# EMBnet.news

Volume 12 Nr. 2
December 2006

- The 2006 EMBnet AGM
- WebLab: a bioinformatics platform
- YeastBASE @ CSC
- MRS - Maarten's retrieval system

and more...

# Editorial

The present issue summarizes the issues arising at the Assembly General Meeting and satellite events held in Sweden and Finland. As you can see the participation was high and EMBnet continues to expand and adapt itself to the evolving needs of its members. Other sections develop in what concern tools and a special mention should be given to MRS, as our community will be seeing more of its functionality and ease of use in the years to come.

The editorial board: Erik Bongcam-Rudloff, Domenica D'Elia, Pedro Fernandes, Kimmo Mattila, Lubos Klucar, and Gonçalo Guimaraes Pereira.

proteinspotlight

> ONE MONTH, ONE PROTEIN <

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 74. Please let us know if you like this inclusion.

Cover picture: Lesser Flamingo (Phoenicopterus minor). taken at Lake Nakuru National Park, Kenya, 2006. [© Erik Bongcam-Rudolf]

# Contents

Editorial ..... 2
The 2006 EMBnet AGM ..... 3
Symposium "Grids and Web Services in bioinformatics" ..... 12
WebLab: a bioinformatics platform ..... 13
YeastBASE @ CSC ..... 14
MRS and the management of biomolecular databanks ..... 17
Using MRS as sequence retrieval mechanism for EMBOSS ..... 21
Comparative genomics approach in promoter analysis ..... 23
Protein spotlight 74 ..... 26
Node information ..... 28

## Editorial Board:

# The 2006 EMBnet Annual General Meeting

## a joint Sweden-Finland initiative

The EMBnet at the CSC, Helsinki

The Annual General Meeting of the EMBnet Stichting of year 2006 was organised jointly by our Swedish and Finnish nodes. As announced, the format of the meeting was likely to attract attention as it was planned in not less than three locations, namely Uppsala, Helsinki and a cruise ship connecting Stockholm and Helsinki on a regular basis.

The trip was going to be an attraction by itself, while the planned collaborative workshop was going to take place in Uppsala, most of the official EMBnet business meeting on the boat and an attractive bioinformatics symposium in Helsinki. This symposium gathered the partners of the EMBRACE (EU funded) scientific project to which the EMBnet members were invited to contribute.

The BMC campus in Uppsala

## Collaborative workshop

On June 15th, all EMBnet members attended an interesting collaborative workshop, chaired by Jose-Ramon Valverde (Spain). Several subjects of interest to the community were highlighted. Short articles from some of the contributors can be found elsewhere in this issue while others will be published in the next issue.

Presentations were as follows:

- Batch queues in wEMBOSS by Martín Sarachu, AR.EMBnet. Reported by Oscar Grau
- Development of DNA microarray resources at CSC, Finland. Reported by Eija Korpelainen
- CSC's Yeast expression database, YeastBASE. Reported by Kimmo Mattila
- Analysis Environment for DNA Microarray Data. Reported by Aleksi Kallio
- MitoRes database: a searchable repository of nuclear-encoded mitochondria sequences. D. D'Elia, F. Licciulli, A. Turi, G. Grillo, D. Catalano, C. Saccone, Italy. Reported by Domenica D'Elia
- p53FamTaG: a database of p53 family members direct target genes integrating in silico prediction and microarray data. E. Sbisà, C. Catalano, G.Grillo, F. Licciulli, A. Turi, D. D'Elia,S. Liuni, G. Pesole, A. De Grassi1, M.F. Caratozzolo, A.M., D'Erchia, R. B. Navarro, A. Tullo, C. Saccone, A. Gisel, Italy. Reported by Domenica D'Elia
- Works on tools, databases and genomics, China. Reported by Jingchu Luo
- Database federation, current status and future prospects, Belgium. Reported by Robert Herzog
- Comparative genomics approaches in promoter analysis, Hungary. Reported by Endre Barta

After the presentations, a brainstorm session tended to collect all good ideas to further promote the EMBnet concept. Discussion items touched, among others, the relations with other regional or worldwide life science or bioinformat-

ics organizations like the Iberoamerican Network, the ONU based networks, HealtGrid, etc.... What can EMBnet do for them? The status of various ongoing projects was briefly discussed, namely EGEE, EMBRACE, SIMDAT, etc. A round table questioning of local status, personal interest, opinion on present and future of EMBnet concluded the workshop.

## Ethnic party

The evening of June 15th was devoted to the traditional "ethnic party", a traditional informal come-together of all delegates each bringing his national delicacies for the other members to sample… Thankfully, there was no Swedish traditional canned fish to indulge into this time, and it was probably better so, as the happening took place in the bioinformatics laboratory, and not outdoors like it was the case in Slovakia last year, and where the effluvium dissipated thanks to the mild central Europe winds… The Belgian delegation had brought its combination of "fromage de Herve" (a cheese known for its characteristic fragrance) and pear syrup which was definitely of a milder nature.

David and Valérie wrapped up in preparing appetizer with "fromage de Herve"

This party was warm and friendly as usual and was the best opportunity to socialise with our new candidate members, Allan from Costa Rica, Etienne from Kenya and Shahid and Raheel from Pakistan.

## Annual General Meeting of the EMBnet Stichting

On June 16th, Erik Bongcam-Rudloff, our acting chairman of the EMBnet Stichting, welcomed all participants to the formal AGM, in the Boström meeting room of the "Biomedicinska Centrum" (or BMC) on the campus of the Uppsala University, the Swedish University of Agricultural Sciences, the Ludwig Institute for Cancer Research and Regionala Etikprövningsnämnden.

### Attendance

Present were the delegates from Argentina, Belgium, China, Colombia, Finland, Hungary, Italy, Mexico, Norway, Poland, Portugal, Slovakia, South Africa, Spain, Sweden, Switzerland while Australia, Austria, Brazil, Chile, India, Israel, Russia and the EBI were represented by appointed proxies. Absences with apologies were noted for Brazil, Canada, Cuba, France, Israel, The Netherlands, EBI, ECTI, and Lion Bioscience while Chile, Germany, IHCP/BGMO and MIPS/GSF were noted as absent. Several observers from Belgium, Sweden, Finland, Switzerland attend the meeting while the applicant new members from Costa Rica (Allan Orozco), Pakistan (Shahid Chohan and Raheel Qamar) and Kenya (Etienne de Villiers) are present to introduce their applications.

Minutes of the 2005 AGM held at the Smolenice castle in Slovakia last year (September 16th and 17th) were approved with an amendment to paragraph 6.7.e from Italy.

### Financial report

Our treasurer Oscar Grau reports about the current status of the EMBnet Stichting funds. Overall good health of our accounting could be reported, even if several fees for the preceding year are still pending.

### Re-election of nodes

A call for re-elections was the next action, as the EMBnet Stichting statutes impose a renewal of membership every three years. This time, it was the turn for Argentina, Brazil, Chile, Colombia, Cuba, Hungary, India, Mexico, The Netherlands, Poland, South Africa, Spain, EBI, MIPS/GSF and UMBER. All these members, except Brazil and

Chile provided reports of their current status and activities over the last year. It was noted that Brazil had expressed his intention to leave EMBnet while the responsibility for a national Chilean node was to be transferred to a new institute whose responsible person is Mr. Juan Asenjo. Our delegate from Cuba had to report temporary financial difficulties impairing his ability to honour his yearly fee. The Dutch bioinformatics node is now the representative of an organised national network grouping five distinct biocomputing institutes, an example structure that our Dutch member suggest to be a better mapping to today's requirements of national interests than the traditional single national node.

## Candidate new nodes

The candidate new nodes were given the opportunity to present their activity and their institute.

**Costa Rica**, presented by the Faculty of Medicine of the University of Costa Rica, is a candidate *national node*. The delegate from Costa Rica, Allan Orozco was invited to present his project. He mentions the strong involvement of his country in the GBIF project and their plans to link biodiversity and biomolecular data. Their group benefits from a 20.000 $ grant and counts 10 persons. They are strong advocates of Open Source.

Allan Orozco, the delegate from Costa Rica

**Pakistan** entered an application as *national node*. The Pakistan's delegates present their institute, sharing several settlements around the country. They are member of several international organisations (COMSATS, COMSTECH, etc.). They are looking into preparing a good quality education in bioinformatics for their institute and the country. They are very eager to obtain support from the EMBnet to help them setting up all that is needed towards this goal. Till now, they were not active in bioinformatics research but are going to concentrate on various tropical and regional diseases. Jingchu Luo, delegate from China, warns them about the risk for scientists not to produce publications, if they are too deeply involved in bioinformatics services.

Raheel Quamar and Shahid Chohan, the delegates from Pakistan

**The BECA institute from Kenya** entered an application as *specialist node*. The presentation was given by Etienne de Villiers. His institute is strongly involved in the molecular biology of several diseases associated with livestock and crop production in Central Africa. BECA benefits from many contacts at the international level, notably with the Canadian International Development Agency (C$ 30 million funding). The biocomputing group has a very powerful computer infrastructure counting a 20 cpu BLAST machine, a GeneMatcher and a 66 cpu HCP cluster. Its Internet connections are limited to satellite transfers. BECA is looking very much in the possibilities

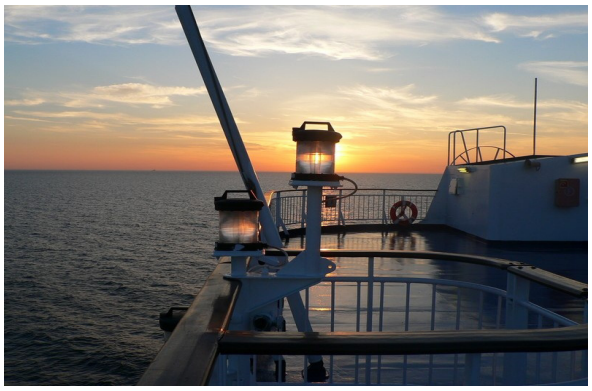
Etienne de Villiers, the delegate from ILRI/Becam

for EMBnet to assist in training. The EMBnet chairman will help them in person already this summer.

Marks of interest to join EMBnet were received from Sri Lanka and the Czech Republic. They are expected to produce the required documentation for application by next year's AGM. *The chairman also reported signs of interest from Trinitad and Tobago and from Paraguay, but these revealed to be fake proposals introduced as forged mail by a facetious member of the meeting, in the wake of the fever produced by the ongoing World Cup Football that was taking place during the same period. The Trinitad and Tobago and Paraguay football team were competing against Mexico and Sweden on the night before the AGM…*

The AGM had to finalise the discharging of the German node, after several years of absence of reporting. Problems of discontinued activity had been received from France, Austria, Canada, ICGEB and the United Kingdom. Formal action about these members will have to be undertaken in due time, according to the dispositions of the Statutes.

Elections for renewal of members produced favourable results for all candidates except Brazil that received 7 votes in favour, 9 against and 6 abstentions. As a consequence, Brazil will be presented for formal discharging at the next AGM. For the three candidate new members, the ballot produced unanimous votes in favour of their designation. From that moment on their delegates from Costa Rica, Pakistan and Kenya joined the meeting as full members with all voting rights.



The midnight sun shining on the Gabriella

## Boarding on Gabriella

Time had come to leave the Uppsala premises for the cruise ship where the AGM was to be continued. A one hour bus trip brought all participants at the harbour of Stockholm where everybody could board of the Gabriella ship from the Viking Line Company. We took the sea at 17:00 with the time of arrival at Helsinki was planned for next morning. The AGM could continue, but this time in the meeting room of the Gabriella, where the meeting was reconvened at 17:10. All meeting facilities, like microphone, beamer and WiFi were available as if we were on land…! Good organisation!

## Reports from EMBnet committees

The *Education and Training Committee* (ET-PC) activities were reported by Valérie Ledent and Vassilios Ioannidis. They reported about building an infrastructure for a repository of training material (courses, video presentations, etc.).

The *Technical Management Committee* (TM-PC) activities were reported by Georgios Magklaras. He mentioned the various actions to transfer the control of the EMBnet mailing lists and domain name. They are looking into the installation of a new computer infrastructure meant to host the various EMBnet network-accessible facilities (list server, web site, ftp server, DNS server, Content Management System, courses repository, etc.), presently distributed on several computers around Europe.

The *Publicity and Public Relations Committee* (P&PR-PC) activities were reported by Pedro Fernandes. Some internal difficulties in the committee did not yet allow to take over the production of EMBnet.news, the quarterly e-publication for which Laurent Falquet was in charge during the preceding years. The chairman insists on the strategic importance of EMBnet.news, whose accesses on the network amount to several tens thousands per annum. The committee commits itself to produce two issues of the newsletter by the end of 2006 and to comply in the future with a tight schedule of publication dates. The EMBnet website is another essential showcase of our activities that deserves an extensive rehearsal. JR Valverde and Robert Herzog insist on

replacing the current structure with a contents management system where it would be significantly easier for all partners to contribute with the large diversity of their activities. The TM-PC is asked to look into the technical possibilities and implement a CMS in the near future.

## Past and future projects

An overview of the past and present projects where member EMBnet nodes are participants is undertaken.

