; these are gene products with similar functions but not necessarily similar sequences.

Usually researchers compare genes by sequence similarity, but similar function does not always correspond to similar sequence; to find functional analogies between gene products, it is necessary to compare them according to the information of their function within the gene description.

Gene Ontology [2] (GO) offers a controlled vocabulary for the description of the gene products: the molecular functions they have, the biological processes they are involved in and the cellular components they are associated to. In the GO database (GODB [3] of 06/08) there are more than 4.000.000 gene products described

by more than 26.300 GO terms creating more than 18.000.000 of associations. Therefore a full comparison of all the gene products annotated is very data-intensive and time-consuming (a single comparison with each other occupies one CPU for 30-45 min). The GO consortium [4] is continuously (weekly) improving GO knowledge, using new GO terms and describing new gene products; thus the data is continuously changing and increasing.

The huge amount of data to analyse is the main problem of the functional analogue search, in fact, performing an analysis of one of the last versions of GODB (go\_06\_08), the whole search would require more than 180 years on a single CPU. However to be able to profit from each new release of the GODB is very important for our analysis, since each improvement of GODB will directly reflect as an increased precision of our algorithm to propose functional analogues.

It is important for our analysis to reduce the processing time in order to offer the whole search each month simultaneously with the release of new gene ontology. Only providing this regular service we can guarantee the maximal benefit for the end user interested in these data.

### Data analysis

A first approach to reduce processing time is to analyse the data set (GOBD) and search for redundancy. However, before analysing the data set for redundant information we applied a before introduced threshold to eliminate gene products which have very low level and few GO terms associated [1]; all gene products which have less than 15 directly or indirectly associated GO terms are excluded from the functional analogue search. 60,6% (2.439.488) of all gene

Ontologies	
	GO:0005741; Cellular component: mitochondrial outer membrane ( <i>inferred from direct assay from UniProtKB</i> ).
	GO:0031965; Cellular component: nuclear membrane ( <i>inferred from direct assay from UniProtKB</i> ).
	GO:0051434; Molecular function: BH3 domain binding ( <i>inferred from physical interaction from UniProtKB</i> ).
	GO:0002020; Molecular function: protease binding ( <i>inferred from direct assay from UniProtKB</i> ).
	GO:0046982; Molecular function: protein heterodimerization activity ( <i>inferred from physical interaction from UniProtKB</i> ).
	GO:0042803; Molecular function: protein homodimerization activity ( <i>inferred from direct assay from UniProtKB</i> ).
	GO:0006916; Biological process: anti-apoptosis ( <i>inferred from direct assay from UniProtKB</i> ).
	GO:0051607; Biological process: defense response to virus ( <i>inferred from direct assay from UniProtKB</i> ).
	GO:0007565; Biological process: female pregnancy ( <i>non-traceable author statement from UniProtKB</i> ).
	GO:0006959; Biological process: humoral immune response ( <i>traceable author statement from UniProtKB</i> ).
	GO:0032848; Biological process: negative regulation of cellular pH reduction ( <i>inferred from direct assay from UniProtKB</i> ).
GO	GO:0051902; Biological process: negative regulation of mitochondrial depolarization ( <i>traceable author statement from UniProtKB</i> ).
	GO:0051402; Biological process: neuron apoptosis ( <i>traceable author statement from HGNC</i> ).
	GO:0046902; Biological process: regulation of mitochondrial membrane permeability ( <i>inferred from sequence or structural similarity from HGNC</i> ).
	GO:0000074; Biological process: regulation of progression through cell cycle ( <i>inferred from direct assay from UniProtKB</i> ).
	GO:0043497; Biological process: regulation of protein heterodimerization activity ( <i>inferred from direct assay from UniProtKB</i> ).
	GO:0043496; Biological process: regulation of protein homodimerization activity ( <i>inferred from direct assay from UniProtKB</i> ).
	GO:0001836; Biological process: release of cytochrome c from mitochondria ( <i>non-traceable author statement from UniProtKB</i> ).
	GO:0010039; Biological process: response to iron ion ( <i>inferred from direct assay from UniProtKB</i> ).
	GO:0035094; Biological process: response to nicotine ( <i>inferred from direct assay from UniProtKB</i> ).
	GO:0009314; Biological process: response to radiation ( <i>non-traceable author statement from UniProtKB</i> ).
	GO:0009636; Biological process: response to toxin ( <i>inferred from direct assay from HGNC</i> ).
	QuickGo view.

Figure 1. List of non redundant GO terms in case of BCL2\_HUMAN.



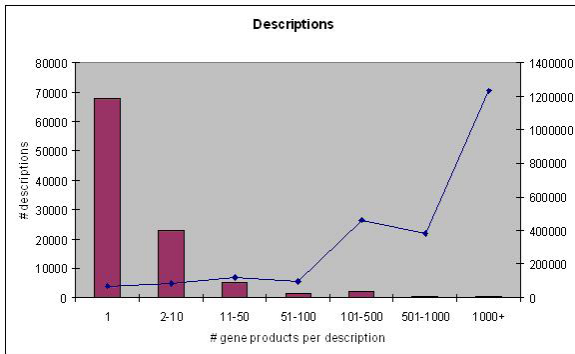


Figure 2. Distributions of descriptions and gene products.

products within GODB are above the defined threshold and entered the functional analogue search. Analysing the remaining data, we observed that many gene products have the identical GO terms associated. For further analysis we define a non-redundant list of directly and indirectly GO terms (Fig.1) to a given gene product as "the description". These 2.400.000 gene products within GODB are described by only 100.029 descriptions.

At this point we do not need to compare each gene product with each other but we can analyse the data set by comparing descriptions with each other.

The description is an information form we are able to analyse the quality of the information within GODB. As we mentioned before there are

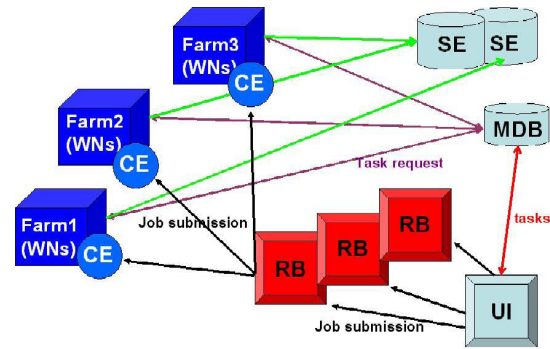


Figure 4. Scheme of GRID distribution.

many gene products which have the same GO terms associated or in other words, which have the same description. From these 100.029 total descriptions 67,8% (67.850) describe only one gene product and therefore are unique descriptions (Fig. 2, bars). Assuming that descriptions which describe up to 50 gene products can be considered still as good quality descriptions we are covering 95,8% of the total descriptions. Therefore we have a large number of high quality descriptions. However, if we visualize the number of gene products described by these good quality descriptions we have to admit that these 95,8% good descriptions represent only 11,1% (271.091) of the analysed gene products (Fig.2, line). All other 2.170.000 gene products are represented by lower quality descriptions. This still high percentage of gene products with low level

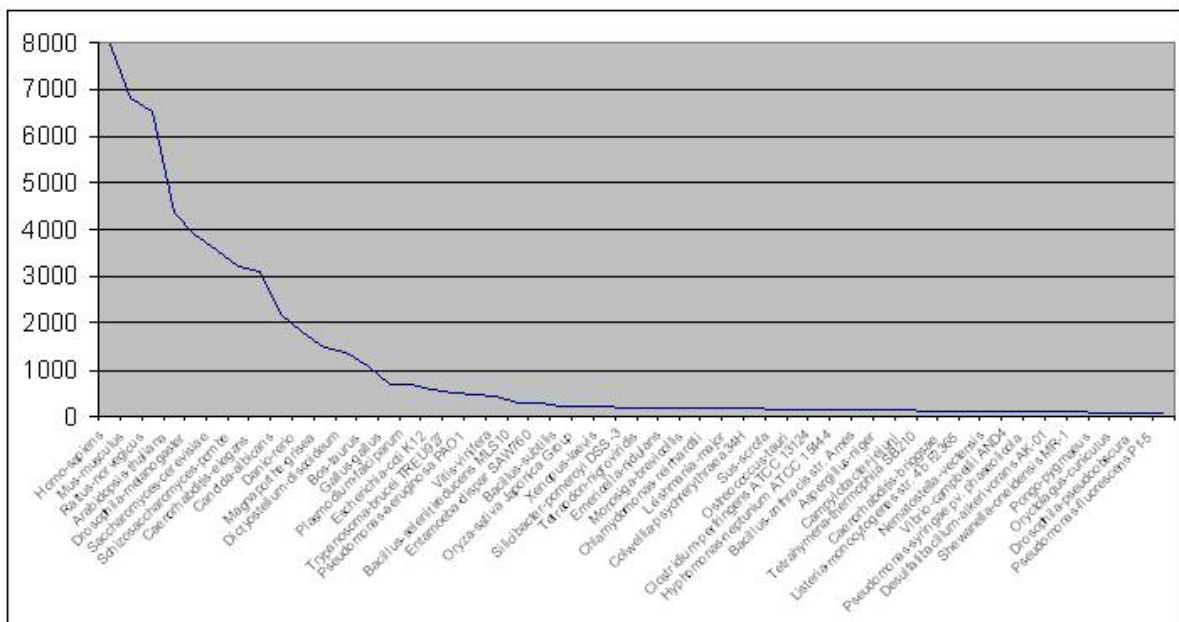


Figure 3. Distributions of descriptions and organisms.

Table 1. Example - BCL2\_HUMAN.

Functional Analogue	Protein family	$\chi^2$ - value	Number common terms	Number non-common terms	Organism	Common name
BCL2_HUMAN	bc12	25789,79	193	0	Homo sapiens	human
Q1WVY1 BORBR		22634,87	170	0	Bordetella bronchiseptica	bacteria
Q1WV3LD BORPA		22634,87	170	0	Bordetella parapertussis	bacteria
Bc12	bc12	20308,28	154	1	Rattus norvegicus	Norway rat
BCL2_CHICK	bc12	13206,1	101	0	Gallus gallus	chicken
Q923R6 CR1LO	bc12	11717,28	90	0	Cricetulus longicaudatus	hamster
Q81008 FELCA	bc12	11717,28	90	0	Felis catus	domestic cat
BCLX_HUMAN	bc12	10374,6	84	4	Homo sapiens	human
Bc12	bc12	9814,676	76	0	Mus musculus	house mouse
BAK_HUMAN	bc12	8649,128	73	6	Homo sapiens	human
BNIP3_HUMAN		8332,739	99	49	Homo sapiens	human
Bnip3		8332,739	99	49	Rattus norvegicus	Norway rat
Q9M2S6 SHEEP	bc12	8053,33	65	2	Ovis aries	sheep
IN12_HUMAN		7954,163	66	4	Homo sapiens	human
debd		7844,056	68	8	Drosophila melanogaster	fruit fly
Q88Q43 FELCA	bc12	7763,406	99	63	Felis catus	domestic cat
Q8HYU5 CANFA	bc12	7763,406	99	63	Canis lupus familiaris	dog
BAXB_HUMAN	bc12	7763,406	99	63	Homo sapiens	human
Nf1		7613,217	115	108	Rattus norvegicus	Norway rat
Bar	bc12	7303,738	102	80	Mus musculus	house mouse
LIRB4_HUMAN		7094,369	63	8	Homo sapiens	human
EIN2		3235,432	46	37	Arabidopsis thaliana	plant
NARJ_ECOLI		2902,651	90	253	Escherichia coli K12	bacteria

descriptions is the major reason why we want to catch each upgrade of the GODB so that we can follow the steady improvement of these gene products.

Having a closer look to the descriptions with only one gene product associated, it is obvious that most of them are originating from model organisms such as human, mouse, rat, Arabidopsis, fruit fly, yeast (Fig. 3). The 20 organisms with the highest number of unique descriptions cover 75,0% (50.875) of the unique descriptions. However, in total we found 2631 organisms that have at least one unique description which is an important fact proving that we have already almost 17.000 descriptions from a knowledge outside the model organisms and this is very often unique information increasing the value of engine.

This non-redundant list of GO information reduces the functional analogue search from 4 million x 4 million to 100000 x 1000000 comparisons; this reflects a reduction of a factor 1600 on the level of number of comparisons only by analysing the data and adjusting the functional analogue search. On the level of CPU occupation we remain with a calculation time of about 1,5 CPU years which is still far too long for a regular update of the gene analogue search data.

In order to solve completely the time problem and reduce the processing to a significantly shorter period than the update time of GODB, we developed a system to divide the full search in a large number of (independent) jobs and to submit those jobs to the GRID infrastructure. In fact the search algorithm allows to examine short lists of descriptions instead of considering the whole list. A tool which submits and monitors the required jobs, the Job Submission Tool (JST) [5] (Fig. 4), processed successfully the full list of descriptions and guaranteed an analysis of the data without missing any information.

## Results

Using the "descriptions" approach and distributing the search on the GRID infrastructure (INFN and EGEE) we were able to terminate the whole search in less than a day.

We have splitted the search in jobs running 20 confrontations each job (each job requires about 2,5 hours including the queue time) resulting in about 5000 jobs processed using 950 different WNs all over Europe and using up to 600 WNs in parallel. The scale of time reduction represents one of the main goals of our research, also for future perspectives, in order to update monthly the GODB with new functional analogue data.

This method finds most of the orthologous gene products and members of the same gene family, but also finds functional analogous gene products not belonging to the same family with sequence low similarity but a high number of common GO terms and sharing therefore similar functions (Table 1).

BCL2\_HUMAN [UniProt: P10415], a well studied player in apoptosis. Table 1 lists the 23 analogous gene products of BCL2\_HUMAN. Firstly, BCL2\_HUMAN was found as the only best hit and therefore is a well-described gene product. Within the 23 best hits we found 12 (54,5%) other gene products which, according to the protein family database (Pfam) belong to the same family, the Bcl-2 family. The rest, 10 gene products (45,5%), belongs to other protein families.

This demonstrates how the engine can find members of the same or of a similar protein family, which is an important confirmation of our algorithm. It is important to mention that members of gene families within Pfam are assigned by sequence similarity and therefore can be found by sequence based approaches. Focusing on the Bcl-2 family, engine found 3 out of 6 BCL2 orthologous gene products present in the SwissProt database with high sequence similarity. Most Bcl2 family members, however, have a significantly lower-than-maximum  $\div 2$  - value due to the fact that via sequence similarity they inherited only the general GO terms. **Engine** has produced a list of potential functionally analogous relations between gene products within and between species using, in place of the sequence, the gene description of the GO. These data are publicly available through engineDB at the url <http://spank.ba.itb.cnr.it/engine>.

## Conclusions

We have demonstrated that, with **engine**, we are able to find most of the gene products' orthologues, but also gene products not belonging to the same gene family and therefore gene products with low sequence similarity which would not have been found with the traditional sequence similarity approaches.

Although we were showing that the descriptions we were using for the analysis are of a good quality we are unfortunately covering only about 10% of the gene products in the analysis. Further it was demonstrated and obvious that most of the high quality descriptions are coming from

the well studied model organisms. However, we have a significant percentage of good quality descriptions of gene products not belonging to model organisms representing often new information about functionality and therefore an important information for **engine**. With time these descriptions from genes products of non-model organism will steadily increase and improve the search results of **engine**. For this reason we were aiming for a solution where we can use each new release of GODB to calculate the functional analogues and provide the information of the last knowledge about gene products.

In future we will offer MySQL dumps for the distribution of the results of the functional analogue search for each GODB release and we are working on an exhaustive web interface to access and download functional analogue data.

