


EMBnet.journal

Volume 16 Nr. 1
December 2010

- 
- The background of the cover is a close-up photograph of five red apples with some yellowing, resting on a weathered blue wooden plank. The plank shows signs of age with some peeling paint and small holes. A few dried, brown leaves are scattered around the apples. The overall aesthetic is rustic and natural.
- **The contribution of the eBioKit to Bioinformatics Education in Southern Africa**
 - **Expanding BioPAX format by integrating Gene Regulation**
 - **EMBnet - the European Molecular Biology Network, Moving forward: 2010 & beyond and more...**

Editorial

Welcome to the inaugural issue of EMBnet.journal: bioinformatics in action. This first issue marks the complete transition from EMBnet.News, an online magazine produced by the EMBnet community since 1994, to EMBnet.journal an international, Open Access, peer-reviewed journal, available from December 2010 at <http://journal.embnet.org>.

EMBnet.journal will provide biomedical scientists with practical information to help tackle routine data-analysis tasks (basic and advanced), implement complex, multi-faceted IT architectural infrastructures and address the many new challenges of modern, data-driven life sciences research.

This new online journal will comprise two main parts: one part will focus on peer-reviewed primary research articles, reviews and technical notes and the other part will contain useful information and resources including commentaries, user-guides, training information and news.

We are now accepting manuscripts for review and publication and look forward to receiving your contributions and feedback. Submit your articles at: <http://journal.embnet.org/index.php/embnetjournal/information/authors>

Or contact us at Erik.Bongcam@hgen.slu.se

EMBnet.journal Editorial Board

Editorial Board:

Erik Bongcam-Rudloff, The Linnaeus Centre for Bioinformatics, SLU/JU, SE, erik.bongcam@bmc.uu.se

Teresa K. Attwood, Faculty of Life Sciences and School of Computer Sciences, University of Manchester, UK, teresa.k.attwood@manchester.ac.uk

Domenica D'Elia, Institute for Biomedical Technologies, CNR, Bari, IT, domenica.delia@ba.itb.cnr.it

Andreas Gisel, Institute for Biomedical Technologies, CNR, Bari, IT, andreas.gisel@ba.itb.cnr.it

Laurent Falquet, Swiss Institute of Bioinformatics, Génopode, Lausanne, Switzerland, Laurent.Falquet@isb-sib.ch

Pedro Fernandes, Instituto Gulbenkian, PT, pfern@igc.gulbenkian.pt

Lubos Klucar, Institute of Molecular Biology, SAS Bratislava, SK, klucar@EMBnet.sk

Martin Norling, Swedish University of Agriculture, SLU, Uppsala, SE, martin.norling@hgen.slu.se

Contents

Editorial	2
News	
Internships initiative of the ISCB-Student Council	3
Reports	
Training Mexican scientists and students on MRS/EMBOSS: A course report	4
EMBRACE Workshop - "Next Generation Sequencing II"	5
EMBnet - the European Molecular Biology Network, Moving forward: 2010 & beyond	8
Report of the EMBnet AGM 2010 Workshop	15
2010 Annual General Meeting – Executive Board Report	17
2010 Annual General Meeting – E&T-PC activities report	18
2010 Annual General Meeting – Publicity & Public Relations Project Committee Report	18
2010 Annual General Meeting – Technical Management Project Committee Report	20
The Finnish EMBnet node: AGM 2010 report	21
The ICGEB EMBnet node: AGM 2010 report	22
The French EMBnet node: AGM 2010 report	23
The Norwegian EMBnet node: AGM 2010 report	24
The South African EMBnet Node: AGM 2010 report ..	25
The Swedish EMBnet Node: AGM 2010 report	27
The contribution of the eBioKit to Bioinformatics Education in Southern Africa	29
Technical Notes	
Efficient functional bioinformatics tools: towards understanding biological processes	31
Research Papers	
Expanding BioPAX format by integrating Gene Regulation	39
Protein Spotlight	44
Node information	46



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

Cover picture: "Apples", Uppsala, Sweden, 2010. [© Erik Bongcam-Rudloff]

Internships initiative of the ISCB-Student Council



Noura Chelbat¹, Avinash Kumar Shanmugam², Venkata P. Satagopam³

¹Institute of Bioinformatics, JKU, Linz Austria,

²Center for Computational Medicine and Biology, UM, Ann Arbor, Michigan,

³Structural and Computational Biology, EMBL, Heidelberg, Germany



Participants in the poster session at the Student Council Symposium, Toronto, 2008

The [ISCB Student Council](http://www.iscb-sc.org/)¹ (ISCB-SC) represents students and young professionals involved in bioinformatics and related fields all over the world as the student organization of the [International Society for Computational Biology](http://www.iscb.org/)².

The idea is to promote young researchers in the field through different initiatives. One such initiative of the ISCB-SC is the developing nation's internships program initiated one year ago. The main aim is to provide a fellowship for students to spend from 3 up to 6 months. Through this initiative we try to find funded internship positions in labs for students from developing countries so as to give them a chance to work with experienced researchers and gain valuable knowledge and

¹ <http://www.iscb-sc.org/>

² <http://www.iscb.org/>



skills. They can then spread these among students and researchers in their home country.

How does it work?

- If a PI has an open position in his/her lab that they would like to offer to a student from a developing country, they can inform us about the position along with their criteria for students to fill that position (skills required, previous experience, stipend amount, duration etc).
- The Student Council will then call for applications from students and screen through these applications as per the given criteria and shortlist the most promising applications.
- These applications will then be forwarded to the PI who can select a student from these to fill the position.

Please note that the screening and short listing will be based on the criteria set by the PI. The PI is free to be involved as much or as little in the process as they wish to be.

Since the beginning of this initiative, about one year ago, we were offered various positions; one of them at the lab of Dr. Reinhard Schneider at EMBL, Heidelberg, which was successfully filled by a student from Estonia. We are currently accepting applications for one more position at the lab of Dr. Burkhard Rost at TU Munich. (Details can be found at <http://iscb-sc.org/content/career-central/>).

We are hoping that more PIs will be willing to offer positions at their labs so that many more students from developing countries can benefit from this!

If you are interested, please get in touch with us at internships@iscb-sc.org. We hope to hear from you!

Regards,
The ISCB-SC internship team

"Far and away, the best prize life has to offer is the chance to work hard at work worth doing"

- Theodore Roosevelt

Training Mexican scientists and students on MRS/EMBOSS: A course report



George Vasilios Magklaras¹, Romualdo Zayas²

¹The Biotechnology Centre of Oslo, The University of Oslo, Oslo, ²UNAM's Bioinformatics National Node, Center for Genomic Sciences, Cuernavaca, Mexico

Between March 22nd and March 26th 2010, the National Bioinformatics Node of Mexico (celebrating its 10th anniversary) at the Center for Genomics Sciences/UNAM, in association with the Norwegian EMBnet node at the University of Oslo/Norway, gave a course entitled "Sequence Mining in Sequence Databases: A case with MRS and EMBOSS". The Norwegian EMBnet node provided the instruction material and access to their services, and the Mexican EMBnet node the workstation laboratory and access to course hand-out material. This was one of the first MRS [1] courses given outside European borders, focusing on teaching the course participants a range of basic concepts about various bioinformatics databases, as well as how to access them in a programmatic way, in order to construct useful sequence processing pipelines for their research.

The course ran as a full day (9:30 – 15:00 hours) activity and it was very much a "hands-on" laboratory. Each day consisted of 1.5/2 hour morning lecture followed by a small break and a carefully prepared practical tutorial session, where students could practice solving real-world cases in the lab and ask questions about the tools. The first day was devoted to the basics of sequence databases and formats, as well as pipeline construction paradigms (comparing the command line approach with other approaches, such as Taverna [2], Accelrys Pipeline Pilot[3]). The next topic was a gentle introduction to key EMBOSS [4] applications and attempts to access specific



Figure 1. Group picture of course attendants

sequence databases. The remaining days were dedicated to introducing MRS as a sequence retrieval tool and the task of combining EMBOSS and MRS applications to construct complex pipelines. During the final day, the students had access to the laboratory, in an attempt to construct pipelines useful for their research problems under the guidance of the course instructor.

In summary, the course achieved its objectives and gave students a useful experience in using both EMBOSS and MRS in combination with scripting languages to construct their own pipelines. This represents a nice example of the EMBnet community at work.

In addition, aside from course instruction, the EMBnet Technical Management Project Committee is in the process of helping the Mexican node to install most of the tools (EMBOSS/MRS 4) in their computational infrastructure, so by the time you read these lines, Mexico will have a working MRS 4 installation serving their community.



Figure 2. "Hands on" laboratory in the undergraduate program classroom of the Center for Genomics Sciences

Acknowledgments

The node managers would like to thank the following people for their support in the course:

- Harald Dahle, for setting up the student accounts and the environment on the EMBnet Norway side.
- Alfredo Hernández Alvarez and the technicians of the LCG UNAM IT unit for recording the course and making the laboratory equipment run smoothly for the entire duration of the course.