• First on the list is the current status of the ***EMBCORE*** EU project (5FP, 2000-2003). Payments of the expenditures engaged by members during the course of this project are still pending. The financial officer will be contacted to obtain an update on these matters, and the members are asked to advise their colleagues whenever clearance of their refunds will happen.

• The ***SRS Federation*** project that started in September 2003 with six partners of EMBnet. It came to an end when the Lion Bioscience company had to withdraw their commitment for continued collaboration, while they were taken over by BioWisdom.

• ***SIMDAT*** is a 6FP EU project where the Belgian node is a partner. The aim is to build a robust distributed network for product development, notably in the field of the Pharmaceutical companies. Many of the outcomes of SIMDAT are very likely to be of interest to EMBnet and progress of SIMDAT will be reported whenever adequate. A replacement of the Lion SRS technology had to be identified and MRS, a product from the Radbout University of Nijmegen was identified. Its implementation as a replacement of SRS proved very successful. A possible outcome could be the development of an ***MRS Federation***.

• Belgium also reported about possible progresses in a ***Federated Blast*** project, where the load for running large BLAST similarity searches could be spread over a distributed computer base.

• ***EMBRACE*** is a 6FP Network of Excellence project where several EMBnet nodes participate (Sweden, Finland, Switzerland, Italy, Spain, UMBER, EBI). Goals are somewhat parallel to SIMDAT, without stressing the industrial requirements.

• ***Vital-IT*** is a 6FP program to provide EU partners with a cost-free use of the *Integrated Computational Genomics Resource* at the Swiss Institute of Bioinformatics. Interested parties can benefit from a short or medium term stay at the institute to be trained in advanced genomics and pursue genomics research.

## Future directions

Following the Collaborative Workshop of the preceding day, Jose-Ramon Valverde proposed the creation of four task forces to promote EMBnet in several directions:

• ***EMBnet kit***: the mission is to prepare the kit of basic requirements for a new node to reach functional level, as well as to provide new users with the minimal setup to be able to work proficiently (Live DVD, Live CDROM, etc.) The volunteers to join this group are: Argentina, Norway, Mexico; project leader : Georgios Magklaras

• ***Federated Systems***: the mission is to manage the federated efforts (MRS Federation, Federated BLAST, etc.). The volunteers are: Norway, Slovakia, Argentina, Sweden, Mexico, Colombia, China and Portugal. Project leader is Robert Herzog

• ***Greedy team*** : the team is in charge of identifying new sources of finances for EMBnet as a whole, or in specialised areas. Volunteers are Sweden, Spain and Argentina. Project leader is Erik Bongcam-Rudloff

• ***Biomedical applications***: the goal is to join effort in medical- or health-oriented projects (genomics, proteomics, biochips, etc.). The precise goals of this group require further definition. As the time left is not sufficient, it is decided to take up this task during the first upcoming Virtual General Meeting (VGM). In the meantime, partners are invited to identify their areas of interest and potential contribution in this area.

## Elections

For the Executive Board (EB), our chairman Erik Bongcam-Rudloff is re-elected with 22 votes.

For year 2006, the EMBnet EB consists of : *Erik Bongcam-Rudloff, chairman, Oscar Grau, treasurer, Robert Herzog, secretary and Laurent Falquet, member*

For 2006, after some election rounds,

- the ET-PC consists of : *Vassilios Ioannidis - chaiman (Switzerland), Valérie Ledent - secretary (Belgium), Lisa Mullan (EBI) and Isabel Marquez (Portugal)*

- the P&PR PC consist of : *Kimmo Mattila (Finland), Lubos Klucar (Slovakia), Domenica D'Elia (Italy) and Pedro Fernandes (Portugal)*

- the TM PC consists of : *George Magklaras (Norway), Nils-Einar Eriksson (Sweden), Jose Valverde (Spain), Cesar Bonavides (Mexico) and David Coornaert (Belgium)*

## Any other business

The chairman invites all committees to designate immediately the positions of chairperson and secretary and publish the decisions as soon as possible on the EMBnet website. EMBnet benefits from a permanent license for the Marratech e-conferencing system. The clients for all platforms are available for free on the Marratech webserver (`www.marratech.com`). The EMBnet server is available at all times (embnet1.bmc.uu.se:8000) for EMBnet members to enter into bi- or multi-lateral discussions. The committees are invited to e-meet at least on a monthly basis. Reporting of each committee meeting should be available on the website and during the monthly VGMs to be invited by the EB.

## Date and place of next meeting

Due to lack of time, the date and place for the next EMBnet AGM is delayed till the next VGM. Candidates are invited to present their projects.

## Concluding remarks

The chairman concludes by asking each committee to lose no time in starting their work, and specially start e-meeting to discuss the various points of action. Much activity has taken place over the last period, which indicates that dedication among the members is higher than ever. However, *EMBnet.news* has to be produced with no further delay. The EB will consider allocating committee projects some funds from the EMBnet Stichting accounts to help them complete their missions. Quick guides have to be produced also. Federated efforts have new areas of collective interest and should be pursued. The chairman thanks his colleague organisers from Finland and all participants for their efforts. He wishes everybody a safe trip back home.

The meeting is closed at 19:00. Participants disperse for the dinner in the nice Smorgasbord restaurant on the boat. Tomorrow, all participants will join the EMBRACE workshop in Helsinki. The return trip to Stockholm on the Gabriella was a delightful experience in serenity and beauty as the nighttimes' landscape at this time of the year is really unique.

Robert Herzog

Discussion on board the ship Gabriella

# Costa Rica Natioanl Node

**Allan Orozco**

Professor in Bioinformatics and Computer Biology, University of Costa Rica, School of Medicine, Department of Pharmacology and Clinic Toxicology (UCR), Office N3., Santo Domingo de Heredia, America Central, Costa Rica.
kelltron@costarricense.cr

## UNIVERSITY OF COSTA RICA

The University of Costa Rica (UCR) was created as an autonomous institution of higher education dedicated to teaching, learning, research, artistic creation and professional development. Its primary mission has been to advance the social, economic and cultural transformations of Costa Rica society, through a policy for establishment of an integral development. In this sense, the University of Costa Rica (UCR) has been an exceptional promoter of innovations and a pioneer in the creation of academic programs that have contributed significantly to national regional development.

The UCR is constituted for 96 different schools that offer almost 100 professions, 25 Centres of Investigation, 11 Institutes of Investigation and many programs of investigation. The UCR also has a TV channel, radio station, a newspaper and many journals.

The School of Medicine was established in 1960 and is one of the most important schools of the University and the most important School of Medicine of Central America. The School of Medicine has almost 600 teachers and more than 800 students, who work together with five different hospitals, three are Class A hospitals. We offer 51 medical specialities and three master's degrees (pharmacology, physiology and biochemistry).

Currently, our institution houses 12 research Institutes, 26 research Centres, 2 experimental stations, 5 biological natural reserves and 2 support units. There is a group of bioinformaticians who are developing different activities in Bioinformatics with the support of the School of Medicine.

Finally, the Department of Pharmacology and Clinic Toxicology with support from the Government of Costa Rica (2004/2006) started a process of national transformation to impulse the Bioinformatics, Pharmacogenomics and Computational Biology in Costa Rica. The same is the Costa Rican national node (national reference) for Bioinformatics. We at the Department of Pharmacology of the School of Medicine also have a server (Cluster's of 4 servers in network) and all necessary conditions for its function (EMBnet node). We provide access to EMBOSS, wEMBOSS and jEMBOSS and the most popular bioinformatics tools. Training and Bioinformatics courses on all the aspects related to Bioinformatics are organized by the Department of Pharmacology and Clinic Toxicology for our scientific community.

Website:
http://www.dftc.ucr.ac.cr/

# ILRI/BECA Specialist Node

**Dr Etienne de Villiers**

Bioinformatics Group Leader, International Livestock Research Institute (ILRI), Biosciences eastern and central Africa, Joint Bioinformatics Platform Nairobi, Kenya
e.villiers@cgiar.org

## The International Livestock Research Institute (ILRI)

ILRI is a non-profit-making and non-governmental organization headquartered in Nairobi, Kenya. ILRI works at the crossroads of livestock and poverty, bringing high-quality science and capacity-building to bear on poverty reduction and sustainable development in Africa, Asia, Latin America and the Caribbean. Agricultural research at ILRI and its partners is aimed at producing healthier livestock and crops to alleviate poverty and hunger in the developing world through exploitation of the latest genome technologies.

A high performance computing (HPC) facility has been established on ILRI's Nairobi campus to serve the bioinformatics computing needs of ILRI and its partners in the Consultative Group on International Agricultural Research (CGIAR), National Agricultural Research Institutes, and African researchers under the Biosciences eastern and central Africa (BecA) umbrella. The ILRI HPC is linked to other HPC facilities at several research centres of the CGIAR in Asia and South America. By sharing the computational power of the HPC system, researchers will be able to conduct more extensive and large-scale genomic research quickly and cost-effectively.

Current ILRI bioinformatics activities are focused on a tick-transmitted intracellular protozoan parasite, known as Theileria parva. This parasite causes East Coast fever (ECF), which kills a million cattle a year in 11 countries in Africa and is responsible for up to half of all deaths of calves kept by pastoralists there. Scientists from ILRI, Nairobi, and The Institute for Genomic Research (TIGR), based in the Maryland, USA, recently sequenced the genome of the T. parva parasite and most of the subsequent annotation work was performed by two bioinformatics specialists at ILRI. This world-class scientific breakthrough was published in top scientific journal Science. *ILRI HPC Facility Website:* `http://hpc.ilri.cgiar.org/`

## Biosciences eastern and central Africa (BecA)

BecAis the first of four regional networks of centres of excellence in biosciences, supported by the New Partnership for Africa's Development (NEPAD). BecA consists of a Hub located on ILRI's Nairobi campus that will provide a common biosciences research platform, research-related services and capacity building and training opportunities; and a network of regional nodes and other laboratories distributed throughout eastern and central Africa for the conduct of research on priority issues affecting Africa's development. BecA is being established amongst a group of cooperating institutions that agree to make their facilities available for regional use.

The BecA Bioinformatics Platform is located at the newly established BecA Hub on ILRI's Nairobi campus. The purpose of the BecA Bioinformatics Platform is to provide scientific informatics support for research projects led by universities and other research institutes in eastern and central Africa that will utilise technologies available through the BecA platform. Current activities are focused on raising awareness of the importance and utility of bioinformatics and building regional bioinformatics capacity by providing introductory training for early-career scientists and giving skilled bioinformaticians ready access to advanced tools, support and expertise. Local universities are working hard to build capacity but are unable to award degrees in bioinformatics at present. Therefore, the current thrust is to explore partnerships with leading bioinformatics institutes to make undergraduate and postgraduate training possible, by linking up with universities with well-established training programs in bioinformatics to offer masters and doctoral degrees, possibly through distance learning, to students affiliated with the BecA platform. *BecA Bioinformatics Platform Website*: `http://www.becabioinfo.org/`

# Pakistan National Node

**Raheel Qamar**

Chairman & Professor of Biosciences, COMSATS University, Islamabad, Pakistan
raheel _ qamar@comsats.edu.pk

## DEPARTMENT OF BIOSCIENCES COMSATS University

The Department of Biosciences was established a few years back at the COMSATS University (CU), Islamabad Campus. The Department started with the BS Bioinformatics programme in 2003 and currently has more than 125 students. This programme is becoming very popular in the country and will fulfil the future needs of Pakistan in this very important field of study. In addition to the BS programme the department has started MS/PhD programme's in Biochemistry/Molecular Biology, Microbiology, Immunology, Molecular Genetics and Developmental Biology. For all these programmes we have a highly qualified staff consisting of more than 30 members of which 12 have PhD's and are actively involved in research in different domains of Biosciences.

The EMBnet Pakistan National Node is situated at the Department of Biosciences, COMSATS University (CU), Islamabad. The purpose of this node is to promote education and training in the field of Bioinformatics within the country. Moreover, we are aiming at providing Bioinformatics services locally. Raheel Qamar is the Node Manager and Shahid Chohan is Node Staff and both are working together towards these goals.

### Promoting the EMBnet

The news of the election of CIIT as the EMBnet Pakistan National Node has been published in three leading English Daily Newspapers of the country, also highlighting the EMBnet and its role in promotion of Bioinformatics internationally. An on-line report can be accessed at the following URL: http://www.paktribune.com/news/index.shtml?155646

### Promoting Bioinformatics

We have written two popular science articles on Bioinformatics for general audience, for publication in the local media. These articles are currently available on-line on Raheel's website: www.raheelqamar.com.

### Providing Training Locally, an International Workshop on Bioinformatics

During August-September 2006 Raheel and Shahid co-organized an international workshop on Bioinformatics in Islamabad with the title: "Use of Bioinformatics in Genomic Research". Twenty participants came from twelve countries and the resource persons from three. The highlight of this two-week long workshop was plenty of hands on practice.

### A CD for the Participants:

All the participants were provided with a CD containing all the course material, the lectures and exercises, along with the necessary software. The content of this CD will be made available on-line in the near future, on Raheel's website mentioned above.