## References

- 1 Tulipano A, Donvito G, Licciulli F, Maggi G and Gisel A - Gene analogue finder: a GRID solution for finding functionally analogous gene products - BMC Bioinformatics 2007, 8:329- 342
- 2 Gene Ontology Consortium - The Gene Ontology project in 2008 - Nucleic Acids Res. 2008 Jan;36(Database issue):D440-4.
- 3 Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R - The Gene Ontology (GO) database and informatics resource Nucleic Acids Res. 2004 Jan 1;32(Database issue):D258-61.
- 4 Gene Ontology Consortium - The Gene Ontology (GO) project in 2006. Nucleic Acids Res 2006, 34: D322-D326
- 5 JST - Job Submission tool - Giulia De Sario, Andreas Gisel, Angelica Tulipano, Giacinto Donvito, Giorgio Maggi, High-throughput GRID computing for Life Sciences, in Mario Cannataro (Ed.), Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine and Healthcare, IGI Global (to appear)

## GRID distribution supporting chaotic map clustering on large mixed microarray data sets



Angelica Tulipano<sup>1</sup>, Carmela Marangi<sup>4</sup>, Leonardo Angelini<sup>3</sup>, Giulia De Sario<sup>1</sup>, Giacinto Donvito<sup>2</sup>, Giorgio Maggi<sup>2,3</sup>, Andreas Gisel<sup>1</sup>

<sup>1</sup> CNR, Istituto di Tecnologie Biomediche, Bari (IT)

<sup>2</sup> INFN, Bari (IT)

<sup>3</sup> Dipartimento Interateneo di Fisica, Università degli Studi e Politecnico di Bari (IT)

<sup>4</sup> CNR, Istituto per le Applicazioni del Calcolo, Bari (IT)

### Introduction

Microarray data are a rich source of information containing the expression values of thousands of genes. The analysis of the expression profile of a cell or a tissue allows to detect its state. It is quite clear that genes showing similar temporal and spatial expression patterns are governed by a common regulatory logic. A vast amount of results of various microarray experiments are already available in public repositories. This allows to combine and compare different data sets of expression values obtained with the same chip design but in different conditions and labs. Studies on this huge amount of heterogeneous information are an interesting and important approach, which can bring new insights in genes behaviour to discover new knowledge about them. This task requires an appropriate data normalization process and, even more important, an efficient method of analysis, such as clustering. Due to the heterogeneity of the data sets, supervised and parametric clustering methods are unsuitable since *a priori* nothing is known about the data structure and its distribution. The simplest way to analyse large and mixed data sets without any loss of information is an unsupervised and non-parametric clustering method, which does not require assumptions on the

data neither a cut above a fixed threshold expression value. This can offer us the possibility to discover unknown correlations between genes or unexpected behaviours in different experimental conditions. Of course an unsupervised method of analysis of a large non-restricted data set produces a lot of noise requiring a very accurate procedure of validation of the clustering results. Resampling, based on a cross-validation method, is an efficient way to evaluate the clustering results, telling us if they are due to a really strong correlation between genes or if they are due to statistical fluctuations or noise.

### Clustering of microarray data

We selected 587 data sets covering 24 heterogeneous biological experiments from a collection of experiments of the Affymetrix microarray 'Human Genome U133 Array Set HG-U133A' related to 22215 genes. All this data has been organized in an expression matrix  $D = n_g \times n_s$ , where  $n_g$  (=22215) is the number of genes and  $n_s$  (=587) is the number of samples. We did not set any threshold for the expression values, considering informative the whole distribution of the data, from the lowest to the largest value.

The simplest way to analyse large and mixed data sets without any loss of information is an unsupervised and non-parametric clustering method, which does not require *a priori* assumptions on the data, neither a cut above a fixed threshold expression value. This can give us the possibility to discover unknown correlations between genes or unexpected behaviours in different experimental situations. We have chosen a hierarchical algorithm which was mutated from the mechanical statistics, the chaotic map clustering algorithm, CMC[1]. CMC does not utilize the distances directly for data partitioning, but it generates 'chaotic' trajectories by assigning a dynamical variable (i.e. a chaotic map) to each data point. Pairs of maps are then coupled by means of a decreasing function of the distance of the corresponding data points. The mutual information between pairs of maps, in the stationary regime, is then used as the similarity index for clustering the data set. By setting different threshold values for the mutual information, a hierarchical partition of the data can be obtained as a tree of clusters. The major problem of a hierarchical clustering algorithm is the lack of a unique criterion to choose the optimal level

of the partition. At this aim we used the modularity[2], a measure of the quality of the division in clusters, looking for its peak in the tree of clusters. We tested the use of the modularity analyzing a data set with a well known classification[3]. We observed that varying a local scale parameter of the algorithm the peak of modularity in the hierarchy of clusters always occurred at the hierarchical level of maximal efficiency. This confirmed that the modularity is a good indicator for the selection of the optimal level of the data set partition. Moreover, we tested the CMC algorithm against another hierarchical algorithm, the Deterministic Annealing (DA). For the CMC we found a maximal efficiency of 87% versus a maximal efficiency of 73% for the DA.

### Cluster validation

Of course this unsupervised method of analysis of a full non-restricted data set, such as the matrix of data 22215x587 we have constructed, produces *per se* a lot of noise which requires a very accurate procedure of validation of the clustering results. We have to be able to evaluate the quality of the results and to find the optimal setting of the clustering procedure. Resampling, based on a cross-validation method [4], is an efficient way to evaluate the clustering results, telling us if they are due to a really strong correlation between genes or if they are due to statistical fluctuations or noise.

In the resampling procedure subsets of the data matrix under investigation, with size  $fN \times S$ , where  $0 < f < 1$  is the reduction factor, are constructed randomly, and the clustering algorithm is applied to each subset. From these results we created a connectivity matrix  $T_{ij}, fN \times fN$  for each resampled matrix, whose elements are:

$$T_{ij} = \begin{cases} 1 & \text{points } i \text{ and } j \text{ belonging to the same cluster} \\ 0 & \text{otherwise} \end{cases}$$

and we compared it to the connectivity matrix of the original data matrix.

Starting from the overlap of the original and resampling connectivity matrices we may define three quantities [5], namely "sensitivity", "specificity" and "positive predictive value", which can be regarded as useful "quality measures" of a clustering result.

To define these quantities we consider the results obtained on the full size data set as the "truth". According to the "truth" we have two

classes: either  $ij$  are in the same cluster (positive) or not (negative). Then we compare the results obtained through resampling.

In Tab. 1 we collect all possible combinations: true positive (TP),  $ij$  are in the same cluster both in the original and in the resampled data set; false negative (FN),  $ij$  are in the same cluster in the original data set but not in the resampled one; true negative (TN),  $ij$  are not in the same cluster both in the original dataset and in the resampled one; false positive (FP),  $ij$  are in the same cluster in the resampled matrix but not in the original matrix.

Tab.1 Comparison of the connectivity matrices.

Classes	positive	negative
$ij$ in the same cluster	TP	FN
$ij$ not in the same cluster	FP	TN

It is now possible to define the positive predictive value as the average, with respect to the resamples, of the number of TP pairs divided by the number of the pairs belonging to the positive class

$$ppv = \left\langle \frac{N_P}{N_P + N_{FP}} \right\rangle \quad (1)$$

with similar notation the sensitivity is defined as

$$sens = \left\langle \frac{N_P}{N_P + N_{FN}} \right\rangle \quad (2)$$

and specificity is defined as

$$spec = \left\langle \frac{N_{TN}}{N_{TN} + N_{FP}} \right\rangle \quad (3).$$

Evaluating these quality measures we are able to identify the stable clustering solutions, which are less likely to be the results of noise or fluctuations and we are also able to evaluate the efficiency of the set of resolution parameters used for the clustering.

### Grid distribution

To validate the clusters obtained by applying CMC clustering algorithm to the original matrix 22215x587, 50 randomly resampled matrices

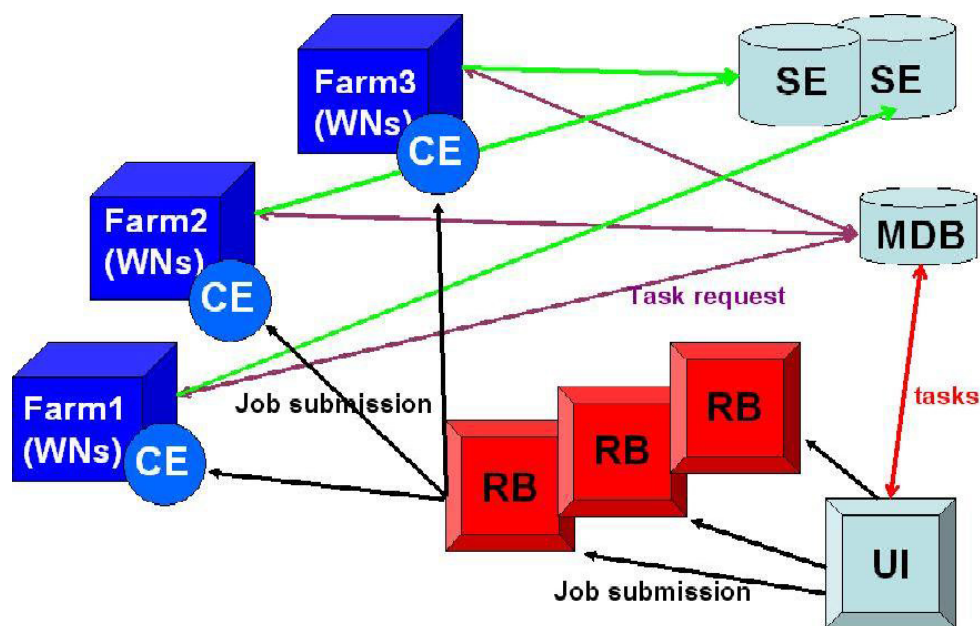


Figure 1. Distribution of the processing over the GRID nodes.

16661x587, with a reduction factor of 25%, were generated. Each matrix was then clustered using CMC with the same set of resolution parameters.

Clustering a single matrix of such a size with CMC is an intensive computational process which requires more than 1.5 GigaBytes of RAM and takes about 2 hours of computing time (one CPU Xeon 3.0GHz). Clustering the whole set of resampled matrices would occupy one single CPU for about 4 days. In this way the clustering validation would be a slow and inefficient procedure. Splitting the whole clustering process of the resampling matrices into several jobs and distributing them on several CPU's speeded up the whole validation procedure giving us the possibility to quickly evaluate the quality of our clustering.

Due to the enormous quantity of computer resources available, the Grid gives the possibility of splitting a large, complex application into many smaller jobs that can run in parallel, greatly reducing the time needed to reach the final results. Our problem of the resampling process is easily dividable in many smaller processes namely every randomized matrix of the resampling process can be launched as an independent job.

The Job Submission Tool (JST) [6], developed to submit and monitor a large number of jobs in the range of hundred thousands in an almost unattended way, was used for the submission

of a small test set of 50 randomized matrices of the size 16661x587. The entire set of resampled matrices was analysed using the grid (Fig.1). As mentioned above each matrix was processed on one WN of the EGEE infrastructure within the virtual organization (VO) Biomed; in this way the analysis was completed in about 3 hours instead of one week. This is a time slightly longer than the time needed for one single run since the grid introduce a certain processing delay due to the queuing time. 9 matrices had to be resubmitted because of a number of different problems. One of them was that not all WN where 32 bit machines and therefore the FORTRAN code was incorrectly executed. Another problem was due to the memory requirement of the clustering algorithm for a matrix of the specified size, more than 1.5 GB of RAM, and this is a requirement which is not satisfied by all the WN available on the Grid.

The results obtained in terms of average values of *ppv*, *spec* and *sens*, as calculated for the given data set, are 0.65, 0.95 and 0.81.

Due to the size of the dataset and the high level of fragmentation (more than 300 clusters) we may expect that the number of true negatives is by orders of magnitude greater than the other quantities defined in Tab. 1. That means that for the case at hand the *specificity* values would be close to 1 even in the case of random or incorrect clustering results. To our purposes we

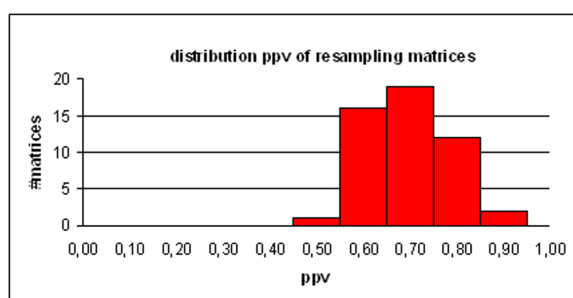


Figure 2. Distribution of the values of *ppv* obtained for the resampled matrices.

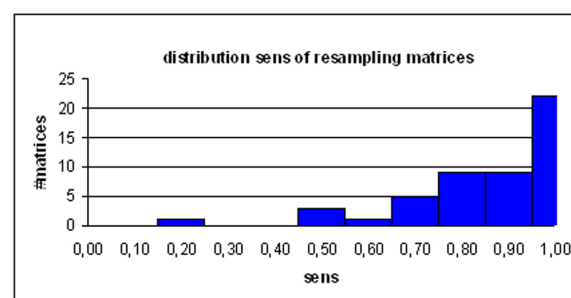


Figure 3. Distribution of the values of *sens* obtained for the resampled matrices.

then consider *ppv* and *sens* as the only relevant quality measures.

To illustrate the results in more details, in figures 2 and 3 we show the distributions obtained for the values of *ppv* and *sens* for the 50 resampled matrices. We can see from the histograms that about a half of the resamples exhibits sensitivity values above 0,95, and that in almost all the cases the *ppv* values are above 0,5. Since it is well known that statistical significance of the quality measures is strongly affected by the size of the data set and by the level and nature of noise in the data, it has long been recognized that there is a substantial intrinsic noise contained in microarray data, we stress that the values obtained are far above what can be considered a good result in such a context and we can conclude that the results of the clustering validation suggest that the approach here proposed is adequate.

## Conclusion

The approach described above, i.e. to distribute on the Grid the validation process of the clustering results, demonstrates to be a very valid implementation for large microarray data sets. The analysis performed with the CMC algorithm is a very complex procedure, requiring an efficient validation method of the results. Since the matrices are so large and the clustering algorithm is quite computational intensive, every resampled matrix was submitted to an independent WN and therefore the total validation was run in parallel terminating the calculation in a fraction of the original time using only one CPU.

We demonstrated that using the GRID infrastructure we are able to validate efficiently and accurately large amount of clusters and that CMC is a potential clustering algorithm for large and heterogeneous microarray data. Moreover this is an application where we clearly differenti-

ated when it is useful to use the GRID infrastructure and when to use local resources: for the analysis of the complete data set we applied the CMC algorithm using a local machine, while for the resampling validation procedure we employed the grid resources speeding up drastically the execution time.

Now we are working on an automatic procedure to perform the analysis. In the future we will enhance the number of data sets from heterogeneous experiments and we will also use a bigger set of resampling matrices to improve the statistical validation. This will require a larger number of worker nodes to distribute the jobs.

## Bibliography

1. L. Angelini, F. De Carlo, C. Marangi, M. Pellicoro and S. Stramaglia, Clustering Data by Inhomogeneous Chaotic Map Lattices, *Phys. Rev. Lett.*, Vol. 85, No. 3, pp 554-557 (2000).
2. Newman M.E.J., Girvan M, Finding and evaluating community structure in networks, *Phys. Rev. E*69, (2004)
3. Khan J, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nat. Med.*, Vol. 7, No. 6, pp 673-679 (2001).
4. Levine E. and Domany E., Resampling method for unsupervised estimation of cluster validity, *Neural Comp.*, Vol. 13, pp 2573-2593, (2001).
5. Van deer Laan, M.J., Bryan, J., Gene expression analysis with the parametric bootstrap, *Biostatistics*, Vol. 2, No. 4: pp 445-461, (2001).
6. <http://webcms.ba.infn.it/cms-software/index.html/index.php/Main/JobSubmissionTool>

## Successful EMBnet-EMBRACE joint Webservices workshop



**Erik Bongcam-Rudloff**

The Linnaeus Centre for Bioinformatics, SLU, Uppsala Sweden



Italian and Belgian nodes at work.

EMBnet and EMBRACE organized the 3th and 4th of November, 2008 a joint workshop in Uppsala-Sweden. The title of the workshop was : " Understanding, creating and deploying EMBRACE compliant WebServices".

The goals of the workshop were: to understand what "EMBRACE compliant" means, to create webservices to databases and tools maintained by the EMBnet community (and their collaborators), how to deploy them and finally how to add the new webservices to the EMBRACE registry located at: <http://www.embraceregistry.net/>.

More about the EMBRACE registry is reported in the Steve Petiffer's article on this issue of EMBnet.news.

The workshop was a great success and gathered together more than 30 participants from 13 different countries including representatives from 11 EMBnet nodes (BE, UK, CH, GR, SK, SE, HU, IT, NO, PK, FI).

The workshop was also visited by Melanie Hilario and Alexandros Kalousis from the newly accepted project e-LICO: "An e-Laboratory for Interdisciplinary Collaborative Research in Data

Mining and Data-Intensive Sciences", starting in 2009-02-01. e-LICO is an FP7 Collaborative Project (STREP) under the theme ICT-4.4: Intelligent Content and Semantics.

A good measure of the success of the Workshop is the fact that more than 20 webservices were deployed on the EMBRACE Webservices Registry during the two days work! This good result encouraged most participants to continue the submission of EMBRACE compliant WebServices to the registry.

### The programme of the workshop was :

#### Monday 3 November

**10:00** - Soap Lab and EBI SoapLab Services (Mahmut Uludag)

**10:45** - Bioclipse and Webservices (Egon Willighagen)

**11:30** MRS updates (Maarten L. Hekkelman)

#### Lunch

**14:00 - 14:45** EMBRACE Webservices Registry (Steve Petiffer)



The participants of the workshop.





Hands-on tutorials by EBI.