References

1. Hekkelman M L and Vriend G. (2005) MRS: a fast and compact retrieval system for biological data. *Nucleic Acids Research* 33 (Web Server issue):W766-W769
2. Hull D, et al. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Research* 34 (Web Server issue):W729-W732
3. Accelrys, Inc., 10188 Telesis Court, Suite 100, San Diego, CA. URL: <http://accelrys.com/products/pipeline-pilot/>
4. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* 16: 276-277

EMBRACE Workshop - "Next Generation Sequencing II"

Rome, Italy, November 2009



Andreas Gisel¹, Erik Bongcam-Rudloff²

¹Institute for Biomedical Technologies, ²Hgen, SLU/UU, SE

After the success of the EMBRACE [1] workshop on "Building Next Generation Sequencing platforms and pipeline solutions", in Rome in November 2009, with more than 60 international participants [2], the organization committee (Domenica D'Elia, Erik Bongcam-Rudloff, Andreas Gisel) decided to organize a second event to keep pace with the fast development of Next Generation Sequencing (NGS) technologies, and the tools and methods necessary to analyse and interpret NGS data.

With financial support from the NoE EMBRACE, the organization committee decided to organize the second NGS workshop in conjunction with the EMBnet Annual General Meeting (AGM) in Ruvo di Puglia in June 2010 [3].

While the workshop in Rome focused on the basic problems arising from the use of NGS technologies, such as data management and storage or mapping and assembly of NGS data, the workshop in Ruvo di Puglia had a different focus,



Figure 1. Workshop participants during the Hackaton



Figure 2. Workshop participants

highlighting areas where NGS technology is useful, such as epigenetics, small RNA analysis and reference guided assembly strategies.

The organisation committee was again able to invite talks from high-profile scientists from all over Europe, to give this event the importance it deserves.

Peter Rice, from the European Bioinformatics Institute (EBI) in Hinxton, presented developments of the EMBOSS [4] software suite to facilitate NGS data analysis. EMBOSS will cope with the new data formats and the tremendous data volumes.

Stefan Marklund, from the Swedish University of Agricultural Sciences in Uppsala (SLU) [5], demonstrated the potential of NGS in epigenetics research, accelerating investigations on a genome wide scale and our understanding of methylation events and their consequences.

Bastien Chevreux, from DSM Nutritional Products AG in Switzerland, presented the powerful assembly suite MIRA [6] and some strategies on how to successfully assemble de novo sequenced bacterial genomes.

Alberto Policriti, from the University of Udine and Applied Genomics [7], Italy who spoke at the previous workshop in Rome, this time presented a very interesting sequence-assembly approach, combining de novo assembly and mapping onto a similar reference as guidance.

Rasko Leikonen, from the European Nucleotide Archive at EBI in Hinxton, presented developments of the Sequence Read Archive (SRA, 8), the necessity for researchers to continue to submit their NGS data as they normally do with single sequences to EMBL, NCBI or DDBJ. The SRA was presented to the NGS society by Guy Cochrane during the Rome workshop.

David Horner, from the University of Milan, Italy, introduced the participants to the world of small RNAs, and demonstrating the potential of NGS technology to understand this new world and its effect on different organisms.

Finally, Massimiliano Gentile from the IT Center for Science (CSC) in Helsinki, Finland, explained the use of NGS technologies for chromatin-immuno-precipitation combined with sequencing, and its power to search for new transcription

factor binding sites. Further, he demonstrated CHIPSTER [9], a user-friendly analysis and workflow tool, originally developed for microarray data analysis, now upgraded to handle NGS data. The ChIP-seq data analysis was the first example to show the potential CHIPSTER has.

The presentations took place during the first day of the two-day workshop, and were followed by a sight-seeing tour to the cathedral of Trani, to the famous castle of Frederic II in Castel del Monte, and finally to a well-earned dinner at the Masseria San Giovanni at Altamura. The whole social event was generously sponsored by the Apulia Region [10].

The second day of the workshop was dedicated to a sort of hack-a-thon, where participants were able to put their hands on algorithms and real NGS data, in order both to understand the various problems in NGS data analysis and to stimulate discussions and collaborations in various fields of NGS technology.

Rasko Leinonen and Vadim Zalunin prepared a hands-on session on how to submit and retrieve NGS data from the SRA, Bastien Chevreux guided the participants through the functionalities of MIRA3, and Alberto Policriti and Francesco Vezzi prepared a session with the read assembler, Velvet [11], the read mapper SOAPdenovo [12], and the Enhanced Reference Guided Assembly (e-RGA) pipeline.

At the end of the event, Erik Bongcam-Rudloff presented to the participants an EU COST-Action [13] proposal that he was coordinating, to build an NGS data-analysis network. The proposal passed the first stage and Erik invited the participants to join preparations for the full proposal.

The event hosted 68 participants from 30 countries all over the world, demonstrating that the topic of NGS data analysis is very important in the bioinformatics and bioscience communities.

Acknowledgements

We would like to thank the institutions who made this workshop possible through their support: in particular, the European Commission through its funding of the EMBRACE Network of Excellence (the EMBRACE project is funded by the European Commission within its FP6 Programme, under the thematic area "Life sciences, genomics and biotechnology for health", contract number LHS-CT-2004-512092); the Regional Council of Apulia



Figure 3. Workshop participants during the Hack-a-thon

(IT) (initiative financed by EU funds POR Puglia 2000/2006); the CNR Institute for Biomedical Technologies of Bari (IT); the National Institute of Nuclear Physics of Bari (IT); and EMBnet (the European Molecular Biology Network).

The organizers would also like to express their deep gratitude to all the invited speakers and all the participants who, through their enthusiasm and interest, made this such a successful workshop.

References

1. <http://www.embracegrid.info>
2. www.nextgenerationsequencing.org
3. <http://www.embnet.org/NGS-AGM2010>
4. www.emboss.org
5. www.slu.se
6. <http://sourceforge.net/apps/mediawiki/mira-assembler>
7. <https://www.appliedgenomics.org/>
8. <http://www.ebi.ac.uk/ena/>
9. <http://chipster.sourceforge.net/>
10. <http://www.viaggiareinpuglia.it/aptbari>
11. <http://www.ebi.ac.uk/~zerbino/velvet/>
12. <http://soap.genomics.org.cn/soapdenovo.html>
13. <http://www.cost.esf.org/>

EMBnet - the European Molecular Biology Network, Moving forward: 2010 & beyond



Teresa K. Attwood¹, Erik Bongcam-Rudloff²



Andreas Gisel³, Etienne de Villiers⁴

¹ Faculty of Life Sciences and School of Computer Sciences, University of Manchester, UK, ² The Linnaeus Centre for Bioinformatics, SLU/UU, SE, ³ Institute for Biomedical Technologies, CNR, Bari, IT, ⁴ EMBnet BecA-ILRI, Nairobi, KE

Executive Summary

The world that gave rise to EMBnet 22 years ago has changed; EMBnet has changed too. Initially focusing on shared *European* community needs, the organisation now embraces partners from around the world, whose aims, aspirations and communities are very different. Over time, its centre of gravity has shifted: no longer needed as a mechanism for distributing the EMBL databases, EMBnet has moved on. In the past, its activities have been funded by membership fees and a series of European grants; now, however, against a background of emerging (government-funded) European infrastructural initiatives, EMBnet's funding mechanisms need to be reviewed, in order to continue to sustain its core activities.

EMBnet is at a cross-roads. Before taking its next steps, it is appropriate to consider how the global bioinformatics landscape is evolving, and how EMBnet needs to adapt. This paper outlines a number of practical steps that could be taken,

tempered by today's funding climate. Its principal recommendations are that EMBnet should:

- i. review and properly define the roles, aims and goals of its Executive Board (EB) and Project Committees (PCs), and consider establishing additional PCs or Special Interest Groups (SIGs), with well-defined roles, aims and goals;
- ii. review how EMBnet and its collection of PCs/SIGs might achieve its goals, with or without further funding;
- iii. identify and exploit its Unique Selling Point (USP);
- iv. review and better understand who its communities are, what their needs are, and how to be more responsive to those needs;
- v. review, streamline and clarify its current membership scheme;
- vi. review how and why it might interact with other networks and organisations;
- vii. establish internal infrastructures that would allow it to make strategic ties to other bioinformatics networks and organisations;
- viii. establish internal infrastructures that would allow it to make more strategic responses to global funding opportunities;
- ix. review the evolving role and internal structure of EMBnet.journal, and consider more tactical publishing strategies; and, in light of these considerations,
- x. review and revamp its current name, brand and Website.

This paper is an open invitation for every member of the constituency to help with this critical evaluation of EMBnet's unique attributes and strengths; to consider how to build on these to create a competent, valuable and focused organisation that complements existing and emerging bioinformatics institutes, networks, associations and societies worldwide; ultimately, to maintain EMBnet's relevance in 2010 and beyond.

Background

Established in 1988, EMBnet served as an organisation for disseminating data, knowledge and services to support and advance research in molecular biology and biotechnology across a broad European community. Its service provision and knowledge sharing was primarily orchestrated by 'National Nodes' with government

mandates to support their local communities; in time, the organisation also attracted a number of 'Specialist' and 'Industrial' Nodes, whose resources and know-how were seen to complement those of its National Nodes.

One of the major drivers for establishing EMBnet was the need for local access to data from centralised sources. In particular, it was intended to function as a distribution network for the EMBL Data Library databases (delivered, in those days, by floppy-disc and/or CD-ROM); consequently, from its inception, EMBnet had a special relationship with the EMBL. However, the growth of the EMBL nucleotide sequence database ultimately rendered this form of data distribution untenable; moreover, as local bioinformatics resources became more commonplace, with increasing compute power and swifter networks, reliance on the central facilities at EMBL was reduced or obviated. Eventually, the EMBL databases moved to the EBI outstation, when that was created in 1995.