### Providing Bioinformatics Services Locally

Currently there is no public server in Pakistan providing Bioinformatics services within the country. We are working closely with George Magklaras, EMBnet Norwegian Node Manager and Technical Management Project Committee(TMPC) member, to set up such a server providing basic Bioinformatics services like: local updating copies of popular nucleotide and protein data bases with SRS, local BLAST searches and the EMBOSS suit. While we are waiting for a high speed machine to arrive, we have set up a test Bioinformatics Server providing the above services. The presence of this server has also been advertised among the local research community and is attracting reasonable interest by the users. This test server was set up with the help of David Judge who was visiting Islamabad as a resource person for a workshop on Bioinformatics.

# Symposium "Grids and Web Services in bioinformatics"

**Eija Korpelainen**

Center for Scientific Computing, Tekniikantie 15 a D, 02100 Espoo, Finland

The mini-symposium "Grids and Web Services in bioinformatics" gathered over 50 bioinformaticians from 25 countries to CSC in June. The symposium provided an update on grids and Web Services, and presented grid user experiences in bioinformatics.

The mini-symposium gave bioinformaticians an update on grids and Web Services. These fast developing technologies offer interesting opportunities for the bioinformatics community, but their usage is hampered by the difficulty of obtaining up-to-date, bioinformatics-focused information. The program therefore pointed out that bioinformatics services are increasingly available also as Web Services, which can be accessed programmatically and combined to workflows using systems like Taverna. Also the newly developed grid infrastructures EGEE and DEISA were presented, with several bioinformatics user experiences. As grid and Web Services technologies are approaching each other, the importance of interoperability, standards, and best practices was emphasized. The program of the mini-symposium is shown below:

- Web Services, interoperability and standards (Christian Bryne, University of Bergen)

- Taverna now and in the future (Katy Wolstencroft, University of Manchester)

- Update on EMBRACE and EMBOSS (Peter Rice, EBI)

- MRS (Gert Vriend, CMBI)

- Introduction to grids (Taavi Hupponen, CSC)

- Large-Scale Profile-HMM searches on the Grid (Laurent Falquet, SIB)

- High-throughput bioinformatics analysis on the grid: GROCK (Jose Valverde, CNB)

- In silico docking on grid infrastructures (Jean Salzemann, CNRS)

- Bioinformatics grid in an industrial environment (Richard Kamuzinzi, Université libre de Bruxelles)

The symposium was kindly sponsored by the EU-funded EMBRACE project, which aims to build a bioinformatics service Grid in Europe. Several speakers were also provided by the EMBnet, a worldwide network of bioinformatics service providers.

For the symposium presentations and further information, please see:

`http://www.csc.fi/molbio/opetus/embrace/gridWS.html`

`http://www.embracegrid.info`

`http://www.embnet.org`

In between the lessons...

# WebLab: a bioinformatics platform

**Jianmin Wu,**
**Xiaoqiao Liu,**
**Ge Gao,**
**Lei Kong,**
**Zhe Li, Jingchu Luo**

Center of Bioinformatics, Peking University, Beijing 100871, China

"Half a day on the Web, saves you half month in the lab" paraphrases a quotation proposed by Alan Bleasby, the former EMBnet UK node manager and one of the key developers of EMBOSS. We developed a bioinformatics platform WebLab (`http://weblab.cbi.pku.edu.cn/`). The main goal of WebLab is to provide the end-users in the molecular biology community with a single and simple Web interface by integrating various sequence analysis packages including the European Molecular Biology Open Software Suite (EMBOSS), the NCBI database search tool BLAST, the phylogenetic analysis package Phylip, the multiple sequence alignment program ClustalW, as well the KEGG orthology annotation system (KOBAS) we developed locally. The Figure 1 shows the WebLab architecture.

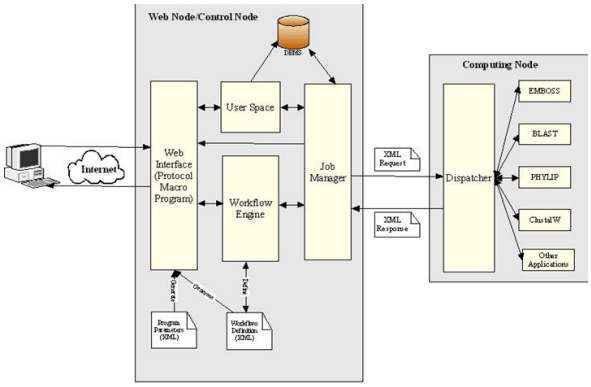



Figure 1. WebLab Architecture, the arrows show how information flows among the system.

In addition to the typical cut-and-paste and browse-and-upload approach, experienced users have an option to retrieve sequence data directly through accession numbers or sequence IDs from the databases indexed on the server. Users can also edit their data such as the description line of FASTA format files within the system. The output results such as ClustalW multiple alignments can be shown either as original plain text, or graphical format with colour fonts (Fig. 2).

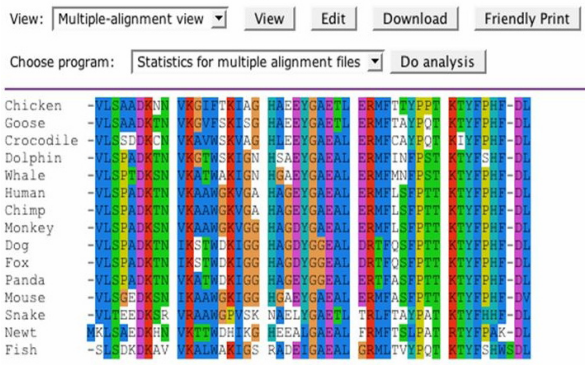



Figure 2. An example of WebLab output.

WebLab creates a user space for registered users to upload input data and store output results. This is extremely useful for novice users since they can manipulate their data in the tree-like folder and file system. In addition, biologists working in the same group or on the same project may share their data with either read-only or read-write permission (Fig. 3).

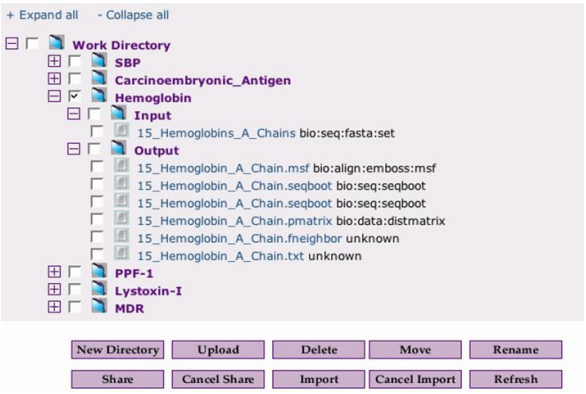



Figure 3. User data space.

WebLab implemented sequence analysis protocols such as phylogenetic tree construction and protein function prediction. Users may follow

the built-in protocols to do step-by-step analysis for single DNA or protein sequence, such as ORF finding, translation, profile scaling, similarity search, motif scanning to predict potential function of the sequence, or start from multiple sequence analysis, then do neighbour-joining tree construction. To make the routine work easier, WebLab also applied the macro approach to make these analyses fully automatic. Other protocols are being added.

WebLab can also be used by bioinformaticians as a workflow-based system for further development. Flexibility is carefully considered in design and adding new tool only needs writing an XML file. Extension mechanism is supported for renders to format input or output data. Bioinformatics developers can plug-in their own extension classes to the platform. WebLab was developed using the platform-independent Java language. Apache tomcat was used as container for Java Servlet and JSP.

WebLab has been publicly available since Oct 2004 with more than 1000 registered users around the world. We have also been using WebLab for our semester graduate course "Applied Bioinformatics Course" (`http://abc.cbi.pku.edu.cn`) with some 200 students each year from Peking University and the Chinese Academy of Agricultural Sciences.

# YeastBASE @ CSC

**Per Harald Jonson**

CSC- the Finnish IT Center for Science, P.O. Box 405, 02101 Espoo, Finland

## Description

YeastBASE is a new service at CSC that was developed during a project financed by CSC and the Finnish Funding Agency for Technology and Innovation (TEKES). The aim was to bring together microarray data scattered in different sources for easy access, search, and retrieval, and to offer this data in a suitable form for modelling and other data analysis purposes.

At present most scientific journals require microarray data to be submitted to a public repository before publication. However, the data from earlier publications are generally not found in these repositories, but on one of the author's home pages, as an appendix on the publisher's pages, or not at all. This makes data hard to find and hence of limited use.

YeastBASE contains microarray data from 86 publications with a total of 1398 hybridisations. YeastBASE is accessible using a standard browser and BASE version 1.2.16 (Saal et al., 2002) was used for data input.

## Content

YeastBASE contains microarray data from 86 publications with a total of 1398 hybridisations. We have got permission for distribution of the data for academic & non-commercial use from the authors of each paper.

## Use

To get to YeastBASE point your browser to `http://yeastbase.csc.fi/`. Clicking the "Accept terms & search" button will then take you top the search

window (Figure 1). Two user cases are shown below in order to illustrate the use of YeastBASE. Each sample is annotated, so information about the genotype and growth condition is searchable. Details about each publication are also searchable. All searches returns the entire experiment, so not all samples necessarily match the search criteria. This is in order to keep the samples in its experimental context. An example of search output is given in Figure 2.

## Case 1 – Data from external stress conditions

All cells respond to changes in their environment. Much of the data in YeastBASE come from experiments where cells have been stressed by changes like elevated temperatures, high/low ionic strength, high osmotic pressure etc. To find e.g. experiments where the media is changed by addition of sorbitol, write sorbitol in the field for media and perform the search. You will find 2 experiments. You will also see that both papers have studied other stress responses as well. It is then possible to download the data after selecting the chips you want. In this case it should be done twice, as it is only possible to download either raw data or author ratio data. The definition of author ratio can be found in the on-line documentation

## Case 2 – Does deletion of sus1 affect the expression of cell wall proteins?

Searches can be made for Gene Ontology (GO) terms in the lower part of the search window. To find genes annotated to be located in the cell

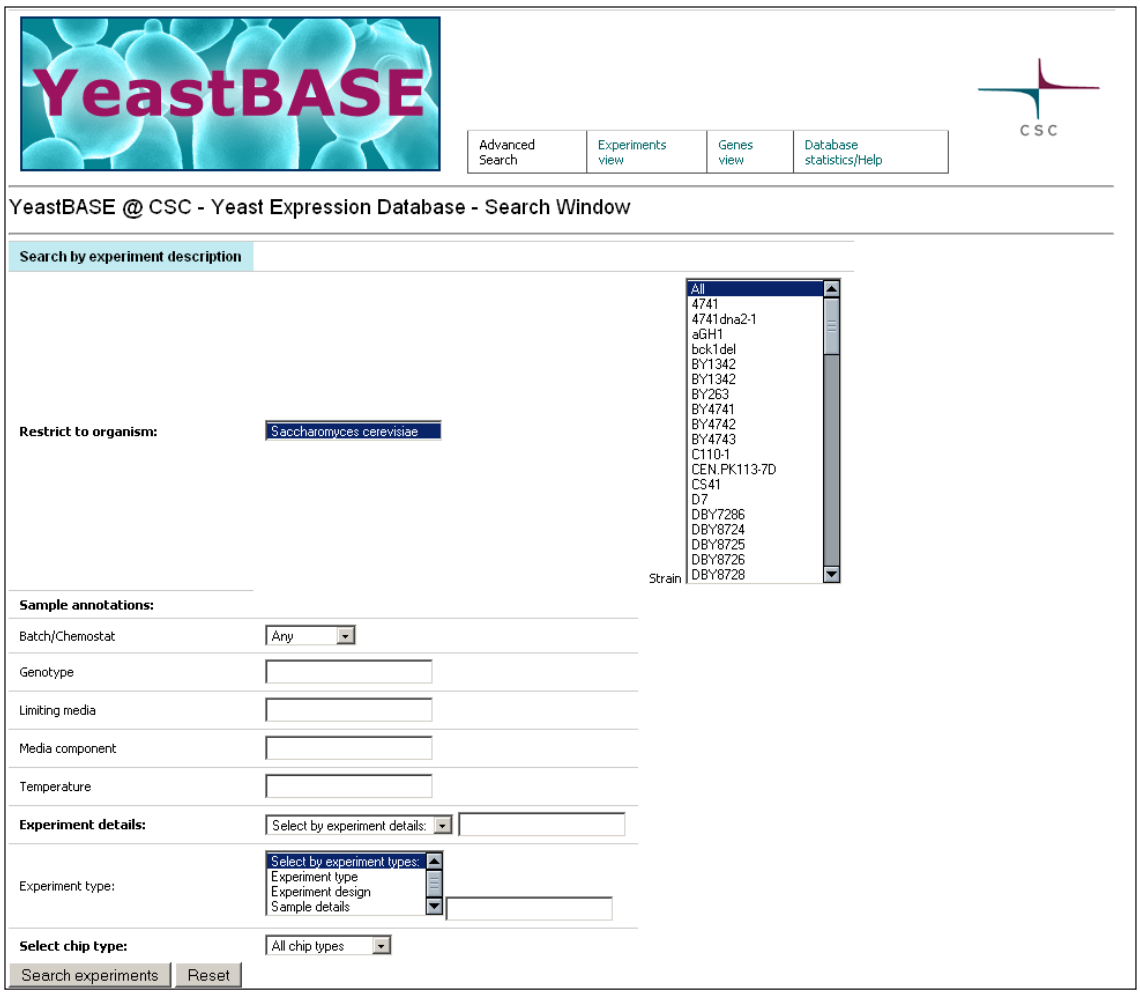



Figure 1: YeastBASE's search window.