**14:45 - 15:30** EMBRACE WP4 test case (Andrew Clegg)

**15:30** Coffee break

**16:00** - Dealing with binary data in the context of Web services (Taavi Hupponen)

**16:45 - 19:00** Hands on Tutorials Webservices (Syed Haider)

**19:30** Dinner

#### Tuesday 4 November

**8:30 - 10:30** Soaplab introduction and hands-on session (Mahmut Uludag)

**10:30** Coffee break

**11:15 - 12:30** Hands on session continuation (Mahmut Uludag)

**12:30 - 13:30** Lunch

**13:30 - 15:00** Hands on Tutorials Webservices (Syed Haider)



Nils-Erikson and Anders Lövgren at the workshop.

The workshop presentations can be downloaded from here: <http://teacher.bmc.uu.se/webservicesworkshop/Presentations.html>

#### Acknowledgements

The EMBRACE/EMBNet workshop was funded by the European Commission within its FP6 Programme, under the thematic area "Life sciences, genomics and biotechnology for health", contract number LHSG-CT-2004-512092.



Steve Pettifer during his talk on the EMBRACE Webservices Registry.

## The EMBRACE Registry



Steve Pettifer,<sup>1</sup>  
 Dave Thorne,<sup>1</sup>  
 Philip McDermott,<sup>1</sup>  
 Terri Attwood,<sup>1</sup>  
 Joachim Baran,<sup>1</sup>  
 Jan Christian Bryne,<sup>2</sup>  
 Taavi Hupponen,<sup>3</sup>  
 Gert Vriend<sup>4</sup>

<sup>1</sup> The University of Manchester (UK)

<sup>2</sup> Computational Biology Unit, Bergen Center for Computational Science (NO)

<sup>3</sup> CSC, the Finnish IT Center for Science, Finland

<sup>4</sup> Centre for Molecular and Biomolecular Informatics, University of Nijmegen (NL)

[www.embraceregistry.net](http://www.embraceregistry.net)

Funded by the European Union's Sixth Framework Programme, The EMBRACE Network of Excellence brings together seventeen institutions including seven EMBnet National and Specialist Nodes, in a concerted effort to provide consistent programmatic interfaces to a multitude of major bioinformatics tools and databases. Now in its fourth year, the EMBRACE partners have produced a growing collection of web services that are freely available for use by the community, and these are now being collected and published via the project's newly released Service Registry. The core function of this publicly-available web site (the front page of which is shown in Figure 1) is to allow service providers to publicise their tools in a form that is easily searched by potential consumers. Moreover, the registry attempts to embody a number of the principles and lessons learned during the early years of EMBRACE.

### Embracing the use of Web Services: lowering barriers to adoption

The primary goal of EMBRACE has been not simply to expose partners' existing tools and resources via web services, but to educate and encourage the entire community in the provision of such resources, and to provide tools to facilitate this task. One of the main lessons learned so far has been that this is not a trivial task: although the theory of web services is perhaps straightforward

enough, in practice, the heterogeneous nature of bioinformatics tools and issues of legacy software present real software engineering challenges. The EMBRACE registry has therefore been set up to accept the registration of web services in a variety of forms, including SOAP-based, REST and DAS services, and even allows the registration of 'arbitrary cgi script' style services. The principle here is that the existence of publicly-available resources, even in a basic form, is better than nothing at all, and that the uptake of such services by the community is likely to inspire the service providers to build more robust and effective interfaces in future.

### Encourage best practice: not all services are created equal

At the same time as wanting to expose existing services of all kinds to a wider audience, the project recognises that some web service technologies are more suitable than others, and aims to encourage providers to adopt 'best practice' techniques in their service provision. The project has produced an extensive technology recommendation ([http://www.embracegrid.info/files/pub/products/D3.1.1\\_v2.0.doc](http://www.embracegrid.info/files/pub/products/D3.1.1_v2.0.doc)), focussing in particular on the use of WS-I compliant technologies to provide robust, interoperable interfaces. The EMBRACE registry includes a number of 'quality control' features that, while accepting services of all types, attempt to highlight and showcase those that follow the more stringent industry-based standards. At the same time, it is intended that the registry should provide automated tools, documentation, user-feedback mechanisms and support via community forums that will encourage all providers to aspire to these standards.

### Automated monitoring

One of the bugbears of working in a distributed computing environment is that automated tasks end up relying on tools and resources provided by institutions located all around the world; the unexpected failure of one of these tools can have dire consequences for an analysis task. Even with today's comparatively reliable network connectivity, the heterogeneous nature of the 'back end' infrastructure that drives most bioinformatics web services means that occasional interruptions to services are inevitable, whether caused by hardware or software problems, or

The EMBRACE Registry

Home Services Documentation Forums My Account EMBRACE Search

All services  
My services  
Top services  
Create service  
My account  
Log out

The EMBRACE Service Registry is a collection of life-science web services with built-in service testing.

This site is a prelude to the internationally supported **BioCatalogue** system that will collect, store, validate, and make available web-services in the biosciences. This registry is mainly meant for the EU projects EMBRACE, BioSapiens and ENFIN, but other users are welcome too. As a potential web service user, you can search or browse the registry for services that match your needs. Furthermore, each entry includes live test data, showing the current and historical status of the service. Each entry can also include example client software to help you include them in your own programs or workflows. As a service provider, the registry helps you build high quality services that conform to industry standards, and gives you a means of advertising your tools to the user community as well as a platform for testing your service. Your service remains your property and published by you – this registry merely advertises its existence and its status.

This system will in due time merge seamlessly into the BioCatalogue, and if you use this registry, you will not have to register your service again with the BioCatalogue.

**Latest service updates**

Service Name	Time Ago	Status Change	Indicator
CLUSTAL W	32 mins ago	status changed to PASSED	Green circle
PairsDB	4 hours ago	status changed to WARNING	Yellow circle
HitKeeper	4 hours ago	status changed to PASSED	Green circle
PRINTS	5 days ago	status changed to PASSED	Green circle
GenomeMatrix	5 days ago	status changed to PASSED	Green circle

**Recent News**

**Having Problems Logging In?**

If you are having problems logging in, make sure that you are using your username, and not your email address, along with your password. Some browsers are a little too keen with autocomplete, and may have put your email address in the username form field.

» [Add new comment](#) Submitted by [admin](#) on Fri, 11/21/2008 - 14:23

**engineDB\_funcnet**

engineDB is a repository for precalculated functional analogue gene products using the gene ontology annotation to calculate the functional analogy by semantic similarity.

engineDB\_funcnet is a web service accessing the engineDB searching for two gene products (UniProt primary accession number) the corresponding raw score and p-value (see <http://funcnet.eu>)

Average rating:  
☆☆☆☆☆

**Current Statistics**

Total Users: 46  
Total Services: 74

Line graph showing growth in users and services over time.

Figure 1. the registry home page. On the right is a table of services that have recently changed their status (e.g. gone from working to broken, or vice versa). At the bottom of the page is a list of news items and recently added services.

routine maintenance. Determining whether a service is working correctly or not at any moment in time has been a real problem. The EMBRACE registry attempts to overcome this by combining automated monitoring mechanisms, which test whether the lowest common factors of services are responding correctly (e.g., is the server currently accessible via the internet?), with high-

level application-specific tests deposited by the service providers (e.g., does this service correctly predict region XYZ on protein ABC). The registry periodically applies these tests, and collects the results to generate easy-to-read reports about the availability and reliability of a service over time.

```

#!/usr/bin/python
# the first line of the script must tell us which language interpreter to use,
# in this case its python

# import the modules we need for this test; SOAPpy is included on the server
# by default, and we'll need the 'sys' module in order to be able to use
# exit to return a value from this script
from SOAPpy import WSDL
from sys import exit

# set up the input values we need for this service...
# an amino acid sequence
seq_str = 'MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPPGGRNRYPPQGGGGWG' \
          + 'QPHGGGWGQPHGGGGWGQPHGGGGWGQPHGGGGWGQGGGTHSQWNKPSKPKTNMKHMAGAA' \
          + 'AAGAVVGGGLGGYMLGSAMSRPIIHFGSDYEDRYRENMHRYPNQVYYRPMDEYSNQNN' \
          + 'FVHDCVNITIKQHTVTTTTKGENFTETDVKMMERVVEQMCITQYERESQAYYQRGSSM' \
          + 'VLFSSPPVILLISFLIPLIVG'

# and the corresponding UniProt ID
ID= 'P04156'

# we'll include the whole main body in a try:except clause so that if something
# unexpected happens, we can return a sensible error value

try:
    # Get web service proxy
    proxy = WSDL.Proxy('http://smart.embl-heidelberg.de/web/service/SMART_webservice.wsdl')

    # Execute the SMART service
    features = proxy.doSMART(protein_sequence=seq_str, protein_ID=ID)

    # Get the features that are returned
    feature_list = features._getItemAsList('feature')

    # Assume that everything will work, so start with a returnvalue of 0 / Ok
    returnvalue = 0

    # we're expecting exactly two features for our input sequence, so test this
    # first
    if len(feature_list) != 2:
        # write a comment to stdout, and return a value of 2. This will be interpreted as
        # a 'warning' status
        print "SMART returned the wrong number of features for this sequence"
        returnvalue= 2
    else:
        # we got exactly two features, so check that the start and end values are
        # exactly what we'd expect for this sequence
        if ((int(feature_list[0].start) != 1) or (int(feature_list[0].end) != 22)
            or (feature_list[0].type != "INTRINSIC")):
            returnvalue = 3
            print "Data Error in first feature"

            if ((int(feature_list[1].start) != 23) or (int(feature_list[1].end) != 240)
                or (feature_list[1].type != "SMART")):
                returnvalue = 4
                print "Data Error in second feature"
except:
    print "Failed to talk to SMART service"
    returnvalue = 1

if returnvalue == 0:
    print "Everything worked fine"





exit(returnvalue)

```

Figure 2: An example test script, written in Python and using the SOAPpy library

## The Traffic Light system

Determining whether a web service is functioning correctly or not turns out to be a more complex task than one might expect. Because the service worked at one moment under a certain set of circumstances using a certain set of support libraries, does not guarantee that it will work later under a different set, and the number of factors influencing such behaviour is enormous. The registry attempts to distil all of this complexity into a simple 'traffic light' system that provides a quick overview of a service's status, as follows:

	Green: the service is working correctly - it should be safe to use this service now.
	Amber: the service is experiencing problems - it may respond, but you should treat any results you get back with caution.
	Red: the service is badly broken - it is very unlikely that you will be able to use this service until the problem is repaired.
	A blue service status icon (admittedly a limitation of the 'traffic light' analogy!) indicates that the service status is unknown, typically because the service provider has not registered sufficient information for regular tests to be carried out.

For users curious to delve deeper into the meaning of the status icons, for example to determine whether a particular problem is likely to affect them in reality, the registry exposes the details of each service's test suite. This includes a log of the behaviour of the tests, and a browser for viewing the actual code used to exercise a service. An example of a typical test program is provided in Figure 2. As well as providing a means of determining whether a service is functioning correctly, the fragments of test code provide the community with a useful starting point for building their own client code against a particular service.

## The future of the registry

In its current form, the EMBRACE registry is an experiment, but one that has already demonstrated some measurable success, enabling service

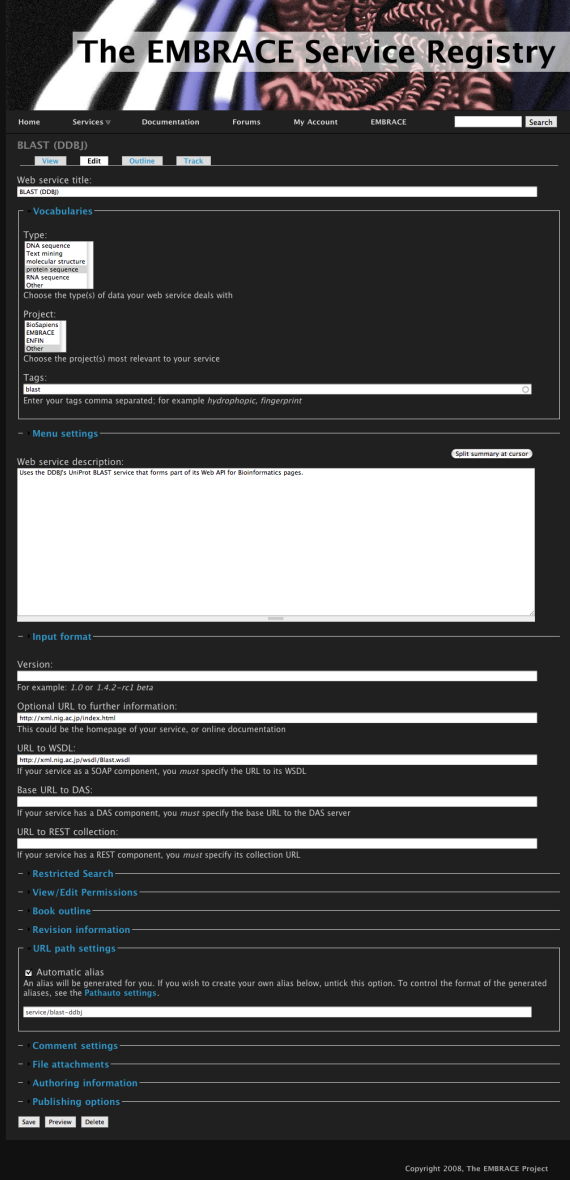


Figure 3. The 'create service' form, with sample inputs.

providers to monitor and improve their provision, even to the extent of exposing bugs in ostensibly 'reliable' services. The UK's Biotechnology and Biological Sciences Research Council has recently funded the BioCatalogue project, which aims to provide a robust long-term solution to the collection and maintenance of life science web services. EMBRACE is working closely with BioCatalogue, aiming to migrate the registry's data and functionality to the Catalogue when it comes on line in early 2009. In the meantime, we encourage you to register your services at <http://www.embraceregistry.net> and di-

rect any questions or problems email to [admin@embraceregistry.net](mailto:admin@embraceregistry.net).

## Acknowledgements

The EMBRACE registry team would like to acknowledge the support of Peter Hallin, from the Center for Biological Sequence Analysis, Technical University of Denmark, during the initial design period, and for the work of EMBRACE's Technology Watch Group, led by Vincent Breton, whose technology recommendations have provided a guiding principle for these developments. We also thank Fred Marcus for his robust support throughout the project. EMBRACE is funded by the European Commission within its FP6 Programme, under the thematic area "Life sciences, genomics and biotechnology for health", contract number LHSG-CT-2004-512092.

### Registering a service

The basic process of registering a service is very easy, and should take only a few minutes. Although the registry is currently under constant development, the details of the interface may well have changed by the time you read this article, but the general principles will remain the same.

#### Create a User

First, create yourself a registry account by following the link underneath the login pane on the front page. You will be emailed your login details, and can request a new password at any time, should you forget it.

#### Register your service!

In the 'services' menu at the top of the page, select 'Create service', and you should be presented with a form similar to that shown in Figure 3. Give your service a sensible name; this will be the main name associated with your service and needs a little thought. Choose something that is descriptive but not too long. The name should capture the essence of the service from a user's point of view. For example, if your database or algorithm has a well known name or acronym, you should include that in the name, as well as a description of what this particular service does. 'PROSITE pattern search' is probably a sensible name; 'prste\_mt\_search1' isn't, and neither is 'Search the PROSITE database of regular expressions using a query sequence in FASTA format'. You can attach additional information later, which will be indexed and searched by the registry, so you don't need to squash everything into this field!

The 'Type' field is intended to capture the primary data-types and functions used by your service. At present, this is quite a limited list, but will be expanded as we learn more about the kinds of service that are being registered. Select from this list one or more types (and please feel free to mail us if something important is missing!)

The 'Project' field only means anything if your service comes from one of the projects in the list. If your project isn't listed, simply select 'other'.

The 'Tags' field is really important. Here you should add any terms that you think describe your service, or would be handy to help users find your service in the first place. It's better to add too many tags than too few, so

include terms that cover the functions of your service, as well as describing its inputs and outputs. You can also add tags for your project, institution, and so on, here.

The 'Description' is your opportunity to document what your service does – this text is also searched by the registry when users look for services. If you have documentation on your own server, you can link to this in the 'Optional URL' field, so it's best here to write an actual description rather than just to provide a link. If you are adding a WSDL-based service, the registry will automatically extract a list of operations/methods, and add them to your description, so there's no need to do this manually (in future, we plan also to extract documentation from the WSDL).

The 'Version' field simply records a version number for your service; here, you can use any naming scheme you like – the text isn't processed by the registry.

The final three fields are vitally important, and you must fill out at least one of them, as they record the actual location of your service. You can include details of WSDL, REST or DAS services at present.

If you'd like to attach a file to the page (e.g., a diagram, additional documentation, a paper, and so on), you can do this via the 'File Attachments' button at the bottom of the page. It is important to note that attaching a file here isn't the same as submitting a test!