Now, more than 20 years on, EMBnet has changed substantially. Not least, for more than a decade, the organisation has been embracing increasing numbers of countries from continents around the world – its constituency is hence no longer European, as illustrated in Fig. 1. Moreover, new research methodologies like 'Next Generation Sequencing' (NGS) are producing data on an unprecedented scale, and are once again driving the need for local services (processing, storage, management, analysis, software and database development, help and advice, etc.). Today, then, EMBnet's remit reaches far beyond the realms of support for molecular biology research.

EMBnet is in a state of transition. From the original handful of European Nodes, it has grown into an expansive international network of bioinformatics and biocomputing centres, the missions of which vary widely: from service centres specialising in scientific computing and communications, to those focusing on the molecular biology and biotechnology of plant, animal and insect viruses, and those dedicated to bioinformatics outreach and training. The nature and size of the local target communities vary enormously, and the range of service provision varies accordingly.

Over the years, one of EMBnet's most prominent developments has been its newsletter,



Figure 1: Snapshot of EMBnet Nodes in 2009, showing representation across the globe (yellow pins are National Nodes, pink are Specialist Nodes). The constituency is no longer European.

EMBnet.news. This has become a significant publication and, for many Nodes, provides an ideal way of reaching their local research communities with mission-relevant information: with its emphasis on presentation of everyday bioinformatics problems and their practical solutions, it fills a gap between bioinformatics theory and data management. Now, as part of EMBnet's expansion, the newsletter has begun to transition from a largely technical publication to a peer-reviewed journal.

While EMBnet has been expanding across the globe, reaching out to wider and more diverse communities with each new Node joining the Network, major initiatives have begun to emerge to try to address the growing need for a European data infrastructure. Against this background, it is therefore timely to review where EMBnet came from and why, what its role is now in both European and global contexts, and how the organisation should move forward, for the benefit of its global membership.

Perspectives on EMBnet

In order to continue running its business beyond its first 5 years, EMBnet was established as a foundation (Stichting), registered in Nijmegen, The Netherlands, in 1993. To operate efficiently, EMBnet initially established 5 Project Committees (PCs), each with a minimum of 3 members: the PCs were i) Technical Developments; ii) Connectivity & Software; iii) Information Services; iv) R&D; and v) Workshops & Training. These were overseen by a Steering Committee, also comprising a minimum of 3 members: a Chair (the spokesman and legal representative in official functions); a Treasurer (concerned with collec-

tion and disbursement of funds); and a Secretary (concerned with keeping official records). Over time, there have been various changes to this structure, such that, today, EMBnet discharges its duties via 3 PCs (Education & Training; Technical Manager; Publicity & PR) and an Executive Board (EB).

In thinking about what EMBnet was and what EMBnet is *today*, it is helpful to consider how it has evolved, what its roles and goals should or could be moving forward, the steps needed to achieve them, and whether and/or how these might be funded. In the sections that follow, we reflect on these questions, both at an over-arching organisational level, and at the level of its internal structures – *i.e.*, its PCs and EB.

EMBnet, the organisation

Current role, and its role and goals moving forward

The current role of EMBnet largely involves the following activities:

- i. sharing data, knowledge and technological knowhow amongst its partners;
- ii. acting as a portal for bioinformatics-related information;
- iii. engaging its members in bioinformatics training activities;
- iv. promoting the activities of its members via its online publication;
- v. attracting new members to the Network.

Moving forward, EMBnet could build on these activities, to become more global and outward looking in terms both of sharing bioinformatics-related information and experience, and of stimulating and/or supporting training activities and capacity building.

Strategically, in the short term, EMBnet should aim to make a commitment both to seek collaboration/cooperation with other bioinformatics networks and organisations, and to seek funding for its future activities via opportunities worldwide.

In the longer term, it should aim i) to be a global hub of bioinformatics networks; ii) to host a range of funded research/training/capacity-building activities; and iii) to nurture its successful online publication as it evolves into EMBnet.journal.

Steps needed to achieve its goals

A) Steps towards short-term goals:

A number of immediate and fairly pragmatic steps could be taken to help EMBnet achieve its short-term goals. These include:

- i. reviewing and properly defining the roles, aims and goals of its EB and PCs;
- ii. reviewing its current structure, and considering whether additional formal PCs or perhaps more 'fluid' Special Interest Groups (SIGs) might be useful (e.g., for NGS discussions); and, if additional or different PCs or SIGs are considered beneficial, firmly defining their roles, aims and goals;
- iii. reviewing what EMBnet as a whole, and as a collection of PCs/SIGs, can realistically achieve, with or without further funding;
- iv. identifying and exploiting its USP!
- v. reviewing and better understanding who its communities are (novice users, bench researchers, tool/resource developers, educators, etc.) and what their needs are, and being responsive to those needs;
- vi. reviewing and streamlining its current membership scheme – e.g., clarifying the rules and goals of personal membership such that the scheme is a) widely understood, b) simple to implement, and hence, ultimately, c) widely used;
- vii. reviewing how and why it might interact with other networks and organisations;
- viii. establishing internal infrastructures that would allow it to make strategic ties to other bioinformatics networks and organisations (e.g., through a dedicated outreach PC and its nominated leader);
- ix. establishing internal infrastructures that would allow it to make more strategic responses to global funding opportunities (e.g., through its SIGs and the designated leaders of those SIGs); and
- x. reviewing the evolving role and internal structure of EMBnet.journal, and considering more tactical publishing strategies (e.g., via themed issues, special conference or workshop proceedings; and so on).

B) Steps towards longer-term goals:

In order to meet its longer-term goals, once again, a number of relatively straightforward and

pragmatic steps could be taken. These include, but are not limited to:

- i. ensuring that the Website properly reflects EMBnet moving forward – e.g., with a meaningful name and ‘brand’, with an easy-to-use interface, with informative content, by embracing social networking technologies, and so on;
- ii. applying for funding from a variety of different sources;
- iii. joining global funding consortia.

As EMBnet moves forward, its internal and external communication mechanisms must work efficiently and effectively, and it should strive to work/cooperate closely with other organisations (ISCB, APBioNet, Bioinformatics.Org, the African Bioinformatics Network and so on.) and, above all, not in competition with them. It should also consider its relationship with international bioinformatics conferences, such as ISMB, ECCB, *etc.*

Importantly, we need to understand what’s different about EMBnet – *i.e.*, what is its USP (or USPs)? Having identified its USP(s), we need to build on it (them)! For example, EMBnet has a growing number of Nodes from developing countries – we need to understand their needs, and to focus efforts there; we need to ask what other networks or organisations currently do, or plan to do, for these countries (and what they will not do); and we also need to understand the role of EMBnet.journal in this context.

Funding requirements and funding mechanisms

There is little here that needs direct funding (aside from a new Website), as the ideas primarily concern the need for internal structural changes within EMBnet. Nevertheless, realising EMBnet’s ambitions will require identifying concrete courses of action, identifying strategic projects and key individuals (champions) who are prepared to help deliver them; moreover, it will require identifying the funding opportunities that exist to make these courses of action/projects possible, and will require concerted efforts to apply for those funds.

In terms of potential sources of funding, these are wide and varied. They might include, but are not limited to: the EU; other national and international funding bodies (e.g., for capacity building, and so on); Bill Gates Foundation; *etc.*

The Executive Board

The following sections outline the EB’s view of its role, aims and goals, set in the context of a vision and mission for EMBnet as a whole, moving forward from 2010.

Everything outlined here hinges on the people within EMBnet, their willingness to support a collective vision and their enthusiasm to act upon it. Our specific, and most urgent, recommendation is the development of a new Website.

Current role, and its role and goals moving forward

The role of the EB is to carry out the decisions of the Board and to run the daily business of EMBnet. It currently works towards achieving this, in close collaboration both with its PCs and with the full Board, in two main ways: i) formally, through monthly Virtual General Meetings (VGMs) and Annual General Meetings (AGMs), and ii) informally, via its email lists. In particular, the EB is responsible for managing the Stichting financial accounts, and for preparing an annual financial report for discussion at its AGMs.

As EMBnet moves forward, the EB should aim to provide vision, leadership and efficient executive coordination of the activities of the PCs and of the growing number of Nodes (and possibly of SIGs, should it establish them).

In the short term, the EB aims to inculcate a spirit of professionalism, dignity and collegiality amongst members of the Board in general, and of the PCs in particular, so that EMBnet can conduct its future business effectively and efficiently.

In the longer-term, the EB should aim to establish internal mechanisms by means of which i) it might secure strategic alliances with other organisations, ii) its members might compete more strategically for funding to support some of its activities, and iii) it might promote its activities more effectively.

Steps needed to achieve its goals

A) Steps towards short-term goals:

A number of immediate and practical steps could be taken to help the EB achieve its short-term goals. These include:

- i. properly defining the roles of EB members (of the Chair, Secretary, Treasurer, *etc.*), and helping to create closer ties between Board members and PCs/SIGs;
- ii. helping to promote the establishment of strategic alliances with other networks;

- iii. rationalising how VGMs and AGMs are conducted, ensuring that the groundwork for these meetings is prepared in advance;
- iv. formalising nomination procedures, so that voting at AGMs (and/or VGMs) is properly informed;
- v. helping to improve the EMBnet brand.