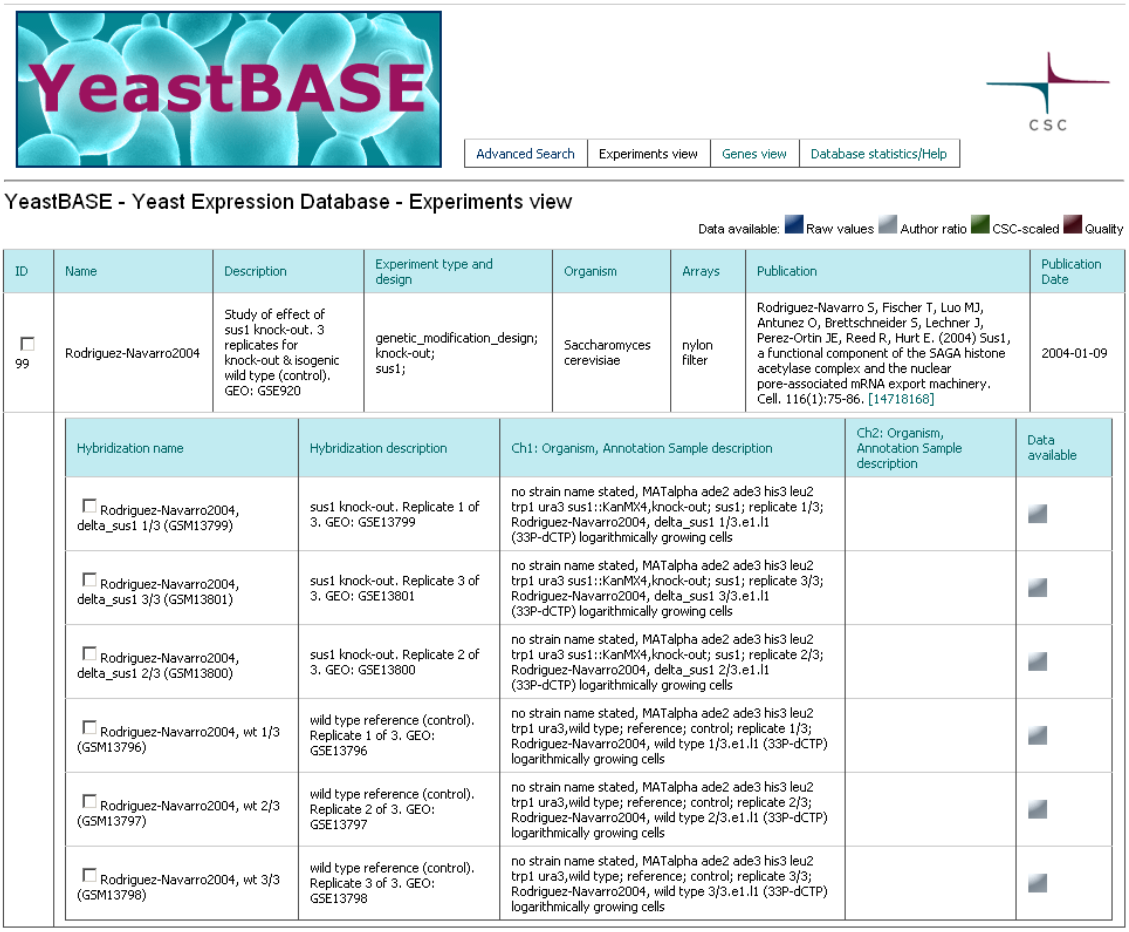



Figure 2: Search result. Select single chips or entire experiments for download.

wall, select the gene annotation "GO Cellular Component", write "cell wall" and press "Look for gene information". The result is a long list of genes. The list, or parts of it, can be saved using the "Download selected data" button.

To find experiments where the gene sus1 was deleted, go to the main search page and write "sus1" in the field for Genotype. The search will return 1 experiment with 6 hybridisations – 3 sus1 deletions and 3 wild type controls. If you want expression data only for your cell wall proteins the list of genes (in the YAL001W form) can be pasted into the bottom field. The downloaded file will then contain the values for these genes only.

## Further information

Questions about YeastBASE should be directed to `yeastbase@csc.fi`

## Acknowledgement

We acknowledge the Finnish Funding Agency for Technology and Innovation (TEKES) for financial support and the authors of scientific papers that allowed inclusion of their data in YeastBASE.

## References

Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg Å, Peterson C. (2002) BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. Genome Biol. 3(8):SOFTWARE0003

# MRS and the management of biomolecular databanks

**Robert Herzog**

Belgian EMBnet Node (BEN), ULB Campus de la Plaine, blv. du Triomphe, 1050 Brussels, Belgium

## Introduction

Offering up-to-date biological and sequence-related databanks is among the most basic functionalities a bioinformatics server has to offer. The hardware and management constraints posed by the large size of these datasets and their continued growth scared off more than one. The excellent software SRS, formerly the main product of Lion Bioscience, now taken over by BioWisdom, was widely used in EMBnet circles and elsewhere, but its updating mechanism imposed a significant learning phase, as did the purpose-built programming language *icarus* invented by SRS creator Thure Etzold.

When Maarten Hekkelman inherited a few years ago the management of the SRS server at the Radbout University of Nijmegen from the hands of former SRS manager Jack Leunissen, he found the hurdle too high. He decided to reconsider the problem from the ground up and started working on his version of sequence retrieval software. "Maarten's retrieval system" or "MRS" was born… Maarten's decisions were to use as much as possible existing Unix tools like *make* and the widely known *Perl* programming language. SRS Achilles' heel was the propensity for data- and index files to get out of synchronicity; the first priority was to alleviate this. Maarten decided to build his system so that every databank would consist on files combining directly the data and the corresponding indexes. It took him some months to reach functional software for the management of bio-sequences, but in the meantime his software was also efficiently used in non biological circles for text indexing and mining.

## Getting in touch with MRS

I will try to give the reader some idea about how MRS feels but I urge anybody interested to go to one of the installed MRS servers[1], as direct contact with the software is worth much more than the length of this writing. For the system manager who is interested in installing the software, several sources of information are available, notably the MRS users list[2]. I wrote an "MRS installation Guide" [3]where hopefully most issues that might come up during installation are also explained.

## Requirements for an MRS installation

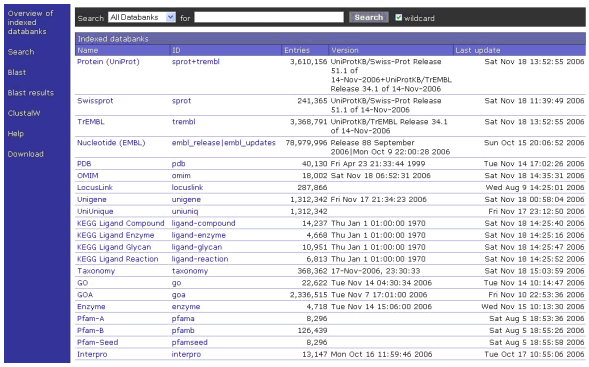
MRS is very much "OS-agnostic". It can be installed on most UNIX infrastructures[4], including the BSD derivative that runs on contemporary MacIntosh. The hardware requirements are proportional to the manager's ambition about which databanks he needs to offer. A server meant to hold the whole of EMBL and UNIPROT databanks for at least the next year should probably not be started if less that one terabyte of disk storage is at hand. Software-wise, MRS requires a development environment comprising a recent version of the *GCC* and *G++* compilers (preferably 4.0 and up), the GNU *make* utility, a recent version of *Perl* and a working *Apache* server. Root rights are not essential to install MRS, except eventually for setting up the *cronjobs* that are to take care of the update runs. MRS is currently accessible on the project SVN server at `http://developer.berlios.de/projects/mrs/` . The author regularly updates his software and the svn security

1 Radbout University Nijmegen: `http://mrs.cmbi.ru.nl`
Belgian EMBnet Node: `http://bendisk.ulb.ac.be/mrs`
SIMDAT ULB partner: `http://springbok.ulb.ac.be/mrs`
2 MRS Users list: `mrs-user@lists.berlios.de`
3 Can be found on `ftp://ftp.be.embnet.org/pub/biosoft/MRS`
4 I installed an MRS server on my tiny VAIO 11in. laptop under Ubuntu 6.10 and manage the protein databanks and a few motif and profile datasets with ease.

and versioning mechanisms allow staying up-to-date in a very comfortable way.

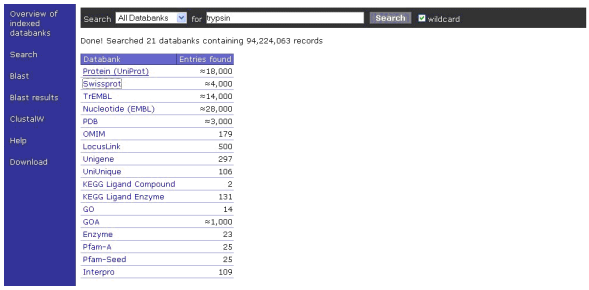
## MRS in action

When opening an MRS server's web interface and clicking on the "Overview of indexed databanks", the next screen capture is typically what you get. You see a list of all installed datasets with the mention of the number of entries, the version and the date of last indexing.



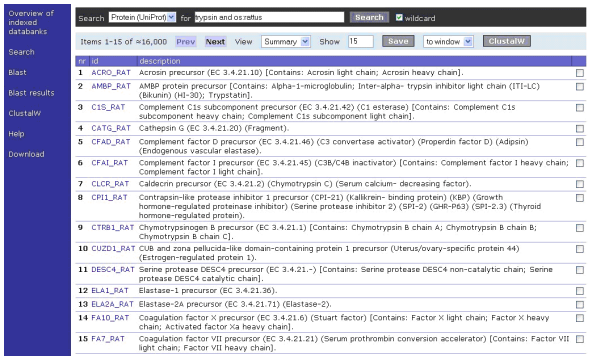

| Name | ID | Entries | Version | Last update |
|---|---|---|---|---|
| Protein (UniProt) | sprot+trembl | 3,610,156 | UniProtKB/Swiss-Prot Release 51.1 of 14-Nov-2006+UniProtKB/TrEMBL Release 34.1 of 14-Nov-2006 | Sat Nov 18 13:52:55 2006 |
| Swissprot | sprot | 241,365 | UniProtKB/Swiss-Prot Release 51.1 of 14-Nov-2006 | Sat Nov 18 11:39:49 2006 |
| TrEMBL | trembl | 3,368,791 | UniProtKB/TrEMBL Release 34.1 of 14-Nov-2006 | Sat Nov 18 13:52:55 2006 |
| Nucleotide (EMBL) | embl_release\|embl_updates | 78,979,996 | Release 88 September 2006\|Mon Oct 9 22:00:28 2006 | Sun Oct 15 20:06:52 2006 |
| PDB | pdb | 40,130 | Fri Apr 23 21:33:44 1999 | Tue Nov 14 17:02:26 2006 |
| OMIM | omim | 18,002 | Sat Nov 18 06:52:31 2006 | Sat Nov 18 14:35:31 2006 |
| LocusLink | locuslink | 287,866 | | Wed Aug 9 14:25:01 2006 |
| Unigene | unigene | 1,312,342 | Fri Nov 17 21:34:23 2006 | Sat Nov 18 00:58:04 2006 |
| UniUnique | uniuniq | 1,312,342 | | Fri Nov 17 23:12:50 2006 |
| KEGG Ligand Compound | ligand-compound | 14,227 | Thu Jan 1 01:00:00 1970 | Sat Nov 18 14:25:40 2006 |
| KEGG Ligand Enzyme | ligand-enzyme | 4,668 | Thu Jan 1 01:00:00 1970 | Sat Nov 18 14:25:16 2006 |
| KEGG Ligand Glycan | ligand-glycan | 10,951 | Thu Jan 1 01:00:00 1970 | Sat Nov 18 14:25:47 2006 |
| KEGG Ligand Reaction | ligand-reaction | 6,813 | Thu Jan 1 01:00:00 1970 | Sat Nov 18 14:25:52 2006 |
| Taxonomy | taxonomy | 368,362 | 17-Nov-2006, 23:30:33 | Sat Nov 18 15:03:59 2006 |
| GO | go | 22,622 | Tue Nov 14 04:30:34 2006 | Tue Nov 14 10:14:47 2006 |
| GOA | goa | 2,336,515 | Tue Nov 7 17:01:00 2006 | Fri Nov 10 22:53:36 2006 |
| Enzyme | enzyme | 4,718 | Tue Nov 14 15:06:00 2006 | Wed Nov 15 10:13:30 2006 |
| Pfam-A | pfama | 8,296 | | Sat Aug 5 18:53:36 2006 |
| Pfam-B | pfamb | 126,439 | | Sat Aug 5 18:55:26 2006 |
| Pfam-Seed | pfamseed | 8,296 | | Sat Aug 5 18:55:58 2006 |
| Interpro | interpro | 13,147 | Mon Oct 16 11:59:46 2006 | Tue Oct 17 10:55:06 2006 |

At the top of the page, a "Search" area allows to compose a search, either in all the databanks installed on a server, or more specifically any of these, chosen from the drop-down list. The text area can receive any simple query, very much like a Goolge search. Note the "wildcard" tick box that allows switching word completion on or off. A search for "trypsin" in all databanks produces the next screen in a matter of seconds.