In the near future we will also add a wizard for registering services with WSDL descriptions. This wizard will tell you if there is anything about the service that would make it harder for WS-I compliant clients to communicate with the service.

### Add one or more tests

Once you've submitted your service, you can add test scripts to validate and monitor its behaviour. While you are logged in, visit your service's page, and at the bottom of the description you'll see a 'Add Test' link. Following this will take you to a new form via which you can describe and upload your test script.

'Web service test title' should describe your test, and again it should be meaningful from a user's point of view: e.g. 'Check that a sample of three identifiers return the correct records'. The 'description' field should describe in simple terms what the test does – in this example, we may want to say which sample of three identifiers are being submitted, and what kind of results are expected.

The 'Binding' field determines what language you are going to use to test your service; Perl, Python and Java are currently supported. Details of which web service support libraries are installed on the server can be found on the registry's documentation page.

If you are testing a DAS service, all you need to do now is supply a correctly formed DAS request – the registry will then periodically check that the server is returning a suitably formed DAS XML document. When the WSDL wizard becomes available it will generate a test client for you by asking for sample request values. This test will check that the output of your service is consistent with its description in the WSDL file.

You can also upload a simple test client for your service. A detailed description of this is included in the registry documentation, but it should involve no more effort than generating a short test script, zipping it up into an archive, specifying the name of the command to execute in your script, and uploading the package to the registry. Once uploaded, the registry will test your script, and set the status icon accordingly. If there are any problems, you can edit and re-upload your scripts at any time.

## Tutorial “Introduction to Bioinformatics”



**Sophia Kossida**

Biomedical Research  
Foundation of the  
Academy of Athens, Athens,  
(GR)

In connection with the 8th IEEE International Conference on Bioinformatics and BioEngineering (BIBE) 2008 (<http://www.bibe2008.org/>) a tutorial entitled “Introduction to Bioinformatics” was organized by Dr. Erik Bongcam-Rudloff (Linnaeus Centre of Bioinformatics, Sweden) (<http://www.anst.uu.se/erikbr/Welcome.html>), president of the EMBnet (<http://www.embnet.org>) and Dr. Sophia Kossida (BRF, Greece) (<http://www.bioacademy.gr/bioinformatics/>), manager of the EMBnet node in Greece.

The tutorial was held on Wednesday, 8 October, 2008 in co-operation with the Biomedical Research Foundation – BRF ([www.bioacademy.gr](http://www.bioacademy.gr)) of the Academy of Athens. The

main objective of the course was to introduce young scientists to, and encourage application of, bioinformatics/computational biology tools in their research and to present some of the biological resources available on the web.

The course included lectures and hands-on tutorials in the following topics: Biological Databases, BLAST, Ensembl. All necessary material for the tutorial is still available at <http://teacher.bmc.uu.se/BIBE2008>.

The technical programme of the tutorial was as follows:

Wednesday, 8 October, 2008 at the BRF Multimedia and Videoconference room.

08.30 - 09.30 Introduction to Biological Databases

09.45 – 10.30 Information Retrieval

11.00 – 12.00 Hands-on tutorial (Group A)

12.00 – 13.00 Hands-on tutorial (Group B)

The tutorials were transmitted live from the Academy of Athens facilities to the BIBE 2008 conference venue at Hotel, Royal Olympic in central Athens. The tutorials were highly appreciated by the participants.

Dr E. Bongcam-Rudloff also delivered a keynote lecture within the actual conference entitled: “Biobanks, Biomolecular Resources and Bioinformatics for Health Care and Medical Research in Europe”.

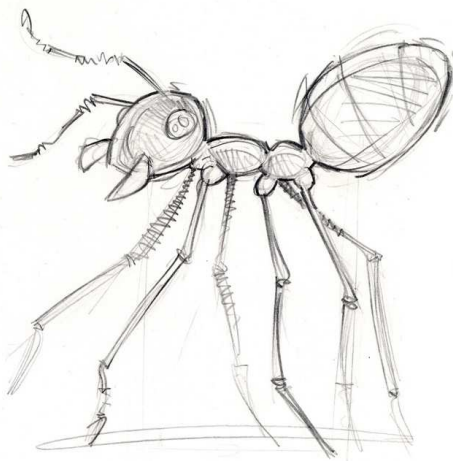


Hands-on tutorials by Erik Bongcam-Rudloff.

## The queen's perfume

Vivienne Baillie Gerritsen

Can a smell affect social behaviour? Without a doubt. Let off an unpleasant one and those closest to you will move somewhere else. Likewise, an agreeable scent will keep them hovering in your vicinity. It's an old trick. Flowers and animals have been using smells for millions of years to ward off predators or to attract individuals for the sake of reproduction. So it does not come as a surprise to learn that ants use the same kind of technique as a means of communication and social interaction. However, it is not so much the odour but the capacity to detect it that is at the basis of two types of social behaviour in a species of red fire ant, *Solenopsis invicta* – the ecological pest. This particular ant either belongs to a colony that has only one queen (monogyne) reigning over it or to a far larger colony which is ruled by several queens (polygyne). In the 1990s, scientists discovered that the basis of a monogyne or a polygyne colony amounted to the existence of only one protein: pheromone-binding protein.



An ant, Nathan Walker illustrations, etc.

Courtesy of the artist

Social behaviour is a complex process. Even in the world of a worm. So can one gene be responsible for a type of royalty? It seems to be the case. What scientists observed is that the *Solenopsis invicta* red ants have two social organisations: colonies with only one queen or colonies with many hundreds. Gp-9 protein had been widely used as a marker to determine population genetic structure in fire ant populations – hence its name: General protein-9. In 1997, a research team noticed that social

organisation seemed to be governed, one way or another, by this particular protein. The question was how?

In an attempt to answer, the scientists collected a number of *Solenopsis invicta* red ants which belonged to mono-queen colonies, and a few from poly-queen colonies. They then had a closer look at their Gp-9 proteins. To cut a long story short, they found two Gp-9 alleles which they called B and b. To their astonishment, they then discovered that the mono-queen *Solenopsis invicta* red ants were all BB homozygotes, while the poly-queen ones were Bb heterozygotes. This implied that the basis of mono- or poly-queen colonies is hereditary. That is to say, the social organisation of ants – with regards to the existence of one or more queens – is orchestrated by a gene.

Gp-9 is in fact a pheromone-binding protein, i.e. it binds these odourless molecules which can govern the behaviour of many creatures. Humans included. Such pheromone receptors are found on the ants' antennae and once its ligand is bound to it, it transfers it to receptors on olfactory sensory neurons, which relay the pheromone message to the brain. And what would be the message such a pheromone relays? In this case, two different messages are relayed depending on the phenotype – BB or Bb – the ant carries. A worker ant will either accept an extra queen ant in its colony or it will not. It sounds simple, yet it is not...



A BB queen has all it needs to found its own colony. It is hugely overweight and can provide for itself and its eggs. Once it has mated with a male ant – in midair if you please – it will fly off and find a nice spot to lay its eggs and look after its brethren, which will subsequently become the workers that will not only look after their queen mother but will also look after the nest and fend for her. All the possible BB queen ants the original queen mother gave birth to will be sacrificed. And if any other BB queen stumbles upon the colony, it will also be killed. This behaviour occurs by way of the ‘BB pheromones’ given off by the BB queens, which signal to the BB worker ants that any other queen must be eliminated. A sorry fate.

In colonies where many queens cohabit, the workers – like the queens – are Bb heterozygotes. How do they behave? It seems that their pheromone-detecting capacities are impaired. As a consequence, Bb red ants can ‘smell’ BB queen ants a mile off and they’ll kill them. However, these same heterozygous ants miss the Bb queens because they can’t detect them. So poly-queen colonies seem to exist simply because Bb ants don’t have the means to smell them. And the bb homozygotes? These ants are not viable – they are also very skinny – and die before they are sexually mature. What is more, Bb queens are far leaner than their BB counterparts and would not be able to form their own colony without the help of workers. So royal chubbiness also seems to be related to Gp-9!

On the molecular level, not much is known about Gp-9. It probably acts as a homodimer.

However, its 3D structure is still misunderstood and besides imagining that the two alleles – B and b – probably engender two different binding structures, their architectural difference is still a blur. A number of the mutations which create the two distinct Gp-9 alleles most probably affect the ligand pocket region. The B allele would be intact, while the b allele would be faulty. This could explain why the BB homozygotes are very good at smelling out the BB and the Bb queens, whereas the Bb heterozygotes can only cope with the BB queens but are not fine-tuned enough to spot the Bb queen pheromones.

Much has yet to be understood but the discovery is paramount and perhaps even a little disturbing. Here is a gene – Gp-9 – that orchestrates social behaviour. Not only does it define whether a red ant will be part of a mono-queen or a poly-queen colony – with all the interactions this implies – but it also plays some part in the chubbiness of the queen. The obvious question is: Is there any form of human social behaviour which is driven by genes solely? Or do we have our say in matters? Does free-will really exist? The usual answer is: Human social interaction is so complex that there is no way that any one form of behaviour could be pinned down to one gene. And the environment also has its say. But before the role of Gp-9 was uncovered, the same was said about ants. The wonderful thing about humans though is that we all own that beautiful talent called imagination which we can use to ensure us of our freedom. Gp-9 or not.

## Cross-references to Swiss-Prot

Pheromone-binding protein Gp-9, *Solenopsis invicta* (Red imported fire ant) : Q8WP90

## References

1. Krieger M.J.B.  
To b or not to b: a pheromone-binding protein regulates colony social organization in fire ants  
BioEssays 27:91-99(2004)  
PMID: 15612031
2. Krieger M.J.B., Ross K.G.  
Identification of a major gene regulating complex social behavior  
Science 295:328-332(2002)  
PMID: 11711637

# 5<sup>th</sup> RIB (Red Iberoamericana de Bioinformática) Congress



## Programme and Abstracts



GOBIERNO DE CHILE

Microsoft Research



PROGRAMA CTE PFB16



XI IFAB



Center for Bioinformatics  
and Genome Biology



FUNDACIÓN  
CIENCIA PARA LA VIDA

**Pontificia Universidad Católica de Chile, Santiago, Chile  
Oct. 15-17, 2008.**

### Conference Chairs

David Holmes (Chair, Chile) and Tomás Pérez-Acle (Co-chair, Chile)

### Local Organizing Committee (Chile)

Marta Bunster  
Danilo Gonzalez  
David Holmes  
Francisco Melo  
Tomás Pérez-Acle  
Cristián Salgado  
Herman Silva

### International Steering Committee

Amos Bairoch (Swiss Bioinformatics Institute, Switzerland)  
Julio Collado-Vides (UNAM, Mexico)  
Oscar Grau ( UNLP, Argentina)  
Nebosja Jovic (Microsoft Research, USA)

### Scientific Session Chairs

Ana Teresa Vasconcelos (Brazil) (Genomics and Systems Biology)  
Goran Neshich (Brazil) (Structural Bioinformatics)  
José Valverde (Spain) (Databases, Networking, Escience)  
Tomás Peres-Acle (Chile) (High Performance Computing)  
Oswaldo Trellis (Spain) (Training, Education)  
Fernando González-Nilo (Chile) (Nanobioinformatics, Emerging Opportunities)

## PROGRAMME

### Wednesday morning, October 15

9:00 – 9:30		Welcome	David Holmes, (Conference chair), Tomás Pérez-Acle (Conference co-chair) and Rafael Vicuña (Dean, Faculty of Biological Sciences, Pontificia Universidad Católica, Santiago, Chile).
<b>Session 1</b> 9:30-13:00		Ana Tereza Vasconcelos (Chair), LNCC/ CNPq, Brazil.	<b>Genomics and Systems Biology</b>
9:30 – 10:15	Keynote	Julio Collado-Vides, Center for Genomic Sciences, UNAM. Mexico.	Computational Biology of Gene Regulation in Bacteria
10:15 – 10:45		Ziomara Gerdtzen, Millennium Institute for Cell Dynamics and Biotechnology. Chile.	Modeling heterocyst pattern formation in cyanobacteria

10:45 – 11:20	Coffee		
11:20 – 11:40		Alvaro Peña, Bioinformatics, Unit, Institut Pasteur, Uruguay.	Evaluation and possible improvements in algorithms for miRNA target prediction, particularly in chronic lymphocytic leukemia (CLL)
11:40 – 12:00		Javier Rivas, CIC, CSIC/USAL, Spain.	Finding the gene signature associated to different disease subtypes by multiclass predictors based on transcriptomic profiles
12:00 – 12:20		Victoria Martin, University of Malaga, Spain.	Integrated genomics' projects management. Handling the genomic projects complexity
12:20 – 12:40		Andrea Mahn, Universidad de Santiago de Chile. Chile.	Discovery of selenium status biomarkers through nutritional proteomics
12:40 – 13:00		David Holmes, (CBGB), Fundación Ciencia para la Vida, Chile.	Alternate Open Reading Frames: a Bioinformatics Nightmare but a Potential Goldmine for Gene Evolution
13:00– 14:15	Lunch		

### Wednesday afternoon, October 15

<b>Session 2</b> <b>14:15- 18:00</b>		Goran Neshich (Chair), Embrapa Informática Agropecuária. Brazil.	<b>Structural Bioinformatics</b>
14:15 – 14:45	Keynote	Nebojsa Jojic, Microsoft Research, USA.	HLA binding models and their role in explaining host-pathogen co-evolution.
14:45 – 15:05		Wendy González, CBSM, Universidad de Talca, Chile	Study of plant voltage gated potassium channels using structural bioinformatics tools.
15:05 – 15:25		Angel González, CBUC, Pontificia Universidad Católica, Chile.	Molecular Modeling and Docking Studies of the Cannabinoid Receptor Type 1 (CB1). Comparison of Rhodopsin and b2-Adrenergic-Based Comparative Models.
15:30 – 16:00	Coffee		
16:00 – 16:20		Carlos F. Lagos, CBUC, Pontificia Universidad Católica de Chile, Chile.	Integrated in silico tools for drug design
16:20 – 16:40		Raul Araya-Secchi, CBUC, Pontificia Universidad Católica, Chile.	Molecular dynamics study of the archaeal aquaporin AqpM
16:40 – 17:00		Bryan Reynaert, CBUC, Pontificia Universidad Católica, Chile.	The small world of thermophilic proteins
17:00 – 17:20		Jardine J.G., Embrapa Informática Agropecuária, Brazil	How did the structure function descriptors of proteins change with introduction of 'Remediated' PDB files?
17:20 – 17:45		Francisco Melo, Pontificia Universidad Católica, Chile.	Scoring Functions for Protein Structure Prediction.
17:45-18:00			Round table discussion
18:00			Welcome cocktail and poster session

## Thursday morning, October 16

<b>Session 3</b> <b>9:15 12:45</b>		José Valverde (Chair), EMBnet/CNB, Spain.	<b>Databases, Networking, E-Science Initiatives</b>
9:15 – 10:00		José Valverde, EMBnet/CNB, Spain.	Running large scale simulations on the Grid
10:00 – 10:30		Goran Neshich, Laboratorio de Biología Computacional Embrapa Informática Agropecuária, Brazil.	Structure Descriptors of Chameleon Sequences
10:30 – 11:00	Coffee		
11:00 – 11:30		Emiliano Barreto, EMBnet Colombian node, Colombia.	A BLAST client for the creation, visualization and analysis of alignments
11:30– 12:00	Guest Speaker	Oscar Grau, Argentina.	Towards a Bioinformatics meta-network
12:00-12:45			Round table discussion
12:45 -14:15	Lunch		

## Thursday afternoon, October 16

<b>Session 4</b> <b>14:15–15:30</b>		Tomás Perez-Acle (Chair), CBUC, Chile.	<b>High Performance Computing</b>
14:15 – 14:45		Ivan Sosa, Microsoft, Chile.	Microsoft HPC Solution Architecture
14:45 – 15:10		Hector Urbina, CBUC, Pontificia Universidad Católica, Chile.	'A web services-based workflow for drug discovery'
15:10 – 15:30		Ricardo Honorato-Zimmer, CBUC, Pontificia Universidad Católica, Chile.	'Conan-COMplex Network ANalysis'
15:30 – 16:00	Coffee		

<b>Session 5</b> <b>16:00–18:00</b>		Oswaldo Trelles (Chair), Spain.	<b>Training and Education</b>
16:00 – 16:45	Keynote	Jaime Puente, Microsoft Research, USA.	Scholarly Communications, Scientific Research & Cyberinfrastructure
16:45-17:10		Paola Arellano, Reuna, Chile.	Academic Networks: e-Infrastructure for e-Science
17:10 – 17:35		Oswaldo Trelles, University of Malaga, Spain.	Extensible-distributed tool for synchronous e-learning environment
17:35-18:00			Round table discussion
21:00			Conference Banquet

Friday morning, October 17

<b>Session 6</b> <b>9:15- 10:30</b>		Fernando González-Nilo, CBMS, Universidad de Talca, Chile.	<b>Nanobioinformatics and Other Emerging Opportunities</b>
9:15 – 9:45		Daniel Luna, Hospital Italiano de Buenos Aires, Argentina.	Medical Informatics Overview
9:45 – 10:00		Stefano Ciesa, Universidad Politécnica de Madrid, Spain.	The ACTION-GRID Project: Nanobiomedical Informatics Support action
10:00-10:15		Fernando González-Nilo, CBMS, Universidad de Talca, Chile.	Nanobioinformatics vs Bioinformatics
10:15 – 10:30		Fabian Avila, CBMS, Universidad de Talca, Chile.	Nanobioinformatics: Dendrimer-Drug Interaction Energy.
10:30-11:00	Coffee		
<b>Session 7</b> <b>11:00-13:00</b>			<b>RIB Organizational meeting and closing of congress</b>
<b>13:00</b>	Lunch		

## ORAL PRESENTATIONS

### WEDNESDAY, OCT. 15, 2008

#### Computational Biology and Modeling of Gene Regulation in Bacteria

Julio Collado-Vides.