B) Steps towards longer-term goals:

In order to meet its longer-term goals, again, various simple practical steps could be taken. These include:

- i. discussing within the EB, with the PCs and the wider EMBnet constituency, the establishment of SIGs, whose role could include, amongst other things, the coordination of members with mutual interest in particular funding calls;
- ii. encouraging and supporting the improvement/re-design of the Website;
- iii. encouraging better use of the Website or, rather, helping to make the Website work better for EMBnet.

Funding requirements and funding mechanisms

Most of these ideas do not have direct funding implications *per se*, as they primarily concern the need for internal structural changes. However, re-development of the Website may have an associated cost, which could probably be met by the Stichting.

The Education & Training PC

Current role, and its role and goals moving forward

The E&T PC is in a state of flux, but is trying to focus its efforts on producing new, and updating old, Quick Guides (QGs). Importantly, it can cover the whole QG-publication workflow, relieving the P&PR PC of this process.

The E&T PC should strive to be a strong, well-structured course-giving PC. It should be able to organise tutorials and workshops, to provide training to local communities through its network of Nodes. The courses could range from basic to advanced topics, and could be delivered in a systematic way: e.g., it could start with 2 courses per semester in different locations, eventually becoming a fund-raising activity for EMBnet. With financial support from local and international donors, the E&T PC could provide both the nec-

essary infrastructure and the course materials. As part of its goal to provide such training to communities throughout the world, the E&T PC could champion the production of 'EMBnet Kits'. For example:

- i. **EMBnet Kits for students.** In order to disseminate bioinformatics history, data, methods, resources, and so on, a student kit, including a professionally produced set of carefully selected materials (texts, videos, training datasets, free programs, *etc.*), could be distributed via DVD, USB stick, *etc.* The kit could be offered (or sold) to Universities and Institutes for distribution to students – this could be done, for example, in collaboration with initiatives like SLING;
- ii. **EMBnet Kits for medical professionals.** A similar resource could be produced for medical practitioners, many of whom are unaware of the benefits of accessing biological data and core bioinformatics methodologies. A kit with illustrated examples could help to motivate medical professionals to look at problems with different eyes, and could have the added benefit of bringing new projects to bioinformaticians, and hence of initiating new collaborations. 'Bio-data aware' medical researchers and practitioners are sorely needed.

Steps needed to achieve its goals

In order to achieve its goals, a number of steps are necessary. The E&T PC should:

- i. better define its role and the roles of its members;
- ii. define the scope and parameters of its courses, especially if it is to embark on travelling workshops/road-shows to different international locations;
- iii. cooperate closely with people outside the PC, to ensure the delivery of relevant, tried-and-tested courses and tutorials;
- iv. keep all of its materials up to date, if it is to deliver useful courses; and
- v. cooperate with the P&PR PC to produce and promote EMBnet Kits.

Funding requirements and funding mechanisms

Organising courses requires funds (e.g., to cover tutors' expenses). National and international funding donors need to be identified (e.g., bio-

informatics companies, Marie Curie grants). The possibility of co-organising courses with other organisations (e.g., FEBS, EMBO) could also be considered.

The TM PC

Current role, and its role and goals moving forward

The role of the TM PC is to:

- i. provide the technical foundation for EMBnet's information flow: this encompasses DNS-services, mail-list services, a Marratech video conferencing system and, most importantly, maintenance of a Web server and its subsystems (Drupal, OJS, etc.);
- ii. provide guidance to Node managers and scientists around the world about setting up IT infrastructure resources to facilitate bio-computing applications (including help with technical specifications, service/software set-up issues, and so on);
- iii. maintain expertise in the development and maintenance of key applications in the field of biological sequence analysis and retrieval systems, including open source applications such as MRS; wEMBOSS; MIRA 3 Whole Genome Shotgun and EST Sequence Assembler for Sanger, 454 and Solexa/Illumina; Galaxy workflows; Taverna; tools that facilitate formatting/indexing of flat-file databases;
- iv. provide technical advice and investigations for core life science computing problems (e.g., for NGS data storage and processing resources).

In addition to maintaining and developing EMBnet's existing information management infrastructure, the TM PC would like to support the following emerging projects:

- i. technical challenges involving the storage and pipeline construction of NGS data;
- ii. GPU computing paradigms in the life sciences.

In the short-term, we would like to deploy improved backup and mirror systems for the www.embnet.org site. We recommend that both the existing and mirror sites should move to a Linux Open Source platform, as MacOS X presents problems when compiling various open source

packages. In the longer term, the TM PC would seek to become a point of reference for technical bio-computing issues.

Steps needed to achieve its goals

To achieve our short-term goals, we seek a modest annual budget to be allocated to various projects and consumables (data drives, memory, software) needed to sustain the basic services. A Mac Mini server is set in Pakistan. Help/guidance in mirroring the EMBnet site will be needed.

Steps required to achieve our long-term goals are more challenging:

- i. the capacity of the present server is unlikely to be sufficient for much longer, and should therefore be upgraded within the next year;
- ii. the issue of Marratech vs other video conferencing solutions might be solved by using freely available university resources;
- iii. the TM PC could become involved with exploratory projects, looking at emerging IT- or scientific developments/trends in partnership with particular Nodes. This would be mutually beneficial to the Node(s) and to EMBnet;
- iv. sharing our expertise via a blog could be useful, and could help to stimulate wider participation in EMBnet. A FAQ-list could be assembled, which, if structured and equipped with a navigation tool, could be useful.

Funding requirements and funding mechanisms

Upgrading the systems, subscribing to a backup service, and projects involving long-term support will need external funding. Projects of general interest could be presented to emerging 'big-science' bio-projects, such as the Science for Life Laboratories in Stockholm and Uppsala, and corresponding initiatives in other countries.

The P&PR PC

Current role, and its role and goals moving forward

The main roles of the P&PR PC are to:

- i. nurture and promote EMBnet's image at large, enhancing its visibility;
- ii. promote EMBnet and its activities by:
 - maintaining and enhancing the Website;
 - supporting the publication and dissemination of EMBnet.journal;
 - liaising with other groups and societies;
 - fostering appropriate, active connections with ISCB, APBioNet, etc.

Future roles could include:

- i. establishing effective links with trans-national organisations that have vested interests in the area, such as the Ludwig Institute for Cancer Research, Sloan Kettering, GBIF (Global Biodiversity Information Facility), TDWG, for example;
- ii. helping to devise new mechanisms for attracting sponsors for EMBnet.journal, and for courses, workshops, conferences, *etc.*

The goals of the P&PR PC include bootstrapping a number of short- and long-term projects and initiatives, including, for example:

- i. Re-designing the Website. The new site needs to be structured to include more informative content and to be more easy-to-use: *e.g.*, it should showcase: EMBnet's people, partners and institutes; its services and resources; its National, Associate and Specialist Nodes; its widely distributed activities; *etc.*;
- ii. Geo-referencing EMBnet. This could be seen either as an addition to the Website (and made available for inclusion in other Websites) or as an alternative view of EMBnet. It should provide an up-to-date overview of the organisation and allow people to readily contact their nearest Node;
- iii. Social networking. EMBnet's LinkedIn group could be used to hold discussions, to generate news feeds and to open new discussions.

All of these initiatives will require help and collaboration with:

- the EB, in decisions relating to public relations and cooperation with other related groups, networks and societies worldwide;
- the E&T PC, to support them in the promotion, organisation and dissemination of QGs, training courses and materials, summer schools, and so on;
- the TM PC, in all issues related to the improvement of EMBnet's Website.

Funding requirements and funding mechanisms

There will be costs associated with printing and disseminating publicity materials. Promotion of EMBnet at conferences will be expensive, but if included in a cooperation agreement, say with

other related networks and societies, the costs could be reduced.

Conclusion

The world that gave rise to EMBnet 22 years ago has changed. As EMBnet has acquired more partners, its operational model has moved from a focus on shared European community needs, to one that embraces disparate Nodes with different aims and aspirations, and serving very different communities with very different requirements. As its partners moved on from their reliance on EMBnet as a mechanism for distributing the EMBL databases and E-GCG/EMBOSS, the organisation moved on too.

Today, it is important for EMBnet to try to understand how the global bioinformatics landscape is evolving, and how it should adapt. This paper has outlined some practical steps that could be taken, all of which need to be tempered by a proper understanding of the current funding climate. Ultimately, the paper is a call to arms: to embrace every member of the constituency in a critical evaluation of EMBnet's unique attributes and strengths; to consider how to build on these to create a focused organisation that complements existing and emerging bioinformatics institutes, networks, associations and societies worldwide; ultimately, to maintain EMBnet's relevance in 2010 and beyond.

Report of the EMBnet AGM 2010 Workshop

Ruvo di Puglia, June 18th, 2010



Teresa K. Attwood¹, Erik Bongcam-Rudloff²



Andreas Gisel³, Etienne de Villiers⁴

¹ Faculty of Life Sciences and School of Computer Sciences, University of Manchester, UK, ² The Linnaeus Centre for Bioinformatics, SLU/UU, SE, ³ Institute for Biomedical Technologies, CNR, Bari, IT, ⁴ EMBnet BecA-ILRI, Nairobi, KE

Executive Summary

More than 20 years after its inception as an organisation for disseminating data, knowledge and services to support molecular biology/biotechnology research across Europe, EMBnet has arrived at a cross-roads. Since 1988, its membership has expanded outside Europe, and the nature of many of the needs it was created to support has changed. Against this background, major (government-supported) initiatives have begun to emerge to try to satisfy the modern European data infrastructure need. The 'EMBnet moving forward' workshop was an opportunity to review the current context and how EMBnet should embrace the future, to the benefit of its global membership. The following pages summarise the workshop discussions and conclusions.