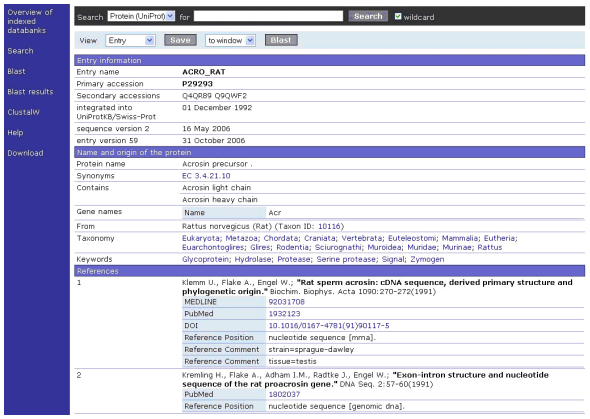

| Databank | Entries found |
|---|---|
| Protein (UniProt) | ≈18,000 |
| Swissprot | ≈4,000 |
| TrEMBL | ≈14,000 |
| Nucleotide (EMBL) | ≈28,000 |
| PDB | ≈3,000 |
| OMIM | 179 |
| LocusLink | 500 |
| Unigene | 297 |
| UniUnique | 106 |
| KEGG Ligand Compound | 2 |
| KEGG Ligand Enzyme | 131 |
| GO | 14 |
| GOA | ≈1,000 |
| Enzyme | 23 |
| Pfam-A | 25 |
| Pfam-Seed | 25 |
| Interpro | 109 |

Once you start knowing the internal structure of the indexed datasets, you can become more specific, like in the following example where the user asked to limit the search to the field « organism species » as defined in UNIPROT. A click on the name of any indexed databank reveals the identifiers of all the indexed fields. Use of the classical Boolean operators allows composing any kind of query string, even quite complicate ones, using parenthesis when necessary.




| nr | id | description |
|---|---|---|
| 1 | ACRO_RAT | Acrosin precursor (EC 3.4.21.10) [Contains: Acrosin light chain; Acrosin heavy chain]. |
| 2 | AMBP_RAT | AMBP protein precursor [Contains: Alpha-1-microglobulin; Inter-alpha- trypsin inhibitor light chain (ITI-LC) (Bikunin) (HI-30); Trypstatin]. |
| 3 | C1S_RAT | Complement C1s subcomponent precursor (EC 3.4.21.42) (C1 esterase) [Contains: Complement C1s subcomponent heavy chain; Complement C1s subcomponent light chain]. |
| 4 | CATG_RAT | Cathepsin G (EC 3.4.21.20) (Fragment). |
| 5 | CFAD_RAT | Complement factor D precursor (EC 3.4.21.46) (C3 convertase activator) (Properdin factor D) (Adipsin) (Endogenous vascular elastase). |
| 6 | CFAI_RAT | Complement factor I precursor (EC 3.4.21.45) (C3B/C4B inactivator) [Contains: Complement factor I heavy chain; Complement factor I light chain]. |
| 7 | CLCR_RAT | Caldecrin precursor (EC 3.4.21.2) (Chymotrypsin C) (Serum calcium- decreasing factor). |
| 8 | CPI1_RAT | Contrapsin-like protease inhibitor 1 precursor (CPI-21) (Kallikrein- binding protein) (KBP) (Growth hormone-regulated proteinase inhibitor) (Serine protease inhibitor 2) (SPI-2) (GHR-P63) (SPI-2.3) (Thyroid hormone-regulated protein). |
| 9 | CTRB1_RAT | Chymotrypsinogen B precursor (EC 3.4.21.1) [Contains: Chymotrypsin B chain A; Chymotrypsin B chain B; Chymotrypsin B chain C]. |
| 10 | CUZD1_RAT | CUB and zona pellucida-like domain-containing protein 1 precursor (Uterus/ovary-specific protein 44) (Estrogen-regulated protein 1). |
| 11 | DESC4_RAT | Serine protease DESC4 precursor (EC 3.4.21.-) [Contains: Serine protease DESC4 non-catalytic chain; Serine protease DESC4 catalytic chain]. |
| 12 | ELA1_RAT | Elastase-1 precursor (EC 3.4.21.36). |
| 13 | ELA2A_RAT | Elastase-2A precursor (EC 3.4.21.71) (Elastase-2). |
| 14 | FA10_RAT | Coagulation factor X precursor (EC 3.4.21.6) (Stuart factor) [Contains: Factor X light chain; Factor X heavy chain; Activated factor Xa heavy chain]. |
| 15 | FA7_RAT | Coagulation factor VII precursor (EC 3.4.21.21) (Serum prothrombin conversion accelerator) [Contains: Factor VII light chain; Factor VII heavy chain]. |

As expected, a click on any line displaying an identifier and description line produces a view of the complete entry, in a nicely enhanced HTML format. Hyperlinks to related databanks are marked up, e.g. to the Taxonomy databank or Medline in the next window.

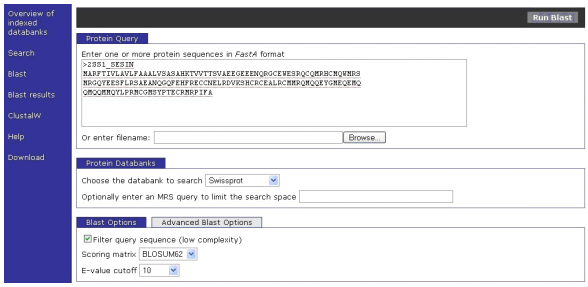



| | |
|---|---|
| Entry name | ACRO_RAT |
| Primary accession | P29293 |
| Secondary accessions | Q4QR89 Q9QWF2 |
| integrated into UniProtKB/Swiss-Prot | 01 December 1992 |
| sequence version 2 | 16 May 2006 |
| entry version 59 | 31 October 2006 |
| Protein name | Acrosin precursor . |
| Synonyms | EC 3.4.21.10 |
| Contains | Acrosin light chain<br>Acrosin heavy chain |
| Gene names | Name: Acr |
| From | Rattus norvegicus (Rat) (Taxon ID: 10116) |
| Taxonomy | Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea; Muridae; Murinae; Rattus |
| Keywords | Glycoprotein; Hydrolase; Protease; Serine protease; Signal; Zymogen |
| 1 | Klemm U., Flake A., Engel W.; **"Rat sperm acrosin: cDNA sequence, derived primary structure and phylogenetic origin."** Biochim. Biophys. Acta 1090:270-272(1991)<br>MEDLINE 92031708<br>PubMed 1932123<br>DOI 10.1016/0167-4781(91)90117-5<br>Reference Position: nucleotide sequence [mma].<br>Reference Comment: strain=sprague-dawley<br>Reference Comment: tissue=testis |
| 2 | Kremling H., Flake A., Adham I.M., Radtke J., Engel W.; **"Exon-intron structure and nucleotide sequence of the rat proacrosin gene."** DNA Seq. 2:57-60(1991)<br>PubMed 1802037<br>Reference Position: nucleotide sequence [genomic dna]. |

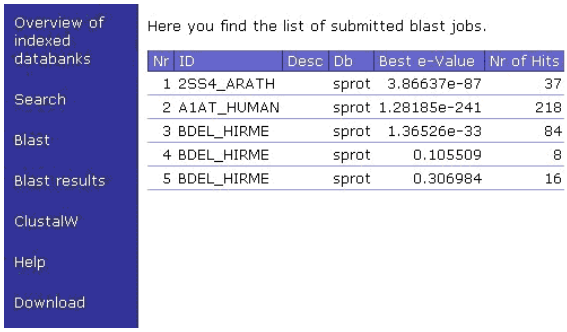
## Built-in BLAST search against protein sequences

Various formats for viewing an entry are available in the top area of this window, but one of the most interesting features of MRS comes with the kind of « built in » BLAST function. Indeed, Maarten produced his own version of the popular BLAST similarity searching program, which does not require purpose-built blast-formatted databanks. Indeed, the search of similarities with your sequence of interest happens directly against the MRS-indexed file. The following screenshots displays the BLAST submission window in the text area of which you either get the entry selected out of the databank or you can paste your own sequence in fasta format. Hitting the "Run Blast" button starts the search. Note that this function-

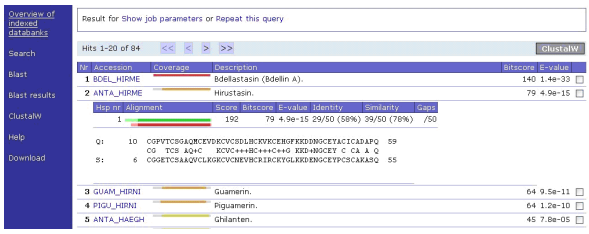
ality is available only for the protein databanks UNIPROT or its sections SwissProt and TrEMBL.




A list of the jobs submitted during the current web session is displayed[5] and hitting the identifier of the submitted sequence opens a nicely enhanced view of the produced hits, as can be seen in the next screenshot.



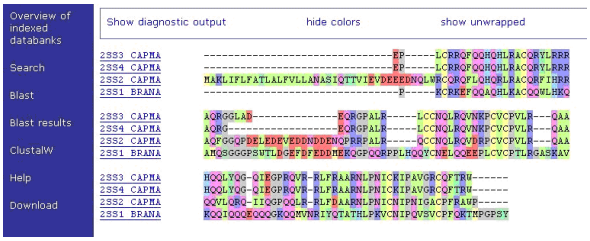

Note that each hit displays a coloured view in the « Coverage » column, the colour of which indicated the level of similarity; it is followed by the description line, the Bitscore and the E-value. Hitting this coloured bar kind of expands the view with the display of the "high scoring pair" in full details.




## Cherries on the cake

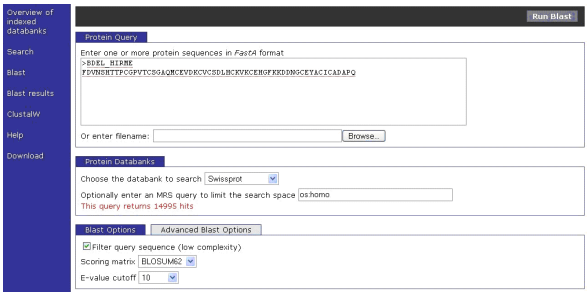
### Cheery # 1

The first of the cherries on the MRS cake is to be found in the last column of the display of this BLAST output, those hits that appear as potentially significant to the user can be marked in the tick boxes, at the right end of each hit. Hitting the "ClustalW" button produces a multiple alignment that users of this software are accustomed with. To get this functionality, the ClustalW package needs to be installed besides MRS on the server.




### Cherry #2

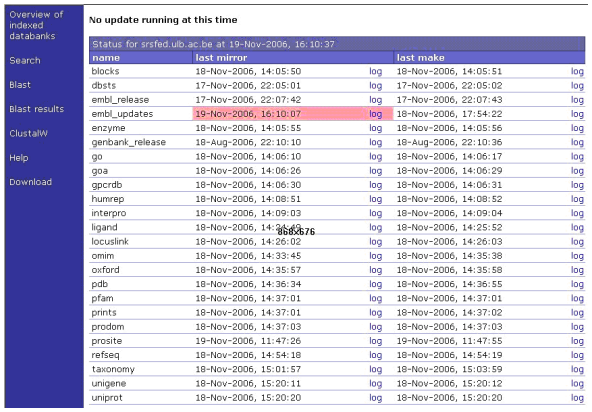
Because the BLAST comparison directly happens against the indexed data file, there is no reason not to take profit of this indexing at run time… In the next screenshot, the user indicates he wants to restrict the BLAST search to only the sequences of a specific organism, in this case *human*. The BLAST windows allows to enter an MRS query which in this case is: "os:homo". The effect of this selection is that the BLAST search is restricted to only the corresponding entries of the databank, with the corresponding changes to the statistics the program produces. This is the reason for the three different e-values in one of the preceding screenshots. Much cuter that to try and keep separate BLAST-formatted databanks for each and every available organism. Note the indication about how many entries will effectively be searched, that appears in, red in the second section of the window




5 If you wonder why BDEL _ HIRME produced various e-values against sprot, read on as the answer is below the subtitle "Cherry 2" below…

## MRS from the system manager's point of view

Space does not allow elaborating on this subject here and now, but let us simply have a look at the next screen that shows MRS during an updating and indexing round. It can be seen that for the embl _ updates, MRS did put the background of the "last mirror" column in red, indicating that this part of the job is currently running . This window gets updated on a regular basis so that following the action on the server is a source of pure delight for the sysadm when everything is working as expected… The updating jobs themselves are managed by a combination of nicely tailored "make" tasks, triggered by simple "cron" jobs. Adding databanks to an MRS server implies writing a corresponding Perl parser and the Perl scripts to mark up the output on the screen. This is mainly an exercise in "informed guessing" when looking at the parsers used for existing datasets.