Centro de Ciencias Genómicas, UNAM. Av. Universidad s/n. Cuernavaca, Col. Chamilpa, Morelos 62210; Mexico. collado@ccg.unam.mx

For many years, we have gathered from original literature knowledge on transcriptional regulation in *E. coli* K-12. This knowledge is electronically encoded in two databases, RegulonDB and EcoCyc. This talk is based on a global perspective of the computational biology and modeling in gene regulation starting from building a database to the diverse analyses of the regulatory network including static and topological statistical properties and finally to the quest for a comprehensive and dynamical description of the behavior of the complete cell. The talk is organized giving some hints in each of these steps either from recent work in our lab, or relevant issues or ideas.

The regulatory network of *E. coli* is currently the most comprehensive and best studied network of any living organism. The maintained effort of curation in our laboratory feeds both RegulonDB and EcoCyc. We estimate to have around 20 to 25% of the full network of regulatory interactions of transcription initiation, given different estimates based on the number of transcriptional factors (TFs) and their interactions. We

have performed several analyses, including a natural language approach that recovers around 40% of the known interactions (3). The modeling of regulation in the database is based on a trilogy composed of tables for the TF, the target DNA binding site, and the promoter. These are linked via a link-table that contains information of their interaction, such as the positive or negative effect on transcription. We believe this to be a core model that might well be applied to the future modeling of regulatory interactions beyond transcription initiation. We have a comprehensive updated collection of models for the TF-target recognition based on position weight matrices for each TF, as well as for the different types of promoters in *E. coli*. We have also built a database of microarray data – obtained from the major public sources of microbial microarrays - that will allow us to interrogate the dynamics of cell behavior in different conditions.

Regulation as a network has been described as a power-free, modular and hierarchical network. We have decomposed the network into physiologically congruent modules based on a mathematical derived threshold to distinguish global regulators from the rest of regulators. Modules are basically generated eliminating all global interactions. The challenge ahead is to integrate the regulatory interactions with physiological and metabolic effects, together with signal transduction pathways. Finally, I present a recent major effort to experimentally map transcription initiation sites in the *E. coli* genome, a research project performed by the laboratory of Enrique Morett at the Institute of Biotechnology, UNAM, in Cuernavaca.

Acknowledgments: This work was funded by NIH, grants number R01 GM071962-05 and GM077678, and by UNAM, PAPIIT grant number IN214905.

### **Modeling heterocyst pattern formation in cyanobacteria.**

Ziomara P. Gerdtzen<sup>1</sup>, J. Cristian Salgado<sup>1</sup>, Axel Osses<sup>2</sup>, Juan A. Asenjo<sup>1</sup>, Ivan Rapaport<sup>2</sup> and Barbara A. Andrews<sup>1</sup>.

Millennium Institute for Cell Dynamics and Biotechnology. <sup>1</sup>Department of Chemical Engineering and Biotechnology, University of Chile, Av. Beaucheff 861, Santiago 837-0456, Chile. <sup>2</sup>Department of Mathematical Engineering, Center for Mathematical Modeling, (UMI 2807-CNRS), University of Chile, Casilla 170/3 Correo 3, Santiago, Chile.

In this paper, we study the process by which cyanobacteria vegetative cells differentiate into heterocysts in the absence of nitrogen. We propose a simple network which captures the complexity of the differentiation process and the role of all variables involved in this cellular process. Specific characteristics and details of the system's behavior such as transcript profiles for *ntcA*, *hetR* and *patS* between consecutive heterocysts are studied. The proposed model is able to capture one of the most distinctive features of this system: a characteristic distance between two heterocysts, with a small standard deviation according to experimental variability. The system's response to knock out and over expression of *patS* and *hetR* was simulated in order to validate the proposed model against experimental observations. In all cases, our simulations show good agreement with reported experimental results. The model also shows that refractability of heterocysts to the action of *PatS* is not required in order to achieve the characteristic differentiation pattern observed in cyanobacteria. Acknowledgments: Millennium Scientific Initiative ICM P05-001F.

### **Evaluation and possible improvements in algorithms for miRNA target prediction, particularly in chronic lymphocytic leukemia (CLL).**

Álvaro Pena, Martín Graña, Hugo Naya.

Bioinformatics Unit, Institut Pasteur, Montevideo, Uruguay.

MicroRNAs (miRNAs) are a class of small noncoding RNA molecules (20-23 nucleotides) that regulate gene expression by inducing degradation or translational inhibition of target mRNAs. More than 500 miRNAs have been reported in the human genome (roughly 1-2% of the coding genes), constituting one of the largest classes of regulatory genes. Aberrant expression of miRNAs has been reported in diverse cancer subtypes. Moreover, increasing experimental evidence implicates miRNAs either as oncogenes or tumor suppressors.

For this reason, different algorithms have been developed for *in silico* miRNAs target prediction. MiRanda and TargetScan are two of the most commonly used algorithms, but in our case only few of the miRNAs targets are predicted by both methods. Nevertheless, when evolutionary constraints are removed, both algorithms predicted several common targets. Recently, a method based on SVMs reported that nearly half of the predicted miRNA target sites in human are not conserved in other organisms. We attempted to detect those targets that were relevant for the study of CLL without the conservation analysis. Under these conditions MiRanda score distribution is reported and analyzed.

### **Finding the gene signature associated with different disease subtypes by multiclass predictors based on transcriptomic profiles.**

Fontanillo C., Risueño A., Prieto C. and De Las Rivas J.

Bioinformatics and Functional Genomics Research Group, Cancer Research Center (CIC, CSIC/USAL), Salamanca, Spain.

Genome-wide expression profiles derived from microarray data are widely used in functional genomics and have become very useful to build "class predictors" using different machine learning (ML) methods. One of the main and most significant applications of such technology is probably disease subtypes classification and disease outcome prediction. However, there is a fundamental problem in these studies, since in most of the ML methods applied to built predictors it is

not clear how to find the best ones, that provide the minimal errors and that reflect a meaningful biological signature of the classes. In fact, many methods (NN, kNN, discriminant analysis) are like “black-boxes” with respect to the biological entities (i.e. genes) that are behind the predictor figures and, therefore, are not useful to allow a proper discovery of the real biology behind the classification.

To address these problems, we have built a multiclass “transparent” disease predictor (applied to leukemia subtypes) which reflects the biological entities behind the classes. We also present a method to assess the biological relevance of the predictors by comparing the error of the optimum classifier with the distribution of errors obtained with series of random predictors. The feature selection (i.e. gene selection) applied to build the classifier is critical, because it provides the way to order the genes by significance. Therefore, we also optimized the feature selection method.

#### **Integrated genomics’ projects management. Handling the genomic projects complexity.**

Victoria Martín<sup>1,2</sup>, Antonio Muñoz<sup>1,2</sup>, Fernando Barranco<sup>1</sup> and Oswaldo Trelles<sup>1</sup>

<sup>1</sup>University of Malaga, Computer Architecture Department, <sup>2</sup>Integrated Bioinformatics at the National Institute for Bioinformatics, Spain.

Technological breakthroughs have driven diverse genomics initiatives producing huge quantities of diverse but complementary, heterogeneous and geographical disperse collections of data from genomic, genetics and transcriptomics experiments. Efficient handling and secure sharing of this information poses additional challenges in project management tasks. In this context, we have developed a genomic projects management system named pUMA (projects at the University of Malaga) to facilitate conduction of genomics projects through a customizable web-portal and personalized underlying database system.

pUMA is a generic, platform-independent, customizable and scalable projects management system specialized in genomics projects, able to quickly deploy a corporative portal to: (a) support project dissemination and coordination activities; (b) delivering data through user friendly interfaces; (c) provide secure

storage, control information sharing and management of different data sources; (d) facilitate the integrated analysis of data; and (e) provide universal access by using standard web-browsers. pUMA software is being successfully used as an essential tool for the progress of several national projects: Tomato, Strawberries, Olive, Asparagus, Porcine, Allergies, etc. and it is available at [www.bitlab-es.com](http://www.bitlab-es.com)

#### **Discovery of Selenium Status Biomarkes Through Nutritional Proteomics.**

Andrea Mahn\*, Héctor Toledo, Manuel Ruz, and Ricardo Vega.

Departamento de Ingeniería Química, Universidad de Santiago de Chile.

Some chemical forms of selenium are considered as chemoprotective. The traditional selenium indexes do no account for the metabolic status of chemoprotective forms of selenium. It was investigated if the dietary supplementation of rats with different chemical forms of selenium was reflected as differences in the abundance of some proteins in blood plasma. Also, the effect of the dose of selenium and the length of the dietary supplementation period was investigated using a proteomic approach to quantify protein expression differences in blood plasma.

Some proteins were significantly affected by the selenium dose; other proteins were significantly affected by the length of the supplementation period. No protein was significantly affected by both factors in the same direction, and no protein reduced significantly its abundance due to supra-nutritional selenium supplementation. The dose of selenium would be the main factor that affects the response of rats to selenium supplementation.

The proteomic response to dietary supplementation with selenium was characterized, resulting in some proteins whose abundance was increased selectively depending on the chemical form of selenium that supplemented the diet. These protein patterns could be proposed as new selenium indexes to assess the metabolic status of this element, and would constitute a new biotechnological tool which could be used in the context of cancer prevention.

Acknowledgements. FONDECYT Project N°1061154



### **Alternate Open Reading Frames: a Bioinformatics Nightmare but a Potential Goldmine for Gene Evolution.**

David S. Holmes, Gustavo Rivera and Francisco J. Ossandón.

Center for Bioinformatics and Genome Biology (CBGB), Fundación Ciencia para la Vida and Depto. de Ciencias Biológicas, Facultad de Ciencias de la Salud, Universidad de Andrés Bello, Santiago, Chile.

Alternate open reading frames (ORFs) that generate overlapping genes have been well documented in viral genomes. However, few examples have been described in other organisms, prompting us to carry out a large scale, multi-genome survey of Bacteria and Archaea in a search for alternate ORFs that might be functional genes. Analysis of over 700 genomes (> 2 million genes) reveals that substantial alternate ORFs (>300 nucleotides without a stop codon) are surprisingly common, especially in G+C rich genomes. This talk will outline the discovery and interpretation of alternate ORFs and will describe a publicly available, searchable database (AlterORF, [www.AlterORF.cl](http://www.AlterORF.cl)) of alternate ORFs.

Many alternate ORFs do not encode proteins and can be misannotated as hypothetical genes exacerbating the so-called "orphan gene" problem and confounding the interpretation of genomes. AlterORF is helping to depurate these incorrect gene identifications. However, alternate ORFs also represent a rich source of potential coding information that can give rise to novel proteins when coupled to appropriate regulatory sequences and, as such, may represent a major contributor to gene evolution. Examples will be described of new genes that are hypothesized to have arisen from alternate ORFs.

Acknowledgements: Conicyt Basal CTE PFB16, Fondecyt 1050063, DI-UNAB 3406-R, DI-02-08/I and a Microsoft Sponsored Research Award.

## **THURSDAY, OCT. 16, 2008**

### **HLA binding models and their role in explaining host-pathogen co-evolution.**

Nebojsa Jojic, Ph.D.

Microsoft Research, Redmond, WA, USA.

With dramatic increases in the amount of data characterizing antigens associated with various diseases, it is now possible to study patterns of pathogen-host co-evolution on a molecular level. While phylogenetic analysis of sequence data can be used to unravel the history of mutations both in viruses and the primate immune system, they typically fail to provide a mechanistic picture of evolutionary forces in action. More recent datasets about immune system targets (epitopes) sampled from various viral proteins has led to substantial improvement in the predictive power of models for immune system targeting. These models relate human allelic variation with the variation in targeted epitopes. In this talk, I will describe how these models can be used to model the evolutionary forces shaping pathogen polymorphisms and the allelic diversity in the HLA region of the human genome.

### **Study of plant voltage gated potassium channels using structural bioinformatics tools.**

Wendy González<sup>1</sup>, Samuel Morales<sup>1</sup>, Jans Alzate<sup>1</sup>, Danilo González<sup>1</sup>, Ingo Dreyer<sup>2</sup>.

<sup>1</sup>Centro de Bioinformática y Simulación Molecular, Universidad de Talca, Chile, <sup>2</sup>Universität Potsdam, Institut für Biochemie und Biologie, Heisenberg-Gruppe Biophysik und Molekulare Pflanzenbiologie, Germany.

Voltage gated potassium (Kv) channels are membrane proteins that allow voltage-driven potassium (K<sup>+</sup>) flux across cellular membranes. Kv channels in plants participate in multiple events, which include stomatal movements and ion uptake from the soil. Plant Kv channels show a similar topology as their counterparts in animals, namely: four subunits each with six transmembrane segments (S1-S6), where S4 is the voltage sensor and S5-S6 constitute the pore.

Electrophysiology and structural bioinformatics tools have partially revealed the molecular mechanisms that regulate Kv channels in animals. To date, Kv channels in plants have been studied experimentally, but structural information about them does not exist. We used Molecular Simulation and Quantum-Mechanics/Molecular-Mechanics (QM/MM) methods to answer unsolved questions in plant Kv channels. Computational approaches helped to clarify the differences between channels that conduct K<sup>+</sup> only into plant cells and other

channels that conduct K<sup>+</sup> only in the reverse direction. We studied also the opening of Kv channels of stomatal guard cells mediated by acidification.

### **Molecular Modeling and Docking Studies of the Cannabinoid Receptor Type 1 (CB1). Comparison of Rhodopsin and $\beta$ 2-Adrenergic-Based Comparative Models.**

Angel González-Wong, Carlos F. Lagos and Tomás Pérez-Acle.

Centre for Bioinformatics (CBUC), Faculty of Biological Sciences, Pontificia Universidad Católica de Chile.

The seven transmembrane helices (TMH) G-protein-coupled receptors (GPCRs) constitute one of the largest superfamilies of signaling proteins found in mammals. GPCRs compose a versatile group of protein sensors that are involved in many important physiological processes, and represent primary targets for drug discovery. In this study, the crystal structure of the  $\beta$ 2-adrenergic receptor was used as template for the development of a molecular model of cannabinoid receptor type 1 (CB1), both members of the class A GPCR family. The resulting CB1 structure was employed in docking experiments with well known cannabinergic compounds. Finally, the ligand binding predictions were compared with previous results on a CB1 model derived from X-ray structure of bovine rhodopsin. Our results suggest that this new derived comparative model can be more appropriate for further virtual screening of chemical databases and rational drug design methodologies involved CB1 receptors. Acknowledgements: Fundación Ciencia para la Vida.

### **Integrated *in silico* tools for drug design**

Carlos F. Lagos<sup>1,2</sup>, C. David Pessoa-Mahana<sup>2</sup>, Patricio Huenchunir<sup>2</sup>, and Tomás Pérez-Acle<sup>1</sup>.

<sup>1</sup>Centre for Bioinformatics CBUC - Faculty of Biological Sciences, and <sup>2</sup>Medicinal Chemistry Laboratory MCL - Faculty of Chemistry, P. Universidad Católica de Chile.

It is generally recognized that drug discovery and development are time and resources consuming processes. Current *in silico* strategies for drug design are playing increasingly larger and more important

roles in drug discovery and development and are believed to offer means of improved efficiency for both the academic and industrial arenas.

We have incorporated several structure and ligand-based techniques into our academic drug discovery programs with promising results. These strategies have been combined to expedite and facilitate hit identification, hit-to-lead selection, optimization of the absorption, distribution, metabolism, excretion and toxicity profile and the avoidance of safety issues of a series of ligands designed towards various pharmaceutical targets.

Here, we describe the use of computational approaches that include ligand-based drug design (pharmacophore), structure-based drug design (drug-target docking), quantitative structure-activity and quantitative structure-property relationships to discover new drug candidates from different chemical scaffolds.