EMBnet structure

This session reviewed EMBnet's structure and whether it needs more PCs or Special Interest

Groups (SIGs). The roles and aims of its EB & PCs, its mission and USP were also explored.

Conclusions:

- no further PCs are required; the EB & PCs will promote new activities via SIGs in future;
- roles of the EB/PCs are well defined, but new roles were proposed for the TMPC;
- SIGs with clear goals should be established and led by key individuals;
- EMBnet's mission is to share knowledge, expertise, training, etc. and should include efforts to break down barriers by seeking funding for cross-border activities;
- EMBnet's history, geographic dispersion & multi-lingual nature are unique.

EMBnet communities

Here, we reviewed EMBnet's communities, who they are, their needs and how EMBnet should respond. We considered why people join and what they get out of it.

Conclusions:

- EMBnet's community is global & diverse;
- community needs differ, but there are common themes: the community needs EMBnet's support to solve its local problems (IT infrastructure, tool development, training, etc.);
- EMBnet should embark on a program of activities to better serve the needs of its communities; a coordination tool to facilitate collaboration would be helpful;
- the reasons for joining EMBnet have evolved over time, but again, there are some common themes – e.g., access to support, to expertise, to standard curricula.

EMBnet strategic alliances

This session examined organisations with whom EMBnet could interact, why it should interact and the form such alliances might take. It also considered if a dedicated outreach PC was needed.

Conclusions:

- EMBnet should aim to form strategic alliances with a range of bioinformatics organisations/networks and developer projects, and should take advantage of global funding initiatives;
- interactions with other organisations are likely to be mutually beneficial and synergistic;
- alliances with other organisations could take different forms, depending on the aims and



Figure 1. EMBnet AGM 2010 – Hotel Pineta, Ruvo di Puglia, Bari (IT)

objectives; joint activities could be funded through mutual cooperation;

- a new outreach PC isn't needed, but a dedicated outreach member on the EB, collaborating with the P&PR PC, could fulfill this purpose; Erik Bongcam-Rudloff is a good candidate.

EMBnet.Journal

Here, we discussed the transition to a peer-reviewed journal. We reviewed the obstacles to article submissions, strategies to overcome these, including the introduction of specific themes.

Conclusions:

- the evolution to a professional, peer-reviewed journal is a good idea;
- EMBnet.news' lack of IF and bibliographic database indexing has been a barrier;
- members would submit articles to EMBnet.journal if/when it has an IF and is registered with a bibliographic database;
- strategies for encouraging article submissions were suggested; fees need careful thought;
- specific themes were suggested, but these need to be introduced strategically.

EMBnet brand

This session examined the EMBnet brand & what changes might be appropriate. We discussed how to improve the website & the cost of changing it.

Conclusions:

- the name EMBnet should be retained, but it should carry a new strapline;
- the logo should remain the same;
- numerous improvements need to be made to the website; the P&PR and PC & TM PC need

to collaborate quickly and effectively to make this happen;

- much of the work to redesign the website can be done by internally, but some aspects might need to be outsourced – these would need to be properly costed.

EMBnet membership

Here, we considered the goals and terms of personal membership, who might become personal members and whether other types of membership were needed.

Conclusions:

- the goals of personal membership are many and varied, but broadening the EMBnet constituency and building loyalties are key;
- the rights and responsibilities of personal membership should be clarified;
- the rules of personal membership need to be clarified and simplified;
- EMBnet should look into establishing Society & Honorary memberships.

EMBnet e-learning resources

This session discussed EMBnet's e-learning resources, the kinds of materials that should be provided, how they should be kept up-to-date and their quality ensured.

Conclusions:

- EMBnet should provide a range of materials appropriate to different courses and audiences; the materials should be well structured, documented and time-stamped;
- authors should be responsible for maintaining their own content; time-stamps could be used to indicate when modules/kits were last updated;
- authors should be responsible for maintaining the quality of their own content; user feedback and satisfaction surveys should also be used;
- EMBnet would benefit from sharing its e-learning resources.

EMBnet and funding opportunities

The last session looked at funding opportunities that could support EMBnet's activities, specifically with respect to its training & research activities and its IT projects.

Conclusions:

- numerous funding opportunities exist, which EMBnet will pursue proactively.

2010 Annual General Meeting – Executive Board Report

May 31st, 2010

Chair: T.K.Attwood

Secretary: J.R.Valverde

Treasurer: E.deVilliers

Member: A.Gisel

Since the last AGM, the Executive Board (EB) has met on a monthly basis; it has also convened monthly virtual meetings for general discussion of issues relating to the Project Committees (PCs), to *EMBnet.News/Journal*, and so on. The meetings have also allowed us to discuss the Stichting financial accounts, to prepare the groundwork both for the annual financial report and, crucially, for the AGM. The EB meetings are closed; the general virtual meetings are open to any members of the EMBnet constituency who are sufficiently interested to participate.

This EB has been in office for a relatively short time. Its election towards the end of last year was somewhat unusual, in the sense that only its Secretary remained on the Board – most of its members are thus new. This has therefore been a period of change, and a learning process for us all. At the same time, the global bioinformatics landscape is also changing. We therefore felt that the 2010 AGM would provide an ideal opportunity to examine those changes, to consider EMBnet's position within this evolving landscape and to explore its aspirations and motivations, moving forward.

To this end, we have begun to take the first small practical steps. Internally, we are trying to better define the roles of each member of the EB, so that the daily business runs more effectively; we are trying to create closer ties between EMBnet members and the PCs (and/or SIGs, should we establish them); we are trying to rationalise how VGMs and AGMs are conducted, by trying to ensure that the groundwork for these meetings is prepared in advance; we are trying to formalise nomination procedures, so that voting at AGMs (and/or VGMs) is properly informed; and we are trying to help to improve the EMBnet 'brand'. Overall, the EB hopes to promote a series of internal structural and behavioural changes within EMBnet both to inculcate a spirit of professionalism, dignity and collegiality amongst all members of the EMBnet constituency, and to

provide efficient and effective executive coordination of EMBnet's activities. In particular, we hope to be able to establish internal mechanisms by which we might: secure strategic alliances with other organisations; compete more strategically for funding for some of our activities; and promote our activities more effectively. Ultimately, we hope to facilitate the evolution of EMBnet into a *respected network of bioinformatics professionals*.

For the forthcoming AGM, our Secretary, J.R.Valverde, will be stepping down, so there will be one vacancy on the EB. At the time of writing, we have received no formal candidacies, so we encourage members of the EMBnet constituency to consider stepping forward to join us! Meanwhile, we would like to thank JR for his dedicated services to EMBnet and especially to the EB during the last 3 years.

2010 Annual General Meeting – E&T-PC activities report

(November 2009 – June 2010)

Chair: Matej Stano

Treasurer: Jingchu Luo

Member: Shahid Chohan

Member: Sophia Kossida

The main mission of E&T PC is to promote training and education in bioinformatics. PC's activities were discussed during three virtual meetings (November 11th, 2009; January 7th, 2010 and April 7th, 2010). As a result, activities of the committee were aimed at publication of Quick Guides. The MySQL (Tables & Queries) Quick Guide was published and is available at EMBnet web portal. This Quick Guide was written by Awais Naseem and Nazim Rahman from Pakistan EMBnet node and reviewed and published by people from Slovak and Chinese node.



Figure 1. Bioinformatics training in Sweden.

2010 Annual General Meeting – Publicity & Public Relations Project Committee Report

(November 2009 – June 2010)

Chair: Pedro Fernandes

Secretary: Lubos Klucar

Member: Domenica D'Elia

Member: Martin Norling

The main mission of the P&PR PC is to nurture and promote EMBnet's visibility and to establish cooperation with other (even dissimilar) major groups, networks and societies.

Starting from November 2009 up to June 2010 the organization of the PC's activities was discussed and agreed during three committee's meetings (March 4th, May 31st and June 18th 2010) and four meetings of the EMBnet.news Editorial Board (February 22th, March 8th, May 26th and June 17th, 2010).

The main issues on which the Committee has worked are the following:

1. supporting the organization of the second edition of the EMBnet-EMBRACE joint workshop on Next Generation Sequencing, held in Ruvo di Puglia (IT) on June 2010 [1];
2. preparation of a P&PR PC's position paper as a basis for structuring themed discussions at the 2010 AGM in Ruvo di Puglia (IT);
3. transition of the EMBnet.news magazine to an Open Access Peer Reviewed journal.

Committee's initiatives proposed to the EMBnet EB and VGM have been:

1. establishment of a Special Interest Group (SIG) to study the input of EMBnet's PCs and generate clear scenarios on the future activities of EMBnet and its societary role (March 2010);
2. re-designing of the EMBnet's web site with a new easy-to-use structure and with more informative content. This work will be lead by the P&PR PC in collaboration with the EB and with the technical support of the TM PC. In addition to TM PC also the P&PR PC will have administrative access to the web portal;
3. "EMBnet Members 2009" geographical distribution published on Google Earth is being updated and a new version will be linked from the new embnet.org pages and regularly updated;
4. managing, updating and publishing Quick Guides will be handed over to the ET-PC at EMBnet AGM 2010.