## MRS and other packages or formats requirements

MRS can easily be used as the source of sequences for other packages. At the Belgian EMBnet Node, we switched already several months ago from SRS to MRS as the sole source of sequences for all the EMBOSS programs. It is also pretty simple to build other datasets while downloading and indexing databanks for MRS. Many examples are already built in and require only uncommenting to jump to live. For instance, we build on a daily basis the derived fasta- and blast-formatted files for the EMBL databank and its updates.

## To conclude

MRS is a very nice system to install and manage. It is a delight both for the end user, by its simple, clear and fast interface, and for the system manager, by its excellent concept and overall ease of maintenance. Its simplicity of concept and its overall tidiness is a welcome change from other databank management systems. And Maarten Hekkelman pursues the developments of his software, notably in areas that should allow easy federation of MRS servers.

# Announcement

## Embrace Workshop on Bioclipse 2007 (EWB '07)

May 23 -25 2007

Uppsala Biomedical Centre (BMC), Uppsala, Sweden

Contents:

Bioclipse is an open source workbench for chemo- and bioinformatics with rich functionality for molecules, sequences, proteins, spectra, and scripts. Bioclipse has advanced plugin architecture which facilitates integration of new functionality, such as algorithms, editors, visualization, Web services, and third party applications. The Embrace Workshop on Bioclipse 2007 (EWB '07) will consist of lectures and hands-on labs to demonstrate the features of Bioclipse, the power of the plugin architecture, and how to integrate new features into the framework.

For more information, see:
`http://teacher.bmc.uu.se/BioclipseWS07`

# Using MRS as sequence retrieval mechanism for EMBOSS

**Guy Bottu**

Belgian EMBnet Node (BEN),
ULB Campus de la Plaine,
blv. du Triomphe, 1050
Brussels, Belgium

## The background

All members of the EMBnet community are without doubt familiar with EMBOSS, a "freeware" software package for sequence analysis, and SRS, a commercial software for indexing and searching databanks. EMBOSS programs can be instructed to operate on sequences residing in databanks. E.g. a user wanting to draw a hydropathy profile of papain could just type the command `pepwindow swissprot:papa1_carpa`. EMBOSS supports a whole variety of methods for retrieving sequences from databanks, which can be located on the same computer as well as on a remote computer accessible over the Internet. EMBOSS of course already from the early days on supported SRS as databank access method. In April 2006 Lion Ltd. decided to sell its Bioinformatics division (together with the copyright of SRS) to Biowisdom Ltd., which greatly reduced the SRS development team. With the further development and even the continuing of academic availability of SRS under a cloud, some people involved in projects as "SRS Federation" and "SIMDAT Pharma" started to search for alternatives. One very promising candidate that popped up was MRS.

MRS stands for Maarten's Retrieval System (the name is an evident paraphrase on Sequence Retrieval System). Its main developer is Maarten Hekkelman, who started working on it while he was at the Centre for Molecular and Biomolecular Informatics of Radbout University (Nijmegen, the Netherlands). After a temporary stay elsewhere he returned on 1 September 2006 with an assignment that allows him to spend a great deal of his time on his favourite brainchild. MRS is "open source" and can be downloaded from berliOS (`http://developer.berlios.de/projects/mrs/`). Just like SRS, MRS has a binary core and a series of auxiliary scripts and configuration files; while SRS uses for these files a proprietary language (Icarus) MRS uses Perl, a scripting language well known to bioinformaticians, which of course makes it easier for local developers. Under MRS the searchable indices are put together with the data (in compressed form) and some extra information in one file. This saves disk space and furthermore, searching and retrieving data has been proven to be faster than with SRS. MRS can be accessed in two different ways: through a Web interface destined to be used interactively by human users and through Web services that allow MRS to be integrated as search engine behind other software[1]. Both access methods rely on a Perl CGI script behind a classic Web server like Apache.

In this article we will explain how MRS can be used as sequence databank access method for EMBOSS.

## Using a WWW server

From release 4.0 on EMBOSS has a databank access method called "MRS", which allows retrieving one or more sequences from a MRS WWW server by entry name or accession number. To define a databank, you must add in the file `.../share/EMBOSS/emboss.default` of the EMBOSS installation (or in the .embossrc file in the home directory of the user) a specification like:

```
DB cmbi_sw [ type: P comment: 'SwissProt at CMBI'
       method: mrs dbalias: sprot format: swiss
       url:  'http://mrs.cmbi.kun.nl/mrs/cgi-bin/mrs.cgi'
]
```

Here "url" is the URL that accesses the MRS home page. In this example we use the "mother" server at the CMBI, but you can choose any other server

1 These Web services are already user for the SIMDAT Pharma project (`http://www.scai.fraunhofer.de/1660.0.html`)

instead. "dbalias" is the name of the databank as it is called in the MRS server. You can find which databanks are available by following the "Overview of indexed databanks" link on the home page.

If you want access to the complete UniProt, you can write instead `dbalias: 'sprot%2Btrembl'`. `sprot+trembl` in MRS syntax means searching both SwissProt and TrEMBL; a further complication is that we must replace the '+' character by its URL escape[2] '%2B'. For databanks that have the release and the updates as separate tables, MRS has a provision for specifying e.g. `emblrelease|embl_updates`. This means that if an entry exists in both the release and the updates, it should be taken from the updates. You must then specify `dbalias: 'emblrelease%7Cembl_updates'`.

Note that this access method works well for retrieving just a few sequences, but if you want to operate on many you might well bump on the time-out of the Web.

## Using the Web services?

An alternative is to use the Web services: it is possible to write an MRS Web Services client in Perl-SOAP and declare it as EMBOSS databank access method of type "application". Our experience is however that the current version of the Web Services is relatively slow and still times out if one tries to retrieve many sequences. At the time of this writing M. Hekkelman is working on a new and more performant version of the Web Services, so that we can better leave this subject in suspense.

## Using a local installation

If you have an installation of MRS on your own computer, you can use this as a very performant sequence retrieval tool. MRS has no equivalent of the SRS "getz" command, but it is not difficult to write one[3]. For example :

```
#!/usr/bin/perl

use MRS;

$ENV{MRS_DATA_DIR} = '/data/mrs';
$db = $ARGV[0];
$allids = $ARGV[1];
$modifier = $ARGV[2];

if ($modifier eq 'gi') {
  $query = "id:$allids|accession:$allids|gi:$allids";}
 elsif ($modifier eq 'id') {
  $query = "id:$allids";
} else {
  $query = "id:$allids|ac:$allids";
}
$dbo = new MRS::MDatabank($db) or die "could not find $db";
$iter = $dbo->Find($query);
while ($id = $iter->Next) {
  $entry = $dbo->Get($id);
  print $entry;
}
```

Note that the environment variable `MRS_DATA_DIR` must be set to the directory where the MRS indices are. Here we do the asignment inside the program (on line 5), so that we do not need to worry about it anymore. If this Perl script is then called `mrsget4emboss` it can be declared as an EMBOSS databank access method of type "application":

```
DB uniprot [ type: P comment: UniProt
             method: app format: swiss
             app: '/usr/local/bin/mrsget4emboss sprot+trembl %s'
]
```

Note that the script supports a third argument, which affects which fields should be searched. For EMBL, where the "id" is actually the same thing as the "ac" it is maybe a good idea to add `id` after the `%s`, in order to win time by searching only one index instead of two. For GenBank the argument `gi` allows to search as well by locus name as by accession number as by NCBI GI number.

2 Since some characters may not be available on the keyboard or have a special meaning in URL syntax (as is the case for the +, which means a space) a character can be replaced by %, followed by its ASCII hexadecimal code.

3 A manual for developers is under construction (`http://mrs.cmbi.ru.nl/doc/`)

# Comparative genomics approach in promoter analysis -

## Orthologous promoter databases and conserved motif search

**Endre Sebestyén[1,2], Tibor Nagy[1], Tamás Pálfy[1], Gábor Tóth[1], Endre Barta[1]**

[1]Agricultural Biotechnology Center, Gödöllö o, Szent-Györgyi Albert u. 4., H-2100, Hungary

[2]Agricultural Research Institute of the Hungarian Academy of Sciences, Martonvásár, Brunszvik u. 2., H-2462, Hungary

## Introduction

The Bioinformatics Group in the Agricultural Biotechnology Center, Gödöllö, Hungary, is especially committed to study the regulation of transcription *in silico*. In the era of comparative genomics it is now possible to compare large genomic regions or whole genomes, both within animals and higher plants.

In order to find evolutionary conserved motifs (putative transcription factor binding sites, TFBSs) in the promoter regions of different plant or chordate genes, we have developed an orthologous promoter database (Database of Orthologous Promoters, DoOP, `http://doop.abc.hu`, Barta et al., 2005). The DoOP database consists of two sections: green plants and chordates. The sections are based on the *Arabidopsis thaliana* and *Homo sapiens* genome annotation, respectively. In both cases we used the annotated first exons to search for orthologous first exons, and extracted the 500, 1000 and 3000 bp upstream (5') regions relative to these first exons as promoters.

Results for keyword homeobox : 182

| No. | Cluster | Description | Motifs | Species |
|---|---|---|---|---|
| 1. | 80100138 | LIM/homeobox protein Lhx4. [Source:Uniprot/SWISSPR | 16 | H P C R M |
| 2. | 80100462 | Pre-B-cell leukemia transcription factor-1 (Homeob | 7 | H P C R F M O |
| 3. | 80100463 | LIM homeobox transcription factor 1 alpha (LIM/hom | 17 | H P C R M |
| 4. | 80100517 | Paired mesoderm homeobox protein 1 (PRX-1) (Paired | 28 | H P C R M O |
| 5. | 80100548 | LIM/homeobox protein Lhx9. [Source:Uniprot/SWISSPR | 17 | H P C G R M |
| 6. | 80100634 | BarH-like 2 homeobox protein. [Source:Uniprot/SWIS | 24 | H P C R M |
| 7. | 80100837 | Homeobox protein aristaless-like 3 (Proline-rich t | 14 | H P R M |
| 8. | 80101012 | Homeobox prospero-like protein PROX1 (PROX 1). [So | 11 | H P C R M |
| 9. | 80101042 | Homeobox protein HLX1 (Homeobox protein HB24). [So | 15 | H P C R M |
| 10. | 80101201 | LIM homeobox 8 [Source:RefSeq_peptide;Acc:NP_00100 | 28 | H P M |
| 11. | 80101688 | diencephalon/mesencephalon homeobox 1 isoform a [S | 3 | H P C G R F M O |
| 12. | 80102549 | Mix-like homeobox protein 1 [Source:RefSeq_peptide | 13 | H M |
| 13. | 80200058 | Zinc finger homeobox protein 1b (Smad interacting | 20 | H R M |
| 14. | 80200138 | Homeobox protein GBX-2 (Gastrulation and brain-spe | 15 | H P R F M |
| 15. | 80200333 | Homeobox protein SIX3 (Sine oculis homeobox homolo | 4 | H C R |
| 16. | 80200340 | Homeobox protein DLX-1. [Source:Uniprot/SWISSPROT; | 18 | H P C G R F M O |
| 17. | 80200341 | Homeobox protein DLX-2. [Source:Uniprot/SWISSPROT; | 24 | H P C R M |
| 18. | 80200366 | similar to Homeobox even-skipped homolog protein | 4 | H P C R M |
| 19. | 80200367 | Homeobox protein Hox-D13 (Hox-4I). [Source:Uniprot | 17 | H P R M |
| 20. | 80200368 | Homeobox protein Hox-D12 (Hox-4H). [Source:Uniprot | 13 | H P C R M |
| 21. | 80200369 | Homeobox protein Hox-D11 (Hox-4F). [Source:Uniprot | 12 | H P C R M |
| 22. | 80200370 | Homeobox protein Hox-D10 (Hox-4D) (Hox-4E). [Sourc | 18 | H P C G R M O |
| 23. | 80200371 | Homeobox protein Hox-D9 (Hox-4C) (Hox-5.2). [Sourc | 17 | H R M |
| 24. | 80200372 | Homeobox protein Hox-D8 (Hox-4E) (Hox-5.4). [Sourc | 27 | H R M |

Figure 1. Search result ("homeobox" keyword search) from the DoOP database with the most important properties of the clusters: cluster ID, a short description, number of evolutionary conserved motifs, and icons representing the most important taxons.