Acknowledgements: FONDECYT, United States Army Medical Research and Materiel Command (USARMRC), Fundación Chilena para Biología Celular & Fundación Ciencia para la Vida.

### **Molecular dynamics study of the archaeal aquaporin AqpM.**

Raul Araya-Secchi<sup>1</sup>, Jose Antonio Garate<sup>1,3</sup>, Ricardo Honorato-Zimmer<sup>1</sup>, Hector Urbina Saavedra<sup>1</sup>, David S. Holmes<sup>2,4</sup> and Tomas Perez-Acle<sup>1,2</sup>.

<sup>1</sup>Centre for Bioinformatics (CBUC), Faculty of Biological Sciences, P. Universidad Católica de Chile; <sup>2</sup>Fundacion Ciencia para la Vida, Avda Zañartu 1482, Ñuñoa, Santiago; <sup>3</sup>The SEC Strategic Research Cluster and the Centre for Synthesis and Chemical Biology, University College Dublin, Belfield, Dublin 4, Ireland; <sup>4</sup>Center for Bioinformatics and Genome Biology (CBGB), Fundacion Ciencia para la Vida, Santiago, Chile.

Aquaporins are a large family of transmembrane channel proteins that allow the passive but selective movement of water and small neutral alditols (i.e. glycerol) or CO<sub>2</sub> across cell membranes. Although aquaporins have been identified in all the domains of life, only bacterial and eukaryotic ones have been described and studied through atomic resolution structures and MD simulation techniques. In 2003 AqpM, the aquaporin from the archaeon *Methanothermobacter marburgensis* was the first archaeal aquaporin functionally characterized,

becoming in 2005 the first archaeal aquaporin to be crystallized and structurally characterized. Here we report the results of the first molecular dynamics (MD) simulation study performed on a fully hydrated model of the AqpM tetramer embedded in a lipid bilayer designed to further characterize, reproduce and understand at a molecular-atomic level, the water selectivity and permeation mechanism of AqpM. From our 20ns MD simulation of AqpM, we obtained time-resolved, atomic-resolution models of the water permeation mechanism across AqpM, which allowed us to calculate key biophysical features and permeation parameters. These features have been compared with other more widely studied aquaporins and aquaglyceroporins (i.e. Aqp0, GlpF, AqpZ) coming from the other domains of life.

Acknowledgments: Fundacion Ciencia para la Vida, Proyecto de Financiamiento Basal PFB-16, Fondecyt 1050063 and a Microsoft Sponsored Research Award.

#### **The small world of thermophilic proteins.**

Bryan Reynaert, Ignacio Vergara, Ricardo Honorato-Zimmer and Tomas Perez-Acle

Centre for Bioinformatics (CBUC), Faculty of Biological Sciences, Pontificia Universidad Catolica de Chile, Chile.

Protein thermostability has long been an active research topic in biotechnology. The main conclusion drawn from these studies is that protein thermal stability is closely related to the number and strength of intramolecular interactions. Within this context, it has been consistently shown that a higher number of salt-bridges in thermophilic proteins distinguish them from their mesophilic orthologs, suggesting that salt-bridges are an important factor for protein thermal stability. However, de novo inclusion of salt bridges and charged residues in mesophilic proteins does not necessarily imply enhanced thermostability. A cutoff of 4 to 6 Å is often used to consider ionic interactions neglecting long-range interactions. In this study, we have characterized ionic networks at different interaction cutoff distances in a 3D database of mesophilic and thermophilic proteins. Our results indicate that networks of ionic interactions occurring around a 10 Å cutoff share topological properties that classify

them as scale-free, small-world networks. In spite of the fact that the basic network structure is common for both types of proteins, differences between the two sets were found. These similarities and differences in the ionic network's structure could be determinant for protein thermostability and have important evolutionary implications.

Acknowledgments: Fundación Ciencia para la Vida and Fundación Chilena para Biología Celular

#### **How did The Structure Function Descriptors of Proteins Change With Introduction of 'Remediated' PDB Files.**

Jardine J.G., Mazoni I., Mancini A., Borro L.C., Alvarenga D., Cecilio P.L., Pelligrinelli T.V. and Neshich G.

Embrapa Informática Agropecuária, Biologia Computacional, Campinas, SP, Brazil.

Due to reasons related mainly to modernization of structure annotation data, the RCSB/PDB implemented in August of 2007, a "remediated" PDB files. The major difference introduced by this format is related to the number of chains present in the PDB file. Namely, files such as the 1cho.pdb, suffered a change in number of chains present in the enzyme structure: from previous 1 to current 3. The reason for such change is based on the fact that the enzyme structure contains two gaps, which are treated with certain degree of difficulty by the software producers. We have invested a significant amount of time and CPU resources in order to curate and analyze what are the consequences for in-silico analysis of the active site 3D environment after implementing remediated PDB files to STING\_DB. For this purpose, we analyzed a number of proteins, chosen to represent both the files which suffered a change in total number of chains as well as those which did not. Results show that the ensemble of protein structure descriptors did not change with respect to a range of values used for example for selecting amino acids into the restricted group of active site residues.

### Scoring Functions for Protein Structure Prediction.

Francisco Melo.

Pontificia Universidad Catolica de Chile, Chile.

In this talk, a general description of scoring functions widely used in protein structure prediction will be provided. This description includes the typical components and structure of scoring functions, and also how these scoring functions are generally derived and used.

Then, some new developments of scoring functions recently carried out in our laboratory will be described in more detail. Three different topics will be covered, which attempt to provide a simple and practical vision about the balance on information quality and quantity of scoring functions for protein structure assessment and prediction. The specific topics that will be presented are:

- 1) The use of scoring functions with a different set of parameters than those adopted to derive them.
- 2) The calculation of effective atom-atom interactions when deriving and using the scoring functions.
- 3) The incorporation of evolutionary information in the derivation of scoring functions.

Finally, a summarized 'future outlook' of the upcoming challenges for the development of improved scoring functions will be given.

### Running large scale simulations on the Grid.

José R. Valverde.

CNB/CSIC, Spain.

In this paper, we present a methodology to use the EGEE Grid to run large scale simulation models in population dynamics and show that this problem is representative of a larger class of massively parallel problems. We have been iteratively building increasingly more complex and hopefully accurate models to study population evolution using large scale Montecarlo simulations. This requires analysis of progressively refined one-time experiments, a process which demands simplified automation methods to run the simulations and deal with technical problems. We have taken measures at various steps in the process to study the efficiency gains obtained. While our simple

approach may arguably be far from achieving optimum efficiency, we were able to reduce running times from years to days allowing us to run progressively refined simulations satisfactorily with minimum effort. We conclude analyzing Grid efficiency and discussing which benefits can be realistically expected with the current technology and provide useful advice for future Grid developers.

### Structure Descriptors of Chameleon Sequences.

Mancini A., Jardine J.G., Mazoni I., Borro L.C., Alvarenga D., Cecilio P.L, Pelligrinelli T.V. and Neshich G.

Embrapa Informatica Agropecuaria, Brazil.

It has been known since the early work by Sander and Kabsch in 1984 that some identical sequence motifs make completely different local (but extended) folds. In that work, Sander and Kabsch showed the existence of a number of pentapeptides, capable of nucleating both alpha helix and beta strand in different sequence and structure constellations. However, the authors had at that time a very limited universe on which to base their conclusions; namely, the size of the PDB at that time (1984) was only 154 protein structures available. In 2008 the number of available structures is 51,079 (May 27th), 47137 containing proteins only. Consequently, the database is 306 times larger and we took full advantage of this fact. In order to analyze chameleon sequences, we established very strict rules and considered only those ones that passed the test of consensus among HSSP and STRIDE definitions for the Secondary Structure Elements. This is a major difference in our approach versus that of Sander and Kabsch.

We then created a data mart of sequences and respective structures with variable size and subjected them to analysis of structure descriptors stored in STING\_RDB. Results are discussed in terms of the influence that a local 3D environment has on SSE nucleation.

### A BLAST client for the creation, visualization and analysis of alignments.

Andrés Pinzón, Emiliano Barreto.

EMBnet Colombian node, Colombia.

NCBI-BLAST is one of the most common bioinformatics tools for the alignment of biological sequences, and is available on several public servers as well as a stand-alone application. Although widely diffused, public BLAST servers present several restrictions for end-users, mainly in the number of sequences that can be analyzed, the availability of user defined databases and the analysis of results. Some of these restrictions can be solved by using the NCBI-BLAST stand-alone version, which is not user friendly.

In order to overcome those restrictions and improve the analysis of NCBI-BLAST results, we have developed BLAME, a BLAST client that allow users to:

- Create personal BLAST databases and perform typical tasks.
- Organize data by projects.
- Analyze data by means of a web or a command line interface (CLI) making it possible to:
  - Generate histograms and export results in several formats (PDF, HTML etc.)
  - Retrieve data by user defined filters or by BLAME pre-defined searches.
  - Visualize and to analyze alignments.
  - Use BLAME as part of workflows (CLI).

BLAME is open software, developed on PHP5, using MySQL 5.0.51a as backend. The current version of BLAME is available for UNIX platforms at: <http://bioinf.ibun.unal.edu.co/software/blame>

#### Microsoft HPC Solution Architecture.

Ivan N. Sosa, Regional Sales Solution Professional for HPC, Microsoft Chile.

Windows HPC Server 2008 (HPCS) combines the power of the Windows Server platform with rich, out-of-the-box functionality to help improve the productivity and reduce the complexity of your HPC environment. Windows HPC Server 2008 can efficiently scale to thousands of processing cores and provides a comprehensive set of deployment, administration, and monitoring tools that are easy to deploy, manage, and integrate with your existing infrastructure. Windows HPC Server 2008, the successor to Windows Compute Cluster Server 2003, is based on Windows Server 2008 and is designed to do the following:

- **Improve productivity of systems administration and cluster interoperability** by dramatically simplifying the overall deployment, administration and management over the entire system lifetime while ensuring interoperability with existing systems infrastructure.
- **Rapid HPC application development through integration with Visual Studio 2008**, which provides a comprehensive parallel programming environment. In addition to supporting standard interfaces such as OpenMP, multiprocessor interconnect (MPI) and Web services, Windows HPC Server 2008 also supports third-party numerical library providers, performance optimizers, compilers and debugging toolkits.
- **Seamlessly scale from workstation to cluster** by allowing end users to harness the power of distributed computing through a familiar Windows-based desktop environment without requiring specialized skills or training.

#### Towards a bioinformatics meta-network.

Oscar Grau, UNLP, Argentina. Abstract not available at time of going to press.

#### A web services-based workflow for drug discovery

Héctor Urbina, Gonzalo Sánchez, Ricardo Honorato-Zimmer, Carlos F. Lagos and Tomás Pérez-Acle  
Centre for Bioinformatics (CBUC), Faculty of Biological Sciences, Pontificia Universidad Católica of Chile.

Web services, an information technology that allows the access to high performance computing environments over the Web, have become an integral part of workflow management in scientific applications. In the life sciences, the web service approach is seen as being a road to standardizing the multitude of tools available from different providers, enabling mass dissemination of resources and knowledge sharing. Within this context, our group is developing different web services suitable to be combined in order to produce novel tools for drug discovery. Here, we present a web service-based tool that offers multiprocessor-distributed docking, binding and quantum single-point

energy evaluation and multi-format conversion. Thus, our implementation offers to the scientist the ability to produce virtual screening workflows using standard SOAP-client graphical interfaces like Taverna. Our work is open source and cross-platform available.

Acknowledgments: Fundación Ciencia para la Vida and Fundación Chilena para Biología Celular.

### Conan - COmplex Network Analysis.

Ricardo Honorato-Zimmer<sup>1</sup>, Tomas Perez-Acle<sup>1,2</sup>.

<sup>1</sup> Centre for Bioinformatics (CBUC), Faculty of Biological Sciences, P. Universidad Católica de Chile, Avda. Portugal 49, Santiago, Chile, <sup>2</sup>Fundación Ciencia para la Vida, Avda. Zañartu 1482, Ñuñoa, Santiago, Chile.

Conan stands for COmplex Network ANalysis, a C++ library created at the Centre for Bioinformatics (CBUC) whose main goal is to produce rapid and accurate analyses of complex networks. Conan allows the scientific developer the inexpensive computation of global and local network properties, such as the average shortest path (ASP), clustering coefficient (C) and degree distribution. It also computes novel network properties like topological entropy and fractal dimension.

The development of Conan takes advantage of the newest features of C++ language's standard (C++0x), parallel programming, generic programming and is heavily based on the Boost Graph Library (BGL), as well as on other Boost libraries. Indeed, almost all its classes derive from one derived from the BGL, which allows the developer, already familiar with the BGL, simply to use the methods provided by this library on Conan's classes. In this sense, Conan can be viewed as an extension of BGL and provides a set of functions that work syntactically similar to those found in BGL.

Acknowledgements: Fundación Ciencia para la Vida and Fundación Chilena para Biología Celular.

### Scholarly Communications, Scientific Research & Cyberinfrastructure.

Jaime Puentes.

Microsoft Research, USA.

This talk will address key topics about

the trends and the potential impact of Scholarly Communications and Virtual Research Environments for scientific research and the academic world in general. Collecting and analyzing data, authoring, publishing, and preserving information are all essential components of the everyday work of researchers. Technology is playing an increasingly fundamental role in supporting this scholarly communication lifecycle. A virtual research environment will be also presented which supports bioscience researchers in managing the increasingly complex range of tasks involved in carrying out research.

### Academic Networks: e-Infrastructure for e-Science.

Paola Arellano.

REUNA, Santiago, Chile.

The National Research and Education Networks (NRENs) allow those who work in research and education, to collaborate and share information and resources through a series of interconnected networks. These NRENs are used to transfer data, support experiments and applications that are crucial for collaborative research and education. These networks are one of the fundamental pillars for the development of e-Infrastructures that support e-Science. In Latin America since 2004, thanks to the European support (through the @LIS Programme), there has been a major increase in the development of NRENs from only four formally constituted networks including the Chilean network REUNA which is one of the pioneers in the region ([www.reuna.cl](http://www.reuna.cl)) to the 12 countries that currently have NRENs, allowing more than 600 institutions to be connected in the region.

Currently, we are working on the consolidation of the Latin American network, RedCLARA ([www.redclara.net](http://www.redclara.net)), and thanks to the support that comes from the regional countries and the European Commission, we are initiating the ALICE2 project. ALICE2 will focus on long term infrastructure, sustainability, the Millennium Goals and inclusion, where we will endeavor to connect the 18 Latin American countries. In addition, new proposals in the field of e-Science are emerging whose objectives are:

- To build an interdisciplinary forum

- To define a strategy for the sustainable implementation of a collaborative infrastructure of National Grid
- To accelerate the adoption of the Grid technologies
- To find support and funds to establish the NGIs and regional grid initiative.

Along with the efforts of each country, there are regional projects in this scope, for example:

1. EELA, "E-Science grid facilities for Europe and Latin America"
2. FEMCID/OAS Project: "Fostering the Use of ICT in Science, Technology and Innovation (e- Science)"

Challenges that must be addressed in order to consolidate the e-Science initiatives in the region are:

- e-Infrastructure (networks, hardware & software)
- Qualified human resources
- Interest and commitment of the research communities
- Country Priorities

#### **Extensible-distributed tool for synchronous e-learning environment.**

Sergio Ramirez, Javier Rios and Oswaldo Trelles.

Department of Computer Architecture; University of Malaga, Spain

In this work, we report a software platform for on-line distance learning tasks that enables teachers to be present in the student space, deploying the same class, in the same schedule and with the same students through the Internet. Thus, the working scenery is a normal student room, in which some students are present and others are virtual students with an interactive presence that allows them to participate by being questioned by the professor and maintaining a fluid but controlled communication with their classmates.

Several important technical aspects have been solved using public domain components: web-browsers, blackboards, on-line messaging-chats and -forums, and real time sound and video delivery. All these components have been integrated by means of a plug-in based library, producing an efficient architecture endowed with a user friendly interface that provides a flexible and customized interface for different age-levels.

**FRIDAY, OCT. 17, 2008**

#### **Medical Informatics Overview**

Daniel Luna.

Hospital Italiano de Buenos Aires, Argentina.

Abstract not available at time of going to press.

#### **The ACTION-GRID Project: Nanobiomedical Informatics Support Action.**

Stefano Ciesa.

Universidad Politécnica de Madrid, Spain.

The ACTION-Grid project is an International Cooperation on healthcare information systems based on Grid capabilities, Medical and Biomedical Informatics and nanoinformatics between Latin America, the Western Balkans, North Africa and the European Union (EU). The main conceptual objective of ACTION-Grid will be to collect the relevant achievements in the fields listed above that can be reused and transferred to Latin America, the Western Balkans and North Africa. By extending the methods proposed in ACTION-Grid such scope could be easily extended to other regions and countries. Such knowledge reuse will be based on previous achievements of the consortium and, of course, from other people and groups. Acknowledgments: Action-GRID FP7-ICT-224176

#### **Nanoinformatics versus Bioinformatics.**

Gonzalez-Nilo Fernando<sup>1</sup> and Cachau Raul<sup>2</sup>.