From EMBnet.news to “EMBnet.journal – Bioinformatics in Action”

The EMBnet.news magazine, edited by the P&PR PC and EB Chair since its first edition (1994) has gone under a strong transformation which has taken much of PC's members efforts and delayed the release of new issues (2010 editions).

One EMBnet.news issue has been published on March 2010 (vol. 15, issue 4). This issue represents the last one of the EMBnet.news edition. Starting from the next one the EMBnet.news magazine is replaced by the “EMBnet.journal – Bioinformatics in Action”, an Open Access peer-reviewed journal published by EMBnet. EMBnet-RIBio 2009 Conference Proceedings are published as the first Supplement issue of the EMBnet.journal (Vol. 16 Suppl. A) [2].

The OJS (Open Journal System), implemented by Lubos Klucar in 2008 for the management of the publishing process, becomes the official web-based journal management system of the EMBnet.journal. In 2009 a new version (OJS 2.3.1.2) freely distributed by the Public Knowledge Project [3] has been implemented. This new version includes additional facilities that better support the OJS editorial and publishing process. Users have different access privileges depending on their own category: Editor, Section Editors, Copyeditors, Proofreaders, Layout Editor, Authors and Reviewers. The system provides users' alert and notification facilities and allows full text search and commenting for all articles published. This new implementation has been tested by the Editorial Board under the guidance of Lubos who held two virtual tutorials workshops on the organization of the OJS editorial process, the procedures to be used and roles for each one of the Editorial Board member.

According to the complexity of the publishing process, the Editorial Board of the journal has agreed on a formal detachment from the P&PR PC, being effective starting from the EMBnet AGM 2010, and on the enlargement of the board to an “Executive Editorial Board” and a “Full Editorial Board”. The Executive EdB will consist of 5-10 people, mostly current members of EdB while the Full EdB will be composed of all EMBnet node managers. Professionals outside the EMBnet will be invited to join the Full EdB depending on specific needs. Inside the Executive EdB each member has been designed for a specific role, such as Section Editor, Copyeditor and so on, and Erik

Bongcam-Rudloff has been officially designed as Editor in Chief. The board has also been working on the Journal's Copyright, Privacy statement, Focus and Scope and on the Editorial policy.

The journal contains two main sections, one that caters for peer-reviewed primary research articles, reviews and technical notes, and one for non-peer-reviewed commentary, reportage, user-guides, training information and news. In addition the journal will also publish Special Issues focusing on 'hot' topics that fall within the scope of the journal and Supplements publishing peer reviewed conference proceedings and meeting abstracts.

EMBnet.news past issues have been moved from the embnet.org site to the new system [4] where all the old issues are collected in a retrievable archive. This work has been carried out by Lubos with the support of a student hired by Nazim Rahman and is now complete, while the transformation of all back issues into OJS is still ongoing. This requires the creation of abstract, HTML and PDF versions for each past published article.

Google Analytics statistics are collected since February 2010. Author's fees for peer-reviewed articles will be charged starting from 2011, and a policy for the article processing charge has also been agreed. The P&PR PC will collaborate with the Executive EdB in promoting EMBnet.journal as a new peer reviewed journal across bioinformatics related web sites and at main Bioinformatics and related conferences. The committee has also been working for retrieving information on how to get indexing in PubMed. DOI is under investigation. The journal is currently indexed in Google Scholar.

References

1. <http://www.embnet.org/NGS-AGM2010>
2. <http://journal.embnet.org/index.php/embnet-journal/>
3. <http://pkp.sfu.ca/>
4. <http://journal.embnet.org/index.php/embnet-news/>

2010 Annual General Meeting – Technical Management Project Committee Report

(November 2009 – June 2010)

Chair and Secretary: George Magklaras

Member: César Bonavides-Martínez

Member: Harald Dahle

Member: Nazim Rahman

Member: Emil Lundberg

Member: Nils-Einar Eriksson.

The TMPC is responsible for developing and maintaining a number of vital IT infrastructure components, namely: i)The embnet.org web server, ii) The embnet.org mailing lists, iii)The embnet.org Domain Name Servers (DNS). During the last year, the TMPC has performed the following activities: César is doing a monthly backup of the Drupal Database and a “sync” of the Drupal installation and files used within Drupal (excluding e-learning files). This backup is performed by one of his servers in Mexico. DVD-backups are also done locally at the webserver. Jerker Nyberg at BMC's computing department has written a script that makes backups to the 1TB external disk that he installed. Steps to establish a mirror site of the MACOSX server in Pakistan have been taken by Nazim. However, recent assessments of MacOS X as a platform with problems compiling properly various open source packages, necessitates a thorough discussion about the best way to establish the planned mirror site.

The eLearning site has been upgraded by (TMPC plus Jose R. Valverde) to the latest version, which now supports CAPTCHAs to avoid flooding by spammers. Measures have been taken to decrease the risk of security breaches (strengthened and revised by George Magklaras and BMC IT). New modules to support Jmol and automatic certificate generation have been added as well.

Members of the TMPC are involved in planning and establishing bioinformatics services locally for the management and analysis of the vast quantities of data coming from the new generation of sequencing instruments (‘Next Gen’ data).

Members of the TMPC were represented on the organizing committee of a workshop in Rome November 18-20, on next generation sequencing related bioinformatics (www.nextgenera-

[tionsequencing.org](http://www.nextgenera-tionsequencing.org)). The secretary of the TMPC (George Magklaras) delivered a talk, outlining the IT/computational challenges of next generation sequencing projects.

In March 2010, the TMPC (George Magklaras) was involved in activities to help the Mexican EMBnet node with the process of formatting EMBOSS datasets. Consultation is also under way to upgrade their server infrastructure in order to be able to install MRS 4. As a result of this activity, the TMPC will soon release a set of scripts as guidance to format EMBOSS datasets. Details of this project will be released before the end of Fall 2010.

In December 2010, the TMPC (Harald Dahle, George Magklaras) was involved in activities to help the Greek EMBnet node with the process of making a web enabling Protein-to-Protein-Interaction application (<http://superclusteroid.uio.no>). The help involved the TMPC in arranging a domain name and a web server instance for them to run their application, as well as integrating their application to a number of standard utilities.

The Finnish EMBnet node: AGM 2010 report



Kimmo Mattila

CSC – IT Center for Science, Espoo, Finland

CSC – IT Center for Science (<http://www.csc.fi>) is a non-profit company owned by the Finnish Ministry of Education, Science and Culture. With a staff of about 200 persons, CSC provides a wide range of information technology support and resources for the Finnish academic and research institutes. Currently, the services include computing, application, networking and data services for a wide variety of sciences, ranging from linguistics to physics.

Bioinformatics services

Bioscience researchers have for a long time been the largest user group of CSC. In 2009, 345 bioscientists used the computing servers of CSC, and approximately 200 more used applications developed or licensed by CSC. The bioinformatics software and databases, installed on the servers of CSC, form the core of our bioinformatics services. The application palette contains sequence databases, tools for sequence analysis, phylogenetics, gene mapping, DNA microarray analysis and structural biology. The support for next generation sequencing data analysis software is just emerging. The complete list of the over 50 available applications and databases can be found from the CSC- BioBox pages: <http://www.csc.fi/molbio>. In addition to the instruction pages, CSC provides on-line support by e-mail and phone, and arranges bioinformatics related training courses and events.

At the moment, the main platform for the bioinformatics services is a HP ProLiant system with 64 computing cores. For heavier computing, CSC has two clusters and a Cray XT-4 supercomputer which together have over 15 000 computing cores. Using the services of CSC requires, in most cases, a CSC user account. During the last

few years, about 4-6 people have been actively maintaining and developing the bioinformatics services of CSC.

Software development

During the last years, CSC has invested significantly in software development, especially in the Chipster microarray data analysis tool (<http://chipster.csc.fi/>). Chipster brings a comprehensive collection of up-to-date analysis tools within the reach of bioscientists via its graphical user interface. The software was originally used for microarray data, but it is a generic tool and it now contains functionality also for proteomics, sequence analysis and next generation sequencing data.

National and EU collaboration projects

CSC is responsible for the National Grid Initiative (NGI) activities in Finland, and was one of the responsible organizations when the first significant grid research network, M-grid, for materials science, was being realized. CSC is also participating in several European Grid projects: these range from technology-oriented Grid projects, like EGEE2 and DEISA, to collaborative projects, like ELIXIR and EMBRACE.

The ICGEB EMBnet node: AGM 2010 report



Sándor Pongor

ICGEB, Trieste, Italy

ICGEB, the International Centre for Genetic Engineering and Biotechnology (<http://www.icgeb.org>) is a non-profit international organization that provides scientific research and an educational environment for the benefit of developing countries. With laboratories in Trieste, Italy (headquarters), New Delhi, India and Cape Town, South Africa, the ICGEB forms an interactive network with Affiliated Centres in 39 of its 60 Member States. ICGEB is part of the United Nations System. At present, more than 400 people from 38 different countries are working in the ICGEB laboratories as research scientists, postdoctoral fellows, PhD students, research technicians and administrative personnel. Research and training programs are focusing on biomedicine, crop improvement, environmental protection/remediation, biopharmaceuticals and biopesticide production. External services include biosafety counselling, biotechnology transfer and bioinformatics.