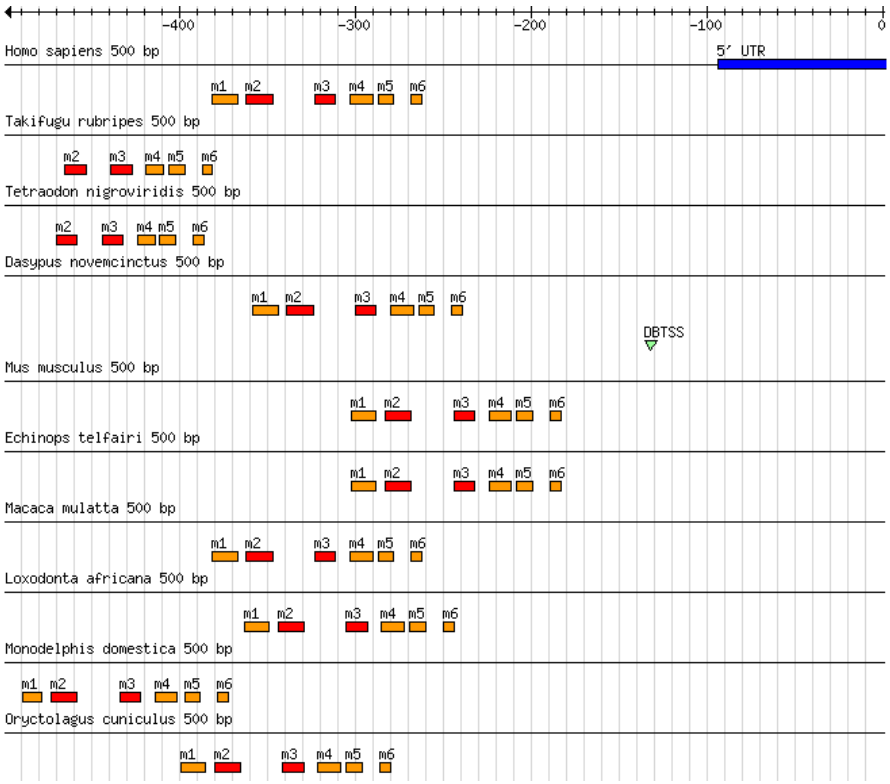



Figure 2. Graphical representation of a cluster (homeobox protein SIX6), showing the orthologous promoter sequences, the annotated 5′ UTRs and transcription start sites, and the evolutionary conserved motifs.

## Construction and content of the database

The two sections of the database are based on the Arabidopsis thaliana and Homo sapiens genome builds from NCBI. We use the annotated first exons of these reference species in the BLAST queries to search for orthologous sequences. The local BLAST databases are generated from the gss, htgs, nt and wgs sections of the NCBI BLAST database and the whole genome sequences available at the ENSEMBL database, the Broad Institute and the Joint Genome Institute. After defining the most orthologous hits, we extract 500, 1000 or 3000 bp upstream sequences if available.

To find the evolutionary conserved motifs in a given set of orthologous promoters, we perform a multiple local alignment with the program DIALIGN. Based on this alignment the motifs are extracted using Perl scripts.

The current plant section (version 1.5) contains 7324 orthologous promoter sets (clusters) with at least two sequences up to 500 bp lengths. There are approximately ~88k conserved motifs in the plant section.

The chordate section (version 1.3) contains 22 846 clusters with 500 bp upstream sequences, with about ~1776k conserved motifs.

## Data access and features of the web interface

The data of the clusters (genes) are accessible through the DoOP website (`http://doop.abc.hu`). After selecting the appropriate section (chordate or plant), users can search in the database with different methods. There are eight search types available: users can enter a specific cluster or sequence id, a keyword to search in the annotation of the clusters, a Gene Ontology ID or term, a scientific species name, an At number in the case of plants or the ENSEMBL gene ID in the case of chordates, and finally a promoter or gene sequence to run a BLAT search. On the search result page (Figure 1.) users can select a given cluster, and examine it in detail. The clus-

| No. | Cluster ID | Description | Score | Ext. score | Query | Hit | Query start | Hit start | Motif type | Alignment |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. | 80700785 | Homeobox protein DLX-6. [Source:Uniprot/SWISSPROT; | 50 | 60 | AtcCacCttAaA | ATCCACCTTAAA | 1 | 61 | | Alignment |
| 2. | 81000435 | KIAA0261 [Source:RefSeq_peptide;Acc:NP_055860] | 50 | 60 | AtcCacCttAaA | ATCCACCTtAAA | 1 | 4 | | Alignment |
| 3. | 80100189 | Cytosolic phospholipase A2 (cPLA2) (Phospholipase | 50 | 55 | tcCacCttAaA | TCCACCTTAAA | 2 | 3 | | Alignment |
| 4. | 80100637 | Zinc finger protein 644 (Zinc finger motif enhance | 50 | 55 | tcCacCttAaA | TCCACCTTAAA | 2 | 10 | | Alignment |
| 5. | 80900656 | Osteomodulin precursor (Osteoadherin) (OSAD) (Kera | 50 | 55 | tcCacCttAaA | TCcACcTTAAa | 2 | 11 | | Alignment |
| 6. | 82400746 | Proto-oncogene DBL (Proto-oncogene MCF-2) [Contain | 50 | 55 | tcCacCttAaA | TCCACCTTAAA | 2 | 27 | | Alignment |
| 7. | 80100498 | Dermatopontin precursor (Tyrosine-rich acidic matr | 50 | 55 | tcCacCttAaA | TCCACCTTAAA | 2 | 6 | | Alignment |
| 8. | 81201263 | Tetraspanin-8 (Tspan-8) (Transmembrane 4 superfami | 50 | 55 | tcCacCttAaA | tCCaCCTTaAa | 2 | 45 | | Alignment |
| 9. | 80100185 | odorant response abnormal 4 [Source:RefSeq_peptide | 50 | 55 | tcCacCttAaA | TCCACCTTAAA | 2 | 203 | | Alignment |
| 10. | 80300391 | Ceruloplasmin precursor (EC 1.16.3.1) (Ferroxidase | 50 | 55 | tcCacCttAaA | TCCACCtTAAA | 2 | 10 | | Alignment |
| 11. | 80500284 | Proteinase activated receptor 2 precursor (PAR-2) | 50 | 55 | tcCacCttAaA | tCCACcTTAaA | 2 | 25 | | Alignment |
| 12. | 81900580 | Cyclic-AMP-dependent transcription factor ATF-5 (A | 50 | 55 | AtcCacCttAa | ATCCACCTTAA | 1 | 1 | | Alignment |
| 13. | 80101418 | Carbonic anhydrase XIV precursor (EC 4.2.1.1) (Car | 50 | 51 | AtcCacCttAaA | AGCCACCTTAAA | 1 | 7 | | Alignment |
| 14. | 81900665 | similar to zinc finger protein 113 | 50 | 51 | AtcCacCttAaA | ATCCACCTTACA | 1 | 6 | | Alignment |
| 15. | 81700157 | Anaphase promoting complex subunit 11 (APC11) (Cyc | 50 | 51 | AtcCacCttAaA | AACCACCTTAAA | 1 | 129 | | Alignment |
| 16. | 80900191 | hypothetical LOC401522 | 50 | 51 | AtcCacCttAaA | ATCCACCTTACA | 1 | 21 | | Alignment |
| 17. | 80900201 | LOC441423 | 50 | 51 | AtcCacCttAaA | ATCCACCTTACA | 1 | 21 | | Alignment |
| 18. | 81200074 | C-type lectin domain family 9, member A [Source:Re | 50 | 51 | AtcCacCttAaA | ATCCACCTTAGA | 1 | 7 | | Alignment |
| 19. | 81701398 | hypothetical LOC388358 | 50 | 51 | AtcCacCttAaA | ACCCACCTTAAA | 1 | 14 | | Alignment |
| 20. | 80900656 | Osteomodulin precursor (Osteoadherin) (OSAD) (Kera | 50 | 51 | tcCacCttAaA | TCcACcTTAAR | 2 | 11 | | Alignment |
| 21. | 80100598 | Synaptotagmin-2 (Synaptotagmin II) (SytII). [Sourc | 50 | 50 | cCacCttAaA | CCACcTTaaA | 3 | 76 | | Alignment |
| 22. | 80800423 | hypothetical gene supported by BX537900 | 50 | 50 | cCacCttAaA | CCACCtTAAA | 3 | 309 | | Alignment |
| 23. | 81100693 | similar to CDNA sequence BC021608 | 50 | 50 | tcCacCttAa | TCCACCTTAa | 2 | 71 | | Alignment |
| 24. | 81900825 | NACHT-, LRR- and PYD-containing protein 11 (PYRIN- | 50 | 50 | tcCacCttAa | TCCACCTTAA | 2 | 106 | | Alignment |
| 25. | 80200250 | Orphan nuclear receptor NR4A2 (Orphan nuclear rece | 50 | 50 | cCacCttAaA | CCACCTTAAA | 3 | 12 | | Alignment |
| 26. | 80300925 | hypothetical LOC389100 | 50 | 50 | cCacCttAaA | CCaCCTTAAA | 3 | 2 | | Alignment |
| 27. | 80800200 | Mdm2, transformed 3T3 cell double minute 2, p53 bi | 50 | 50 | AtcCacCttA | AtcCaccTtA | 1 | 9 | | Alignment |
| 28. | 81101509 | Plakophilin 3. [Source:Uniprot/SWISSPROT;Acc:Q9Y44 | 50 | 50 | tcCacCttAa | TCCACCTTaA | 2 | 39 | | Alignment |
| 29. | 82000407 | Tyrosine-protein phosphatase non-receptor type sub | 50 | 50 | tcCacCttAa | TCCACCTTAA | 2 | 2 | | Alignment |

Figure 3. Result of a MOFEXT search against the DoOP motif collections. The table shows the ID and description of the cluster from which the conserved motif originates, the score, sequence and position of the hits, and the type of the motif.

ter page starts with the 500 bp sequence collection, but the 1000 or 3000 bp collection also can be selected. It contains detailed information on the cluster in question. Besides the annotation of the reference species' sequence, a graphical representation of the cluster is available (Figure 2.). The sequences of the clusters, the multiple alignment and the sequences of the conserved motifs are also available for download.

To facilitate the search through the promoter collection of the DoOP database, or the millions of conserved motifs, we have developed a new program, called MOFEXT (MOtiF search and eXTension). The MOFEXT program source code or binary version is available upon request. MOFEXT is a new tool designed specifically to search in the conserved motif collections (consensus sequences) of plant and chordate orthologous promoters derived from the DoOP database. The program searches in the collections of consensus sequences, performing a gapless local alignment. Although the primary aim of developing MOFEXT was to find motifs similar to experimentally verified transcription factor binding sites, the program is also capable of searching other DNA or protein motif collections. To facilitate the use of MOFEXT in transcription regulation studies, we have built a specific web site (`http://doopsearch.abc.hu`) to search and analyse conserved motifs available from the DoOP database.After choosing the chordate or plant section on the start page, users can define the search parameters and select different promoter and motif collections. The result of a MOFEXT search (Figure 3.) contains a list of similar motifs with details on the cluster from which the motif originates. For a pattern search in the whole 500, 1000 or 3000 bp promoter regions, the FUZZNUC program from the EMBOSS software package is also available in the website. The FUZZNUC result page is similar to the MOFEXT result page.

## Future development

We continuously update and develop our website. Due to the growing number of whole genome annotations and available genome sequences, the number of orthologous promoter sets, sequences and quality of the conserved motif collections will improve. We try to keep up with these changes and integrate them into our website. New features like the annotation of known transcription factor binding sites are planned to be implemented in the near future. Suggestions and comments are welcome at the `doop@abc.hu` address.

## References

Barta E, Sebestyén E, Pálf y T B, Toth G, Ortutay C P, Patthy L. DoOP: Databases of Orthologous Promoters, collections of clusters of orthologous upstream sequences from chordates and plants. *Nucleic Acids Res* 33: D86-90 (2005).

# Skin-deep

**Vivienne Baillie Gerritsen**

**The colour of human skin has been – and still frequently is – at the heart of violent controversy. Political, social and physical. Yet, as the science of human genetics unfolds, we are reminded over and over again that any given human population cannot be defined according to its pigmentation since any skin hue blends gradually into another. However, there is no doubt that there are dark skins, and there are light ones. The darkness – as the lightness – of skin depends on the amount of melanin present in the epidermal cells. And the amount of melanin depends directly – though not solely – on the existence of a protein that has been christened 'solute carrier family 24 member 5' or 'SLC24A5'.**

The Immigrants, Sue Jarvis

Scientists have been searching for genes which have a direct influence on skin pigmentation for over a century. Though many involved in various forms of human albinism[1] have been discovered – and over one hundred involved in the coat colours of model organisms such as mice – to date no one had found one gene that had a direct effect on skin colour. The colour of our skin depends on the amount of the pigment melanin present in our epidermal cells. The pigment is found in organelles known as melanosomes, which in turn are found in melanin-producing cells: the melanocytes. Epidermal melanocytes use dendritic protuberances to fire melanosomes into our skin cells, or keratinocytes. And it is the concentration of these skin cell melanosomes which will lend a dark or not so dark hue to the bearer.

Is there a biological advantage in owning a dark skin, or a light skin? Melanin-rich skins – i.e. dark skins – have a greater protection against the sun's harmful ultraviolet rays than melanin-poor skins do. Which is the reason why darker skins are found where the sun hits most. Melanosomes in darker skins actually cap the nucleus of skin cells, thereby protecting the cell's DNA from UV irradiation. On the other hand, a lighter skin – that would be found in the more Northern latitudes – is needed to let sufficient ultraviolet rays through to synthesize vitamin D, which is important for bone growth.