<sup>1</sup>Universidad de Talca, Center for Bioinformatics and Molecular Simulation, Chile and <sup>2</sup>National Cancer Institute, SAIC, USA.

Everyday, biologists, chemists, physicists, mathematicians and engineers are working together to create new nanoparticles that require exhaustive structural characterization. These analyses demand intensive use of computational chemistry, which can be very time consuming. Such revolutionary milestones in Pharm, Medicine, and other fields are solely possible with collaborative grid computing and the use of new emerging fields like nanoinformatics. Nanoinformatics can be defined as the application of information

technology to the field of Nanotechnology or more specifically Nanobiology. Nanoinformatics entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of nanotechnology data.

Numerous tasks in nanoinformatics are very similar to bioinformatics, and an analysis and the comparison of these two areas could help accelerate the development of nanoinformatics. For example, whereas in bioinformatics mapping and analysis of DNA and protein sequences is an important task, in nanoinformatics it is the use of SMILES (Simplified Molecular Input Line Entry Specification) code to compare the sequence of different nanoparticles using matching algorithms. Similarly, in bioinformatics 3-D model visualization is a critical step to analyze structural properties of proteins, and in nanoinformatics such analysis is also very critical and many of the tools and methods used for proteins could also be used for the analysis of nanoparticles in order to rationalize their physical-chemistry properties. In this work a brief comparison between bioinformatics and nanoinformatics will be discussed.

Acknowledgments: Anillo Científico PBCT ACT/24 and Action-GRID FP7-ICT-224176

### **Nanoinformatics: Dendrimer-Drug Interaction Energy.**

Avila Fabián, Gonzalez-Nilo Fernando, Caballero Julio, Aguayo Daniel.

Universidad de Talca, Center for Bioinformatics and Molecular Simulation, Chile.

Nanoinformatics is an emerging discipline where some tools related to bioinformatics converge with nanotechnology. Several applications of nanotechnology in medicine require a fast analysis of thousands of data inputs to understand the nanoparticle properties required to some specific application.

Drug delivery is probably one of the most popular applications of nanotechnology to medicine. The impact of this technology could be very important in human health. By using this technology, it should be possible to reduce drug doses or increase the bioavailability of these drugs. Different systems have

been study in this field, such as entrapment of drugs inside dendrimers, involving electrostatic, hydrophobic and hydrogen bond interactions. The energy analysis of these interactions using experimental techniques is costly and difficult to interpret; so, it is necessary to use or develop efficient computational methods to obtain structural information from these interactions. Our group has implemented an efficient methodology to compute interaction energies of pairs of molecules (dendrimer-molecule) using a sampling algorithm based on Monte Carlo and semi-empirical quantum chemical calculations. To compute the interaction energy, we implemented a pipe-line, where the MOPAC package is used for accurate energy calculations. Our method allows us to correlate these energies with experimental drug-dendrimer affinities.

Acknowledgment: Anillo Científico PBCT ACT/24 and Action-GRID FP7-ICT-224176.

## **ABSTRACTS OF POSTERS**

### **A Mathematical model for metal stress response in *Halobacterium* NRC-1.**

Guillermo Espinoza, Alejandro Maass and Elisabeth Pécou.

University of Chile and University of Nice Sophia Antipolis, France.

Our work is based on the biological results of professor Baliga and co-workers. In their article they have reconstructed the physiological behaviors of *Halobacterium* NRC-1 and they have proposed a biological model using a system level study, including the principal genes and strategies to withstand stress from transition metals, identifying four mechanisms that play a central role in conferring resistance to excess.

In this poster, we propose a mathematical model for the uptake, efflux, storage and trafficking of transition heavy metals in *Halobacterium* NRC-1 based on a framework of differential equations and the power law formalism. Since there are many variables, we have sub-divided the whole system into two parts, one solving the trafficking of Cu(II) and Zn(II), and the other dealing with the uptake of Mn(II) and Fe(II).

We have proved, in a formal way, that the system presents stationary states, homeostatic behavior depending on the different constants (degradation,



synthesis and affinity) and monotonicity conditions that imply global steady states responses independent of either the dynamics or the choice of parameters. Finally, we have made several simulations to obtain numerical solutions for the systems of equations and graphical representations for the trajectory of each one of the elements in the model. This proves that it is possible to measure qualitatively the adaptability of the system to its environment.

### **A maximum entropy reconstruction of gene interaction networks during skeletal muscular differentiation in *C. elegans*.**

José Antonio Barros and Tomás Pérez-Acle.

Centre for Bioinformatics (CBUC), Faculty of Biological Sciences, Pontificia Universidad Católica de Chile, Chile.

Several mathematical approaches have been used to reconstruct the network of gene interactions taking as a source microarray data. From these methods, the Maximum Entropy approach offers a unique opportunity to obtain the interaction network without any *a priori* knowledge. Using previously reported data regarding skeletal muscle differentiation in *C. elegans*, we have reconstructed the network of gene interactions. Our results identify important genes that control cell differentiation, tissue development and skeletal muscle production.

### **C32-alkane films adsorbed onto amorphous and hydrophilic SiO<sub>2</sub> surfaces. Towards organic-inorganic hybrid bionanodevices.**

Ignacio Vergara<sup>1,2</sup>, Raul Araya-Secchi<sup>1</sup>, Hector Urbina<sup>1</sup>, Valeria del Campo<sup>2</sup>, Edgardo Cisternas<sup>2</sup>, Ulrich Volkmann<sup>2</sup>, Haskel Taub<sup>3</sup> and Tomás Pérez-Acle<sup>1</sup>.

<sup>1</sup>Centre for Bioinformatics (CBUC), Faculty of Biological Sciences, P. Universidad Católica de Chile, <sup>2</sup>Laboratorio de Superficies, Facultad de Física, P. Universidad Católica de Chile, <sup>3</sup>Department of Physics and Astronomy, University of Missouri Columbia, USA.

Alkane molecules play an important technological role as the principal constituents of

commercial lubricants and as fundamental models for more complex organic molecules. Similarly, SiO<sub>2</sub> is a key interfacial material found in many semiconductor devices. Thus, knowledge of the structure and dynamics of alkanes adsorbed on SiO<sub>2</sub> surfaces could have a great impact on the semiconductor industry and biological physics. In particular, the study of alkane films on SiO<sub>2</sub> surfaces could lead to improvements in the lubrication and protection of microelectromechanical (MEMS) devices and also to support advances in electronic and organic/inorganic hybrid bionanodevices.

In this study we present molecular dynamics simulations of C32 films (n-C32H66), supported on an amorphous and hydrophilic SiO<sub>2</sub> surface. Our study reveals the biophysical fundamentals related with the structure formation, phase transitions, and wetting/dewetting phenomena of these films. Simulation results are compared to those experimentally obtained on C32 films by ellipsometry, Atomic Force Microscopy, and x-ray reflectivity measurements.

Acknowledgments: Fundación Chilena para Biología Celular and Fundación Ciencia para la Vida.

### **Determination of pK values on PAMAM dendrimers using quantum mechanical methods.**

Alzate-Morales Jans, Caballero Julio and González-Niilo Danilo.

Centro de Bioinformática y Simulación Molecular, Universidad de Talca, 2 Norte 685, Casilla 721, Talca, Chile.

Since their discovery in the late 1970s, dendrimers have provided considerable impetus for research because of their unique properties as well as their potential applications as metal complexing agents, nanoreactors for particle synthesis, light harvesting devices, or as gene vectors. PAMAM (poly(amidoamine)) dendrimers represent an exciting new class of macromolecular architecture called "dense star" polymers. Unlike classical polymers, dendrimers have a high degree of molecular uniformity, narrow molecular weight distribution, specific size and shape characteristics, and a highly-functionalized terminal surface. The manufacturing process is a series of repetitive steps starting with a central initiator core. On the other hand, dendrimers can accumulate positive

charge by protonation of the primary amines at the rim and the tertiary amines in the interior. Their charge is thus pH dependent, whereby they are positively charged at low pH and neutral at high pH. The microscopic charging mechanism of these dendrimers is not immediately obvious and has prompted discussion in the literature.

In this work, we deal with the determination of the pK values on PAMAM G0 and G1 dendrimers by means of some quantum mechanical methods. The results are compared with the experimental data available in the literature and are used to understand the mechanism of protonation.

Acknowledgments: Grants PBCT PSD-86 and ACT-24 from CONICYT and Government of Chile.

### eBiopipeline on line resource for EST data analysis.

Fernandez Paula, Principi Dario, Delfino Santiago, Heinz Ruth Amelia, Farber Marisa, Lew Sergio and Paniego Norma.

Instituto de Biotecnología, CICVyA INTA Castelar Pcia, Buenos Aires and Instituto de Ingeniería Biomédica, Facultad de Ingeniería, UBA, Argentina.

The characterization of expressed sequence tag collections represents an affordable approach to investigate organisms, especially for orphan genome species. We present eBiopipeline, a web-based tool for processing and annotating sequence data. The pipeline accepts raw chromatogram trace files or fasta-formatted files. The system offers a user-customize analysis tool including the following processes: base calling and cleaning, clustering and assembling, and annotation. Each component comprises highly reliable open-source tools (Phred-Phrap; Cap3; Blast algorithm, HMMER, etc) and non-redundant public and custom-made databases (SwissProt, TrEMBL, NR, InterPro, GO, etc). All the process components can be run consecutively. Default or specific parameters are allowed for each program in the pipeline. Components can be run independently according to user-specified options. The results of all steps are available to users for downloading as FASTA and/or text files. Application using an EST collection of wild *Helianthus* will be presented.

### Bioinformatics analysis provides insight into life in extremely acidic environments (pH1).

Valdés J., Quatrini R. & D.S.

Holmes. Center for Bioinformatics and Genome Biology, Fundación Ciencia para la Vida and Universidad Andrés Bello, Santiago, Chile.

*Acidithiobacillus ferrooxidans*, *A. thiooxidans* and *A. caldus* are chemolithoautotrophic  $\gamma$ -proteobacteria that thrive at pH 1-2 and derive energy from the oxidation of iron and/or sulfur. They fix CO<sub>2</sub> and nitrogen. They are important in industrial bioleaching operations for the recovery of copper and gold and play important roles in the biogeochemical cycling of metals and nutrients in naturally acidic environments. We have annotated their genomes and have carried out comparative genomics in order to shed light on their metabolism and genetic regulatory circuits. Bioinformatics and genomic approaches to unraveling their biology are particularly valuable given the low growth rates and yields of these microorganisms and difficulties encountered in their genetic manipulation.

Draft genome sequences were generated from *A. thiooxidans* and *A. caldus* and were assembled, annotated and compared to the publicly available genome sequence of *A. ferrooxidans*. Eleven metabolic pathways and phenotypic characteristics of the three acidithiobacilli have been reconstructed: CO<sub>2</sub> fixation, TCA cycle, sulfur oxidation, sulfur reduction, iron oxidation, iron assimilation, quorum sensing, hydrogen oxidation, flagella formation, chemotaxis and nitrogen fixation. Predicted transcriptional and metabolic interplay between pathways pinpoints potential coordinated responses to environmental signals such as energy source and nutrient limitations.

**Conclusion:** This study provides preliminary metabolic and regulatory models for each species and predicts important interactions that may occur between them in the environment (ecophysiology). Several responses appear to be especially characteristic of autotrophic microorganisms and may have direct implications for metabolic processes of critical relevance to understanding how these microorganisms survive and proliferate in extreme environments.

Acknowledgements: Conicyt Basal CTE PFB16, Fondecyt 1050063, DI-UNAB 34-06, DI-UNAB 15-06/I, BHP Billiton Initiation Award and a Microsoft Sponsored Research Award.

**From an EST Database to functional molecular markers in *Prunus* species.**

Diez-de-Medina, S., Latorre, M. and Silva H.

Millennium Nucleus in Plant Cell Biotechnology, Plant Functional Genomics and Bioinformatics Lab, Andrés Bello University, Republica 217, 837-0146 Santiago, Chile.

One of the Chile's major problems for exporting fruits is the distance to the foreign markets and reaching those markets with good quality fruits. The prolonged storage of fruits such as peaches and nectarines can have negative effects on their quality. Prolonged storage triggers a physiologically disorder know as chilling injury. This phenomenon affects especially the varieties that Chile grows for export. One of the most important problems associated with chilling injury is woolliness. Woolly fruits lack juice, a condition that it is unacceptable for consumers, leading to an economic problem for exporters and the exporter country. In order to address this problem, a functional genomics peach (*Prunus persica*) project was developed. ESTs sequencing and gene expression studies under different postharvest conditions were performed resulting in a significant universe of unigenes that were shown to be regulated under different condition of post harvest treatment. After analysis of ESTs and contigs obtained from four different DNAs libraries, a set of 32 genes where selected due to their association with metabolic pathways related to cell wall metabolism and stress response. Using this selected group of unigenes, new intron flanking PCR molecular markers where developed with the goal to locate them on the *Prunus* reference map and to perform evaluations of peach populations that segregate for the woolliness phenotype. Acknowledgments: ICM P06-065-F and Consorcio Biofrutales SA.

**Functional association networks derived from the structural information embedded in the genome of *Escherichia coli* K-12.**

Felipe I. Lazo, Cristopher A. Oporto, Gustavo A. Rivera, David S. Holmes, José L. Jara and Raquel Quatrini.

Center for Bioinformatics and Genome Biology, FCV, Andrés Bello University and Universidad de Santiago

de Chile, Chile.

As has been observed for many complex networks, protein-protein interaction (PPI) networks are modular, reflecting the relative independence and coherence of different functional units in a cell. Recently, it has been claimed that modularity also occurs within functional association networks. Functional association is an *in silico* measure of the likelihood that two given genes operate in the same metabolic pathways. Such information can be derived from structural information embedded in a genome, such as the physical distance between proteins or conservation of proteins pairs across genomes. In the light of the significant progress in sequencing projects and the need for annotation on a large scale, inference of modules in functional association networks have an important potential for predicting functions for hypothetical genes, improving functional annotations of known genes and extending metabolic interactions.

Using functional association data for *Escherichia coli* K12 derived from the Prolinks and String databases, we evaluated how well current metrics for defining functional association reflect curated functional modules stored in Ecocyc and KEGG databases. Next, we reconstructed *E. coli*'s functional association network, we applied three different clustering strategies and defined which performed best in identifying true functional modules, namely, operons, protein complexes and metabolic pathways. Based on this analysis, we redefined and augmented some of the published functional association metrics, resulting in improved predictability and we confirmed the modular nature of functional association networks.

Acknowledgements: Conicyt Basal CTE PFB16, Fondecyt 11060164 & 1050063 & Microsoft Sponsored Research Award.

**In Silico Metabolic Network Analysis of Implicated Factors in the Development of *Pseudomonas aeruginosa* (PAO1) Biofilms.**

Rehder, Jan Christian Otto. & Estévez-Bretón, Carlos Manuel.

Grupo de Investigación de Bioquímica Computacional y Estructural y Bioinformática, Pontificia Universidad

Javeriana, Cra. 7 No. 43 – 82, Edificio 52 Lab.108. Tel. (+571) 320 8320 Ext. 4136, Colombia.

The present work collected information inherent to gene regulatory pathways involved in morphogenesis of structurally complex microcolonies that constitute *Pseudomonas aeruginosa* PAO1 biofilms. A graphical model of the molecular pathways was constructed using CellDesigner, permitting the development of a chemo-kinetic functional model written in SBML language through Systems Biology Graphic Notation (SBGN).

Possible congruencies were traced between the kinetics of two *P. aeruginosa* PAO1 quorum sensing signals, the production of surfactants from non-motile cells in the microcolonies and the relevance of an endosymbiotic prophage in the generation and maintenance of this complex structure, simulating in continuous time the variation of the concentrations from the components. Through simulation, a clear dependency in the production of surfactants respect to the concentrations of quorum sensing signals was observed whereas, on the other hand, prophage behaviour could not be studied due to the limitations in the construction of the functional model. It was confirmed, that the response of surfactants to the quorum sensing signal concentrations was an authentic property of auto-organized systems and that systemic research presents a fundamental tool to understand morphogenesis and biodesign.

### **Prediction of protein behavior in hydrophobic interaction chromatography and aqueous two-phase system using only their amino acid composition.**

J. Cristian Salgado, Juan A. Asenjo and Barbara A. Andrews.

Centre for Biochemical Engineering and Biotechnology, Department of Chemical Engineering and Biotechnology, University of Chile, Beauchef 861, Santiago, Chile.