Bioinformatics services

Bioinformatics was the first service started at ICGEB in 1990, established with the aim of providing access to sequence databases, computing facilities (GCG). Gradually, the emphasis has shifted; today ICGEBnet mainly offers courses and consultation in bioinformatics, as well as open bioinformatics services - these are mostly based on web servers developed within the Protein Structure and Bioinformatics Group which is responsible for the project. The use of ICGEBnet is available free of charge to all ICGEB Signatory Country scientists. Protein Tools include protein domain prediction from sequence and structure, as well as access to the SBASE protein domain library, a Benchmark collection for protein clas-

sification, and various services for the analysis of protein 3D structures. DNA Tools include servers to predict and visualize bent regions in DNA sequences, as well as plots of various structural parameters along the sequence. With a total of over 1300 students since 1990, the annual course "Bioinformatics: Computer Methods in Molecular Biology" offers lectures and practical sessions on biological sequence analysis as well as an introduction to the major bioinformatics services available worldwide. Participation is free for selected applicants from ICGEB Member States. ICGEBnet provides tutorials and consultation for the students of the Ph.D. course at ICGEB-Trieste.

At the moment, the main platform for our bioinformatics services is a Sun Fire V20z system, with 22 computing cores equipped with 8GB RAM. During the last few years, 1-2 people have been actively maintaining and developing the bioinformatics services at ICGEB. At present, there are 327 registered users.

Software development

Bioinformatics tools developed by the group:

DNA tools	
Bend.it ¹	This server predicts DNA curvature from DNA sequences.
Plot.it ²	This server plots various physicochemical, statistical or locally computed parameters along DNA sequences (1-D or sequence-plots) or against each other (2-D plots).
Model.it ³	This server produces a 3D model of a DNA molecule, given a sequence of maximum 700 nucleotides.
Introns ⁴	Intron phase pattern conservation server.

1 http://hydra.icgeb.trieste.it/dna/bend_it.html

2 http://hydra.icgeb.trieste.it/dna/plot_it.html

3 http://hydra.icgeb.trieste.it/dna/model_it.html

4 <http://hydra.icgeb.trieste.it/~kajla/introns/>

<u>Protein tools</u>	
Sbase ¹	Support vector machines, domain prediction system
Fthom ²	Predict domains in your sequence
P450 ³	A directory of p450-containing systems
CX ⁴	This server calculates the atomic Protrusion Index from a three-dimensional protein structure.
Pride ⁵	This server calculates the PRobability of IDentity between three-dimensional domains (or whole structures)
DPX ⁶	This server calculates the atom depth from a three-dimensional protein structure.
Benchmark ⁷	A protein classification benchmark collection for testing machine-learning algorithms
prideNMR ⁸	NMR Protein fold similarity server
theGPM ²	This is an in-house modified version of the Gpm global proteome machine, a tandem mass spectrometry data analysis server.

1 <http://hydra.icgeb.trieste.it/sbase/>

2 <http://hydra.icgeb.trieste.it/sbase/sbase.php?sec=analyse&sub=predict>

3 <http://www.icgeb.org/~p450srv/>

4 <http://hydra.icgeb.trieste.it/cx/>

5 <http://hydra.icgeb.trieste.it/pride/>

6 <http://hydra.icgeb.trieste.it/dpx/>

7 <http://net.icgeb.org/benchmark/>

8 <http://net.icgeb.org/pridenmr/>

9 http://proteome.icgeb.trieste.it/tandem/thegpm_tandem.html

The French EMBnet node: AGM 2010 report



Guy Perrière

Laboratoire de Biométrie et Biologie Évolutive, Université Claude Bernard, Lyon, Villeurbanne Cedex - France

The ReNaBi (*Réseau National des plates-formes en Bioinformatique*) is the French EMBnet node since 2008. Present head of ReNaBi is Claudine Médigue (Génoscope, 2 rue Gaston Crémieux, 91057 Evry Cedex) and its delegate for EMBnet is Guy Perrière. This structure is a national network of 13 bioinformatics platforms officially commissioned by IBISA a national French agency. IBISA individually evaluates each platform every four years in order to determine if it still can be labelled as such. Among the requirements for a labelling are:

- self-funding of the platform
- dedicated personnel
- regular formation and teaching activities
- public on-line services offered

At this date, all the ReNaBi platforms met those requirements. Those platforms are also affiliated with many research laboratories and universities all around France (in Bordeaux, Lyon, Marseille, Montpellier, Paris, etc.) and are routinely used to assist research, mainly at academic level.

Due to the fact ReNaBi gathers many sites having a broad range of activities, the computing services offered cover the whole spectrum of bioinformatics:

- access to general or specialized sequence databases
- alignment and similarity search programs
- general sequence analysis package
- biostatistics packages
- molecular phylogeny programs
- tools for proteomics and transcriptomics data analysis

- tools for protein structure prediction and modelling
- Next Generation Sequencing (NGS) specific programs and pipelines

As all the platforms are independent, it is not possible to give a global financial assessment. The ReNaBi itself receives a recurring funding of 50,000€ from IBSA in order to support scientific animations such as workshops, conferences or thematic networks. As for the conferences supported, the main one is the French national conference in bioinformatics: JOBIM (*Journées Ouvertes en Biologie Informatique et Mathématique*). Among the different thematic networks, one is devoted to the use of grid computing and one to NGS users.

Again, due to its very own structure, it is difficult to give the complete list of machines available through ReNaBi. Standard equipment for a ReNaBi platform consist usually in a small computing cluster with about 100-500 cores, a mail server, one or two databases server(s), and a set (of variable size) of micro-computers. If we take the example of the PRABI (*Pôle Rhône- Alpes de Bioinformatique*), we have:

- one Dell PowerEdge 2950 (2×quadcore CPU@2.66 GHz, 8 Gb RAM, 146 Gb disk)
- one Dell PowerEdge 2950 (2×quadcore CPU@2.66 GHz, 8 Gb RAM, 600 Gb disk)
- one Dell PowerEdge 2950 (2×quadcore CPU@2.66 GHz, 32 Gb RAM, 900 Gb disk)
- three Sun Fire X4500 M2 (8×quadcore CPU@2.3 GHz, 64 Gb RAM, 3×145 Gb disk)
- two Sun Fire X4500 M2 (8×bicore CPU@2.8 GHz, 64 Gb RAM, 3×145 Gb disk)
- one Sun Fire V490 (4×bicore CPU@1.5 GHz, 16 Gb RAM, 2×146 Gb disk)
- one Sun Fire 880 (8 CPU@900 MHz, 28 Gb RAM, 6×36 Gb disk)

This, for the sole genomic aspects covered by this platform. Indeed, all aspects related to protein structure prediction and biostatistics have also their own sets of dedicated computers.

Lastly, the ReNaBi is involved in the Elixir European initiative. Particularly, we plan to modify its legal status in order to apply for being a node in the forthcoming Elixir infrastructure.

The Norwegian EMBnet node: AGM 2010 report



George Magklaras

Biotek - UiO, Oslo, Norway

The Norwegian EMBnet node has 65 members, and offers a range of life science computing services:

- official mirror of the EMBL, UniProt and Genbank databases
- MRS 4 install (command line and web)
- EMBOSS (command line and web)
- GCG (holders of license for legacy users)
- R tool
- dedicated application hosting (provision of relational databases and web services for science groups)

We have our own dedicated machine room, with nightly incremental backup, offering a variety of systems to our members:

- several memory intense systems (up to 64 GB RAM)
- total of 40 CPU cores
- Multi-Tbyte capable filesystem
- GPU cluster (work in progress, not yet operational)

At present, the node employs two staff members and is active with the EMBnet TM PC, specifically with:

- technical issues
- HTS IT (published HTS IT draft report, as well as articles on EMBnet News)

Our users are affiliated with Universities all over Norway, as well as private companies working in the field of Life Sciences. As such, our services assist both in the research at the University level as well as the development of commercial products.

At present, we are offering a course on sequence mining using MRS and EMBOSS, with the aim of:

- introducing students to some commonly used sequence databases
- introducing students to sequence mining tools, primarily MRS and EMBOSS

- familiarizing the students with the command line interfaces to the tools, and the possibilities that they open up with regards to creating pipelines.

This course has been offered at the Mexican EMBnet node in March 2010, and will be offered at the University of Oslo later this summer. Feedback from the course in Mexico has been very favorable. The course material has been made available online, as has video recordings of the course.

In the past 3 years, we have also offered courses on the following subjects:

- Bioperl (July 2008)
- R (January 2009)
- EMBOSS/GCG (June 2009)

In terms of new areas of research, we believe that GPU computing is an exciting new field, and will provide significant improvements over CPU-based systems in terms of computational power for the price. Certain types of bioinformatics algorithms can be run in massively parallel mode using GPU cores and thus benefit from this technology.

To enable us to work in this new field, we have applied for, and received, funding to purchase a small GPU cluster. We have acquired a server equipped with 4 NVIDIA Tesla c1060 GPU cards, for a total of 960 processing cores. Two of these cards will be replaced by the next generation Fermi c2050 cards as soon as these are available on the market, bringing the total up to 1376 cores. The current configuration gives a processing power of approximately 3600GFLOPs. After the upgrade, this will be increased to approximately 4400GFLOPs. This system will enable us to both build competence in this important field, and allow our members to run their algorithms in an adequately powerful system. Our plan is to port various bioinformatics algorithms to the GPU processors and make them available to the node members by the end of 2010. We also plan to share our expertise with the EMBnet community.