SLC24A5 was discovered in the zebrafish while searching for genes involved in skin cancer. Zebrafish are small freshwater fish found in India and Burma. They have been part of laboratory research since the 1970s because – besides being vertebrates like us – their epidermis and their eggs are transparent, and embryonic development can be observed and tinkered with at every stage. Likewise, even melanocytes and their melanosomes are visible to the naked eye. Whilst dabbling with zebrafish DNA, scientists observed a mutation which turned the zebra-like scale pattern of the fish into a smooth golden one, and the mutation was subsequently named the 'golden gene'. The scientists then turned to humans to see whether

[1] See Spotlight issue August 2004

our genome embraced the same gene. And it does. Moreover, it is so similar that the human version can actually be engineered into the golden zebrafish to give it back its original stripy appearance, proving that SLC24A5 has a direct influence on pigmentation.

Dark skins are loaded with melanin, whilst white skins are not. From an ultrastructural point of view, this means that there are fewer and smaller melanosomes per melanocyte in light skins. What decides on a melanosome's morphology and its melanin content in the first place? SLC24A5 seems to have a pivotal role in melanosome morphogenesis as well as melanin synthesis. The protein is probably an ion exchanger lodged in the melanosome membrane, where it is involved in calcium transport. Calcium is used in the process of melanosome maturation. And if its uptake is lessened, melanosomes present a lower content of melanin, giving skin a paler shade of brown.

So what is the difference observed between a Scotsman and a Nigerian in terms of SLC24A5? One amino acid. Threonine replaces alanine at position 111 in the human sequence. And it is this difference which results in a shift in skin colour. 'White-skin' SLC24A5 probably causes the melanosomes to be not only fewer but also smaller in size, and melanin synthesis is suppressed – all because of a change in calcium transport. The exciting discovery is that the light-skin version of SLC24A5 is found in Europeans, whilst the dark-skin version is found in Africans. What is more, populations that are a mixture of white and black 'blood' – such as the African-American and the African-Caribbean populations – share a mixture of the two versions of SLC24A5. In fact, SLC24A5 accounts for 25 to 38% of the European-African skin melanin index.

Everything seems to fit in very nicely. However, light-skinned East Asians share the same version of SLC24A5 as dark-skinned Africans – which only goes to show that SLC24A5 is part of a far more complex molecular pathway. Like all biological processes. Skin pigmentation cannot be reduced to the doings of one sole protein. Melanin-pigment abnormalities – such as skin cancer or some forms of albinism – should benefit from research on genes such as SLC24A5 which could become a novel target in biotechnology. But one of the most interesting aspects of SLC24A5 is the discovery of a gene that sheds a brighter light upon our past and on how our ancestors sauntered from one continent to another.

## Cross-references to Swiss-Prot

Golden protein, *Brachydanio rerio* (Zebrafish) : Q49SH1
Solute carrier family 24 member 5, *Homo sapiens* (Human): Q71RS6

## References


**1.** Lamason R.L., Mohideen M.-A..P.K., Mest J.R., Wong A.C., Norton H.L., Aros M.C., Jurynec M.J., Mao X., Humphreville V.R., Humbert J.E., Sinha S., Moore J.L., Jagadeeswaran P., Zhao W., Ning G., Makalowska I., McKeigue P.M., O'Donnell D., Kittles R., Parra E.J., Mangini N.J., Grunwald D.J., Shriver M.D., Canfield V.A., Cheng K.C.
SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans
Science 310:1782-1786(2005)
PMID: 16357253

2. Sturm R.
A golden age of human pigmentation genetics
Trends Genet. 22:464-468(2006)
PMID: 16857289

3. Müller J., Kelsh R.N.
A golden clue to human skin colour variation
Bioessays 28:578-582(2006)
PMID: 16700060

# National Nodes

## Argentina
Oscar Grau
IBBM, Facultad de Cs. Exactas, Universidad Nacional de La Plata
Email: grau@biol.unlp.edu.ar
Tel: +54-221-4259223 Fax: +54-221-4259223
http://www.ar.embnet.org

## Australia
Sonia Cattley
RMC Gunn Building B19, University of Sydney,NSW, 2006
Email: scattley@angis.org.au
Tel: +61-2-9531 2948
http://www.au.embnet.org

## Austria
Martin Grabner
Vienna Bio Center, University of Vienna
Email: martin.grabner@univie.ac.at
Tel: +43-1-4277/14141
http://www.at.embnet.org

## Belgium
Robert Herzog, Marc Colet
BEN ULB Campus Plaine CP 257
Email: rherzog@ulb.ac.be mcolet@ulb.ac.be
Tel: +32 2 6505146 Fax: +32 2 6505124
http://www.be.embnet.org

## Brazil
Gonçalo Guimaraes Pereira
Laboratório de Genômica e Expressão - IB
UNICAMP-CP 6109
13083-970 Campinas-SP, BRASIL
Tel: 0055-19-37886237/6238
Fax: 0055-19-37886235
Email: goncalo@unicamp.br
http://www.br.embnet.org

## Chile
Juan A. Asenjo
Centre for Biochemical Engineering and Biotechnology (CIByB). University of Chile
Beauchef 861, Santiago, Chile
Tel: +56 2 6715140
Fax: +56 2 6991084
Email: juasenjo@ing.uchile.cl
http://www.embnet.cl

## China
Jingchu Luo
Centre of Bioinformatics
Peking University
Beijing 100871, China
Tel: 86-10-6275-7281
Fax: 86-10-6275-9001
Email: luojc@pku.edu.cn
http://www.cn.embnet.org

## Colombia
Emiliano Barreto Hernández
Instituto de Biotecnología
Universidad Nacional de Colombia
Edificio Manuel Ancizar
Bogota - Colombia
Tel: +571 3165027 Fax: +571 3165415
Email : ebarreto@ibun.unal.edu.co
http://www.co.embnet.org

## Costa Rica
Allan Orozco
University of Costa Rica (UCR), School of Medicine,
Department of Pharmacology and ClinicToxicology
San Jose, America Central
Costa Rica
Email: allanorozco@gmail.com
Tel: +506 2074489
http://www.dftc.ucr.ac.cr/

## Cuba
Ricardo Bringas
Centro de Ingeniería Genética y Biotecnolgía,
La Habana, Cuba
Email: bringas@cigb.edu.cu
Tel: +53 7 218200
http://www.cu.embnet.org

## Finland
Kimmo Mattila
CSC, Espoo
Email: kimmo.mattila@csc.fi
Tel: +358 9 4572708
Fax: +358 9 4572302
http://www.fi.embnet.org

## France
Jean-Marc Plaza
INFOBIOGEN, Evry
Email: plaza@infobiogen.fr
Tel: +33 1 60 87 37 11 Fax: +33 1 60 87 37 96
http://www.fr.embnet.org

## Hungary
Endre Barta
Agricultural Biotechnology Center
Szent-Gyorgyi A. ut 4. Godollo,
Email: barta@abc.hu
Tel: +36 30-2101795
http://www.hu.embnet.org

## India
Akash Ranjan
Laboratory of Computational Biology & Bioinformatics facility, Centre for DNA Fingerprinting and Diagnostics (CDFD), Hyderabad
Email: akash@cdfd.org.in
Tel: +91 40 7155607 / 7151344 ext:1206
Fax : +9140 7155479
http://www.in.embnet.org

## Israel
Leon Esterman
INN (Israeli National Node) Weizmann Institute of Science
Department of Biological Services, Biological Computing Unit, Rehovot
Email: Leon.Esterman@weizmann.ac.il
Tel: +972- 8-934 3456
http://www.il.embnet.org

## Italy
Cecilia Saccone
CNR - Institute of Biomedical Technologies
Bioinformatics and Genomic Group
Via Amendola 168/5 - 70126 Bari (Italy)
Email: saccone@area.ba.cnr.it
Tel. +39-80-5482100 - Fax. +39-80-5482607
http://www.it.embnet.org

## Mexico
Cesar Bonavides
Nodo Nacional EMBnet, Centro de Investigación sobre Fijación de Nitrógeno, Cuernavaca, Morelos
Email: embnetmx@cifn.unam.mx
Tel: +52 (7) 3 132063
http://embnet.cifn.unam.mx

## The Netherlands
Jack A.M. Leunissen
Dept. of Genome Informatics
Wageningen UR, Dreijenlaan 3
6703 HA Wageningen, NL
Email: Jack.Leunissen@wur.nl
Tel: +31 317 484074
http://www.nl.embnet.org

## Norway
George Magklaras
The Norwegian EMBnet Node
The Biotechnology Centre of Oslo
Email: admin@embnet.uio.no
Tel: +47 22 84 0535
http://www.no.embnet.org

## Pakistan
Raheel Qamar
Department of Biosciences, COMSATS Institute of Information Technology, Park Road, Chak Shahzaad Campus, Chak Shahzaad
Islamabad, Pakistan
Email: Raheel _ qamar@comsats.edu.pk
Tel: +0092-333-5119494
http://www.ciit.edu.pk/Departments _ & _ Faculties/Link=DeptDetail&f=Departments%5F%26%5FFaculties&SMID=10

## Poland
Piotr Zielenkiwicz
Institute of Biochemistry and Biophysics
Polish Academy of Sciences Warszawa
Email: piotr@pl.embnet.org
Tel: +48-22 86584703
http://www.pl.embnet.org

## Portugal
Pedro Fernandes
Instituto Gulbenkian de Ciencia
Unidade de Bioinformatica
2781-901 OEIRAS
Email: pfern@igc.gulbenkian.pt
Tel: +351 214407912 Fax: +315 214407970
http://www.pt.embnet.org

## Russia
Sergei Spirin
Biocomputing Group, Belozersky Institute Moscow
Email: sas@belozersky.msu.ru
Tel: +7-095-9395414
http://www.genebee.msu.ru

## Slovakia
Lubos Klucar
Institute of Molecular Biology SAS Bratislava
Email: klucar@embnet.sk
Tel: +421 2 5930 7413
http://www.sk.embnet.org

## South Africa
Heikki Lehvaslaiho
SANBI, University of the Western Cape, Bellville
Email: heikki@sanbi.ac.za
Tel: +27 (0)21 959 2096
http://www.za.embnet.org

## Spain
José M. Carazo, José R. Valverde
EMBnet/CNB, Centro Nacional de Biotecnología, Madrid
Email: carazo@es.embnet.org,
jrvalverde@es.embnet.org
Tel: +34 915 854 505 Fax: +34 915 854 506
http://www.es.embnet.org

## Sweden
Nils-Einar Eriksson, Erik Bongcam-Rudloff
Uppsala Biomedical Centre, Computing Department, Uppsala, Sweden
Email: nils-einar.eriksson@bmc.uu.se
erik.bongcam@bmc.uu.se
Tel: +46-(0)18-4714017, +46-(0)18-4714525
http://www.embnet.se

## Switzerland
Laurent Falquet
Swiss Institute of Bioinformatics, Génopode-UNIL, CH-1015 Lausanne Email: Laurent.Falquet@isb-sib.ch
Tel: +4121 692 4078 Fax: +4121 692 4065
http://www.ch.embnet.org

# Specialist Nodes

## EBI
Rodrigo López
EBI Embl Outstation, Wellcome trust Genome Campus,
Hinxton Hall, Hinxton, Cambridge, United Kingdom
Email: rls@ebi.ac.uk
Phone: +44 (0)1223 494423
http://www.ebi.ac.uk

## ETI
P.O. Box 94766
NL-1090 GT Amsterdam, The Netherlands
Email: wouter@eti.uva.nl
Phone: +31-20-5257239
Fax: +31-20-5257238
http://www.eti.uva.nl

## ICGEB
Sándor Pongor
International Centre for Genetic Engineering and Biotechnology
AREA Science Park, Trieste, ITALY
Email: pongor@icgeb.trieste.it
Phone: +39 040 3757300
http://www.icgeb.trieste.it

## IHCP
William Moens
Institute of Health and Consumer Protection
Via E. Fermi 1 - 21020 Ispra (Varese), Italy
Email: william.moens@jrc.it
Phone: +390332786481
http://ihcp.jrc.cec.eu.int/

## ILRI/BECA
Etienne deVilliers
International Livestock Research Institute
PO Box 30709, Nairobi 00100, Kenya
Email: e.villiers@cgiar.org
Phone: +254 20 4223000
www.becabioinfo.org

## LION Bioscience
Thure Etzold
LION Bioscience AG, Heidelberg, Germany
Email: Thure.Etzold@uk.lionbioscience.com
Phone: +44 1223 224700
http://www.lionbioscience.com

## MIPS
H. Werner Mewes
Email: mewes@mips.embnet.org
Phone: +49-89-8578 2656
Fax: +49-89-8578 2655
http://www.mips.biochem.mpg.de

## UMBER
Terri Attwood
School of Biological Sciences, The University of Manchester,
Oxford Road, Manchester M13 9PT, UK
Email: attwood@bioinf.man.ac.uk
Phone: +44 (0)61 275 5766
Fax: +44 (0) 61 275 5082
http://www.bioinf.man.ac.uk/dbbrowser

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`
Tel: +46-18-4716696

## Submission deadline for the next issue:
February 20, 2007