The prediction of the partition behavior of proteins in hydrophobic interaction chromatography (HIC) and aqueous two-phase systems (ATPS) using mathematical models based on amino acid composition was investigated. Predictive models were based on the average surface hydrophobicity (ASH), which is

estimated by means of models that require the 3D structure of proteins and by models that use only the amino acid composition of proteins. These models were evaluated in a set of 12 proteins with known experimental retention time in HIC and 11 with known partition coefficient in ATPS. Our results indicate that the prediction based on the amino acid composition is feasible for both separation systems, even though the quality of the prediction depends strongly on the operational conditions. In the case of ATPS the best results were obtained by the model which assumes that all of the amino acids are completely exposed. An increase in the predictive capacity of at least 54% with respect to the models which use the 3D structure of the protein was obtained in this case. However, best prediction in HIC was obtained by the model based on a linear estimation of the amino acidic surface composition. This model required additional tuning, but its performance was 5% better than the one obtained by the 3D structure model.

Acknowledgments: FONDECYT PostDoctoral Research Project 3070031.

### **Pre-processing Optimization of RNA Immunoprecipitation Microarrays Data**

Emiliano Barreto-Hernandez<sup>1,3</sup>, Margarida Gama-Carvalho<sup>2</sup>, and Lisete Sousa<sup>3</sup>

1 Instituto de Biotecnología, Universidad Nacional de Colombia Bogotá, Colombia. 2 Inst. de Medicina Molecular, Faculdade de Medicina, Univers. de Lisboa Av. Professor Egas Moniz, 1649-028 Lisboa, Portugal. 3 DEIO, Faculdade de Ciências da Universidade de Lisboa and CEAUL Bloco C6 -Piso 4, Campo Grande, 1749-016 Lisboa, Portugal.

Pre-mRNA splicing is an essential step in the post-transcriptional gene expression control involving protein splicing factors like U2AF, among others, which are known to be exported to the cytoplasm and may be implicated in additional cellular functions. Identification of U2AF-associated mRNAs under native conditions was performed by immunoprecipitation and hybridization to Affymetrix GeneChip. Normalization and gene selection methods were performed, but the results were not reliable as they were quite different for different procedures, mainly because more than

20% of the mRNAs detected are differentially enriched between both samples and the common normalization methods are based on small differences between them. We implemented a background correction method inspired by a non-specific hybridization method used for pre-processing data from ChIP-Chip technology. In this work, linear regression models are used to model the non-specific hybridization in each array, accounting for interactions between each three consecutive nucleotides in the probe sequence. Each probe intensity on the array was standardized using its predicted intensity and the variance of the probe for similar predicted intensities. The standardized probe intensity values showed no need for further normalization and could be directly compared. Following Johnson et al. (2006), we propose a probe set score, and a probe set enrichment value (ENRval) and its respective p-value for gene enrichment selection.

#### **Studies of oligomerization of Human Liver Arginase I, using kinetics and bioinformatics approach.**

\*Jaña-Pérez, Natalia.; +Uribe, Elena; +Carvajal, Nelson.; \*Martínez-Oyanedel, José.

\*Laboratorio de Biofísica Molecular, +Laboratorio de Enzimología. Departamento de Bioquímica y Biología Molecular. Facultad de Ciencias Biológicas. Universidad de Concepción, Chile.

Arginase is manganese-dependent enzyme, consisting of a trimer of identical subunits, that catalyzes the hydrolysis of L-arginine to L-ornithine and urea. Arginase has a wider role controlling arginine in metabolic process, including production of creatine, polyamines, proline and nitric oxide. An S-shape structural motif, the last 19 C-terminal residues, plays a role in oligomerization, contributing more than 54% of the interaction area of the subunit in Human Liver Arginase I (HLAI). The *B. caldovelox* hexameric arginase reveals a conformational change induced by occupancy of an external site by guanidine or arginine. This occupancy is equivalent to the arginine 308 in HLAI.

To characterize the oligomerization process, the mutation R308A and deletion 309Del have been constructed, expressed and partial purified using DEAE cellulose chromatography. Km and kcat, Mn dependence, pH profile have been analyzed.

In parallel, a bioinformatics approach using molecular simulations NPT, SDM and SA was used to reproduce the conformational changes induced by the mutation and deletion and the oligomerization of the monomers.

The kinetics data shown that both variants present hyperbolic kinetics in contrast to cooperative effects presented by the wild-type enzyme. The bioinformatics models reveal a trimeric state for 309Del and monomeric states in R308A. Our results agree, and support the importance of the residue Arginine 308 in the process of Arginase oligomerization.

#### **Study of Energetic Features in Protein-protein Interaction Interfaces for Identifying Transient and Obligat Complexes.**

Tatiana Gutiérrez-Bunster<sup>{3}</sup>, Marta Bunster<sup>{2}</sup>, JoséMartínez-Oyanedel<sup>{2}</sup> y Luis Rueda<sup>{1}</sup>

<sup>{1}</sup>Department of Computer Science. <sup>{2}</sup>Department of Biochemistry and Molecular Biology. Universidad de Concepción. <sup>{3}</sup>Universidad del Bío-Bío. Concepción. Chile

One of the crucial issues for understanding and classifying protein-protein interactions (PPI) is to characterize their interfaces in order to discriminate between transient and obligate complexes. Energy features for PPI interfaces were extracted from databases of protein complexes in terms of their known three-dimensional structure, and which are already classified as transient or obligate.

The FastContact package was used to obtain the energy contribution for each of the interaction surfaces. 642 feature were extracted for each complex, and a feature selection algorithm was applied to a dataset of 296 complexes, which was performed by following the forward selection strategy and the Chernoff distance to measure the class separation. To study the accuracy of the classification, different linear methods were used to reduce dimensions, including the heterocedastic, the homocedastic and Chernoff discriminant analysis, combined with Bayesian quadratic and linear classifiers. Also, support vector machine with polynomial kernels was used. The information on discrimination led to a ranking of the most influential features in the interface that discriminate between obligate and transient complexes. The best results obtained in this study

showed 81% classification accuracy in a 10-fold cross validation setup. The analysis of the most discriminating features shows that desolvation energies make a more important contribution to the separation of the classes than the electrostatic energies.

### **The Dependence of SVM-RFE Gene Selection on Microarray Data Noise.**

Elizabeth Tapia, Pilar Bulacio and Laura Angelone.  
CIFASIS-Conicet Institute, Bv. 27 de Febrero 210 Bis, Facultad de Cs. Exactas e Ingeniería, Riobamba 245 Bis, Rosario, Argentina.

A simulation approach to study the dependence of SVM-RFE gene selection with respect to the signal-to-noise ratio of microarray datasets is presented. It is revealed that SVM-RFE gene selection depends on both the signal-to-noise ratio and the policy of gene elimination. Specifically, for SVM-RFE implementations removing a constant fraction of genes per step, smaller signal-to-noise ratios lead to the selection of smaller sets of genes with lower rates of false discoveries. Conversely, for the native SVM-RFE implementation removing one gene per step, larger sets of genes with higher rates of false discoveries, are selected. The results of these studies correlate well with those on real data. We conclude that one should be very careful with SVM-RFE gene selection conclusions on microarray data.

### **The use of AlterORF to deplete mis-annotated genes. The case study of 14 *Staphylococcus aureus* genomes.**

F. J. Ossandon<sup>1</sup>, F. Lazo<sup>1</sup>, G. Rivera<sup>1</sup>, A. Auchincloss<sup>2</sup>, A. Bairoch<sup>2</sup>, D. S. Holmes<sup>1</sup>

<sup>1</sup>Center for Bioinformatics and Genome Biology (CBGB), Fundación Ciencia para la Vida ([www.cienciavida.cl](http://www.cienciavida.cl)) and Depto. de Ciencias Biológicas, Facultad de Ciencias de la Salud, Universidad Andrés Bello; <sup>2</sup>UniProtKB/Swiss-Prot, Swiss Institute of Bioinformatics, Switzerland.

Alternate open reading frames (ORFs) have been well documented in viral genomes. However, few examples have been described in other organisms, prompting us to carry out a large scale survey of Bacteria and Archaea in a search for alternate ORFs

that might be functional genes. Analysis of over 2 million genes reveals that substantial alternate ORFs are surprisingly common, especially in G+C rich genomes. Many of them do not encode proteins but are often misannotated as genes, exacerbating the so-called "orphan gene" problem. AlterORF is helping to deplete these incorrect genes.

We analyzed 14 *Staphylococcus aureus* strains available from the NCBI using the web based tool AlterORF ([www.AlterORF.cl](http://www.AlterORF.cl)) in which every gene and alternate ORF was compared with 10 public domain databases. *S. aureus* was chosen because of its medical significance and low G+C genome (32% G+C), which presents fewer alternate ORFs and gives the opportunity for detailed examination. Despite its low G+C content, examples were found where the wrong ORF appears to be misannotated as a gene. In addition, although many genes were annotated as "hypothetical, no known function", AlterORF was able to predict significant hits with known domains.

It is proposed that AlterORF should be included as part of the "gold standard" computational package that is being developed to provide high quality genome annotation.

Acknowledgements: Conicyt Basal CTE PFB16, Fondecyt 1050063, UNAB DI-34-06/R, UNAB DI-02-08/I and a Microsoft Sponsored Research Award.

### **Use of clustering studies to improve the prediction of protein behavior in hydrophobic interaction chromatography and aqueous two-phase system.**

Jorge E. Ugarte, Barbara A. Andrews and J. Cristian Salgado.

Centre for Biochemical Engineering and Biotechnology, Department of Chemical Engineering and Biotechnology, University of Chile, Beauchef 861, Santiago, Chile

The aim of this study is the improvement of mathematical models used to predict the behavior of proteins in hydrophobic interaction chromatography (HIC) and aqueous two-phase systems (ATPS) based on the amino acid composition of proteins. This problem was tackled by carrying out clustering analysis over a large database of amino acid properties (APV): Self Organizing Maps, k-means, Simulated Annealing, Growing Neuronal Gas, Growing Grid, and Hierarchical

Clustering were used. This analysis allows us to generate new APVs from those found in the literature, which were used to improve prediction models. Three of these models require only the amino acid composition of proteins and different assumptions regarding the amino acids tendency to be exposed to the solvent; the other requires the proteins' three dimensional structure. These models were adjusted using the new APVs and were evaluated on a set of 12 proteins with known experimental retention time in HIC and 11 with known partition coefficient in ATPS. We found that the best APVs were generated by the Growing Neuronal Gas algorithm. In fact, two vectors that significantly improve the performance of the prediction models were found. By using these vectors the prediction performance of the model based on the 3D structure and the best model based on amino acid composition were improved in 38% and 31%, respectively.

Acknowledgments: FONDECYT PostDoctoral Research Project 3070031 and Millennium Scientific Initiative ICM P05-001F.

### **Opportunities for Collaborative Research in Bioinformatics at the Fundación Ciencia para la Vida, Santiago, Chile.**

David S. Holmes.

Center for Bioinformatics and Genome Biology, Fundación Ciencia para la Vida, Santiago, Chile.

The Fundación Ciencia para la Vida (FCV) hosts a number of computer intensive bioinformatics and genomic projects that provide excellent opportunities for e-Science collaborations. In addition, it is involved in outreach programs to the Latin American scientific community and general public that could be significantly advanced by cooperation with other Ibero-Latin American Universities and Institutes. Therefore, we are seeking partners in these endeavors and welcome suggestions for additional collaborations. This poster indicates the projects available for e-Science collaborations, identifies existing partners and provides additional information for others who may wish to join us.

The development of e-Science projects has allowed us to: accelerate bioinformatics research, build human resource capacity, leverage additional funding,

advance HPC, transcend the critical mass problem, overcome geographic limitations, harness the power of large scale, collaborative e-science, improve the visibility of Latin American bioinformatics and become involved in international outreach.

Acknowledgments: Conicyt Basal CTE PFB16, Fondecyt 1050063 and a Microsoft Sponsored Research Award.

### **The LEGO World of Alternative Splicing.**

Felipe A. Veloso, Felipe Lazo and David S. Holmes.

Center for Bioinformatics and Genome Biology (CBGB), Fundación Ciencia para la Vida and Andrés Bello University, Chile.

This work presents an analogy between the phenomenon of Alternative Splicing and the LEGO game. This game consists in combining interlocking blocks of different colors to build different structures. Fundamentally, alternative splicing follows a similar logic, only adding some constraints to the combinations. Both systems rely upon the abundance and diversity of the blocks or pieces in order to generate distinct functional combinations. From this perspective, the eukaryotic gene structure is concordant with the proposed analogy. In fact, the more exons a gene has, the more diverse its exons are, promoting biological complexity. This correlation is also observed in orthologous gene families with variable exon number during evolution, and in experimentally validated alternative splicing events. As a negative control, prokaryotic gene/protein sequences were analyzed. If these sequences are divided *in silico* into 'pseudoexons', the diversity of the resulting subsequences is significantly lower. These results highlight the role of alternative splicing as an evolutionary pathway and selective pressure on compositional profiles of eukaryotic gene products.

Acknowledgments: Conicyt Basal CTE PFB16, Fondecyt 1050063, UNAB DI-16-06/I and a Microsoft Sponsored Research Award.

## National Nodes

### Argentina

IBBM, Facultad de Cs. Exactas, Universidad Nacional de La Plata

### Australia

RMC Gunn Building B19, University of Sydney, Sydney

### Belgium

BEN ULB Campus Plaine CP 257, Brussels

### Brazil

Lab. Nacional de Computação Científica, Lab. de Bioinformática, Petrópolis, Rio de Janeiro

### Chile

Centre for Biochemical Engineering and Biotechnology (CIByB), University of Chile, Santiago

### China

Centre of Bioinformatics, Peking University, Beijing

### Colombia

Instituto de Biotecnología, Universidad Nacional de Colombia, Edificio Manuel Ancizar, Bogota

### Costa Rica

University of Costa Rica (UCR), School of Medicine, Department of Pharmacology and ClinicToxicology, San Jose

### Cuba

Centro de Ingeniería Genética y Biotecnología, La Habana

### Finland

CSC, Espoo

### France

ReNaBi, French bioinformatics platforms network

### Greece

Biomedical Research Foundation of the Academy of Athens, Athens

### Hungary

Agricultural Biotechnology Center, Godollo

### India

Centre for DNA Fingerprinting and Diagnostics (CDFD), Hyderabad

### Italy

CNR - Institute for Biomedical Technologies, Bioinformatics and Genomic Group, Bari

### Mexico

Nodo Nacional EMBnet, Centro de Investigación sobre Fijación de Nitrógeno, Cuernavaca, Morelos

### The Netherlands

Dept. of Genome Informatics, Wageningen UR

### Norway

The Norwegian EMBnet Node, The Biotechnology Centre of Oslo

### Pakistan

COMSATS Institute of Information Technology, Chak Shahzaad, Islamabad

### Poland

Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warszawa

### Portugal

Instituto Gulbenkian de Ciencia, Centro Portugues de Bioinformatica, Oeiras

### Russia

Biocomputing Group, Belozersky Institute, Moscow

### Slovakia

Institute of Molecular Biology, Slovak Academy of Science, Bratislava

### South Africa

SANBI, University of the Western Cape, Bellville

### Spain

EMBnet/CNB, Centro Nacional de Biotecnología, Madrid

### Sri Lanka

Institute of Biochemistry, Molecular Biology and Biotechnology, University of Colombo, Colombo

### Sweden

Uppsala Biomedical Centre, Computing Department, Uppsala

### Switzerland

Swiss Institute of Bioinformatics, Lausanne

## Specialist Nodes

### CASPUR

Rome, Italy

### EBI

EBI Embl Outstation, Hinxton, Cambridge, UK

### ETI

Amsterdam, The Netherlands

### ICGEB

International Centre for Genetic Engineering and Biotechnology, Trieste, Italy

### IHCP

Institute of Health and Consumer Protection, Ispra, Italy

### ILRI/BECA

International Livestock Research Institute, Nairobi, Kenya

### MIPS

Muenchen, Germany

### UMBER

School of Biological Sciences, The University of Manchester,, UK

for more information visit our Web site  
[www.embnet.org](http://www.embnet.org)



**EMBnet.news**  
**ISSN 1023-4144**

Dear reader,

If you have any comments or suggestions regarding this newsletter we would be very glad to hear from you. If you have a tip you feel we can print then please let us know. Please send your contributions to one of the editors. You may also submit material by e-mail.

Past issues of EMBnet.news are available as PostScript or PDF files. You can get them from the EMBnet organization Web site:

<http://www.embnet.org/download/embnetnews>

### Publisher:

EMBnet Executive Board  
c/o Erik Bongcam-Rudloff  
Uppsala Biomedical Centre  
The Linnaeus Centre for Bioinformatics, SLU/UU  
Box 570 S-751 23 Uppsala, Sweden  
Email: [erik.bongcam@bmc.uu.se](mailto:erik.bongcam@bmc.uu.se)  
Tel: +46-18-4716696

Submission deadline for the next issue:  
February 20, 2009