The Norwegian EMBnet node wishes to acknowledge its user base and the Molecular Life Science committee of the University of Oslo for providing funding to achieve these goals. (<http://www.uio.no/forskning/tverrfak/mls/kjernefasiliteter/miniplattform/>).

The South African EMBnet Node: AGM 2010 report



Winston Hide, Alan Christoffels

The South African National Bioinformatics Institute (SANBI), University of the Western Cape, Bellville

[The South African National Bioinformatics Institute](#) (SANBI) conducts high-quality scientific research focused upon delivery of translatable biomedical discoveries, primarily through local and international collaboration with partner organisations. SANBI is part of the University of the Western Cape, situated outside Bellville near Cape Town. The Institute is headed by a Director, who reports through the Faculty of Natural Sciences, and provides overall leadership to the organization. The Institute consists of a group of faculty supported by technical and administration staff, guiding research of a group of Masters and PhD students and Post-Doctoral scientists.

SANBI became a member of the [European Molecular Biology Network](#)² in 1997. It developed close relationships with faculty at the University of Witwatersrand and University of Pretoria, supporting training and research there and at other sites around the country.

SANBI is well recognised in the areas of gene expression and host-pathogen disease research, including HIV, Trypanosomes and Malaria; and in the provision of bioinformatics and biomedical informatics training. This recognised expertise and proven capacity development has enabled the Institute to secure additional funding from a number of high-profile international agencies, e.g., National Institutes of Health, to expand its training programmes with the aim of developing faculty capable of producing NIH-funded research.

1 <http://www.sanbi.ac.za/>

2 <http://www.embnet.org/>

The Institute provides long-term skilling to impact on diseases prevalent in Africa, in particular the discovery of genetic factors that contribute to disease resistance in hosts, e.g., HIV and Malaria, and the genetic relationship contributing to cancers. SANBI's first major scientific breakthrough was in 1999, in collaboration with US investigators, and resulted in the discovery of a genetic cause for a type of blindness in humans called retinitis pigmentosa. SANBI became the bioinformatics research centre for the Centre for AIDS Programme in South Africa in 2003, and for the South African AIDS Vaccine Initiative in 2005, and has been integral to HIV research in South African vaccine development. The Institute has recently delivered high-impact publications in the area of mammalian gene regulation, with a resulting expansion in knowledge of how genes deregulate in cancers. The Institute has had a strong influence on the development of health biotechnology locally, and internationally, through development and deployment of software tools for genetic research, adopted by international biotechnology companies, such as Affymetrix, and by over 400 research institutions worldwide.

Staffing

SANBI recently went through a change in management as a result of the resignation of the founding Director, Winston Hide, who moved to Harvard School of Public Health, and now holds a visiting professorship at SANBI. In addition, owing to the creation of the King Abdullah University of Science and Technology, a group of faculty and staff also moved from SANBI, in 2009, to establish bioinformatics in Saudi Arabia.

New Faculty have subsequently been appointed, including a National Bioinformatics Research Chair holder, the interim director, Alan Christoffels, and Simon Travis (HIV Specialist). Owing to the change in management, and movement of several members of the institution to new positions both internally and elsewhere, continuity with EMBnet has had to be renewed. SANBI has completed the process of hiring three additional research staff members, one of which will become our new EMBnet Node manager.

Impact of EMBnet on South Africa and SANBI

The EMBnet node and management model has been of great impact in Africa. First, EMBnet

actively supported the establishment of SANBI and provided training to our staff and subsequently to several of our institutions. Second, the model of management and node structure was adopted by SANBI in its drafting of a successful proposal to establish the South African National Bioinformatics Network. Although the network took up an aggressive 3-year training programme, it has subsequently been disbanded, as government priorities have changed. Now, more than ever, SANBI and African sites need continuity with European programmes such as EMBnet, in particular in terms of training opportunities and co-development. SANBI is currently a development site for the Galaxy system, providing a development nexus for annotation systems, in addition to being a site for deployment of EMBOSS.

Service

Since becoming a member of EMBnet, SANBI has provided online and remote support for bioinformatics activities throughout South Africa. It provides online specialized tools through its website, and complements these with specialised training and databasing services, such as those for HIV, and focused support of national and local research efforts in trypanosomes, tsetse fly genomics, cancer, food safety, and related biotechnology projects. Faculty and staff at SANBI have provided onsite trainings for other institutions, and a staff member also has been seconded to the national health laboratory service, to provide expertise and training remotely over longer periods of time.

Training

SANBI hosts bioinformatics training workshops for African scientists funded through the WHO, the UK Wellcome Trust, the SA National Research Foundation, the US National Institutes of Health and the Centres for Disease Control. With the disbanding of the National Bioinformatics Network, SANBI continues to fill the gap in national training, by coordinating 6-week introductory bioinformatics courses across the country. In February 2010, the national course was attended by 30 delegates, representing 5 universities. Annually, SANBI hosts a regional Ensembl training course for 25 delegates, which is presented by an EBI trainer. Most recently, SANBI hosted a national workshop, for 45 attendees, on genomics data interpretation, with presenters from South Africa,

Harvard and Stanford. In addition, SANBI offers formal degrees at PhD and Masters level. SANBI is a founding member of the African Society for Bioinformatics and Computational Biology, and will be a host organisation for the 2011 regional meeting of the International Society for Computational Biology in Cape Town.

Facilities

The Institute has adequate scientific computer infrastructure, and is the site for a pair of high performance 32 CPU IBM P-690 servers and, in the next month, an 8 CPU Xserve cluster, which provide a significant proportion of the research compute infrastructure for bioinformatics in Africa. SANBI scientists have workstations and Internet access, as well as backup and disk storage. In late 2010, SANBI will move its premises to a new building, offering extensive training facilities, a visiting scientist facility and meeting rooms, in addition to the research and service provision currently performed.

Research

- Delivery of an African driven analysis and annotation of *Glossina*, the vector for the tsetse fly.
- Through capacity developed from genome annotation, we apply expertise and technologies developed to other relevant organisms to African health, with a particular emphasis upon integration of HIV clinical, immune and sequence diversity outcomes, and pathways analysis in Malaria.
- We integrate comparison of vector, host and pathogen genomes to deliver unique African knowledge of HIV and Malaria.
- Develop and apply understanding of normal and diseased human gene expression to diseases relevant to South Africans.
- Develop capacity of scientists through tightly defined research projects that have high impact on health in South Africa.

EU Collaborative projects

SANBI faculty and post docs have enjoyed funded collaborations through the EU FP funding programmes and currently we serve on the SYSCO programme, together with Institut Pasteur Tunis and Paris, and Max Planck Berlin.

The Swedish EMBnet Node: AGM 2010 report



Nils-Einar Eriksson

Computing Department of Uppsala Biomedical Centre (BMC), Uppsala, Sweden

People working for EMBnet Sweden are Nils-Einar Eriksson (TM PC), Emil Lundberg, Martin Norling (PR PC) and Erik Bongcam-Rudloff (EMBnet chairman 2003-2009).

Services

EMBnet Sweden has a web-site connected to several unique tools produced by members of EMBnet together with their associated researchers. EMBnet's video conference system (Marratech) is located and managed at the Swedish node. Uppsala University is providing mail-list services for EMBnet. The lists are managed by node personnel. DNS master services for EMBnet are also provided by the Swedish node.

SeqScoring

The tool can be used to up-load your SNP- and indel- files from large re-sequencing projects, and get the data scored by conservation across species: www.seqscoring.net.

EVALLER

EVALLER™ is a web-tool wherein you can electronically test (e-Testing) a protein's potential allergenicity/cross-reactivity based on its amino acid sequence (<http://bioinformatics.bmc.uu.se/evaller.html>).

MolMeth

MolMeth is a structured database that provides free access to methods used in molecular biology and molecular medicine. Submitted methods and contributions are subject to curation. www.molmeth.org.

RetroTector online

A web-based tool for analysis of retroviral elements in small and medium size vertebrate genomic sequences: <http://retrotector.neuro.uu.se/>.

EMBnet Sweden has also been involved in the development of several bioinformatics packages and tools for the MacOSX platform:

- eBiotools: A collection of bioinformatics tools (>200): EMBOSS, Staden, T-coffee, ClustalW, and many others
- BioX: a graphical interface for eBiotools
- eBioKit: a server loaded with bioinformatics tools and databases, a solution for small and medium size research groups. The eBioKit is now used in more than 10 countries worldwide.

The node was one of the main organizers of the Next Generation Sequencing technologies workshop organized in Rome, November 2009, and of part II of the workshop in Bari-Italy, July 2010. (www.nextgensequencing.org).

Selection of Publications (2009-2010)

1. Pettifer S, Ison J, Kalas M, Thorne D, McDermott P, Jonassen I, Liaquat A, Fernández JM, Rodriguez JM, Partners I, Pisano DG, Blanchet C, Uludag M, Rice P, Bartaseviciute E, Rapacki K, Hekkelman M, Sand O, Stockinger H, Clegg AB, Bongcam-Rudloff E, Salzemann J, Breton V, Attwood TK, Cameron G, Vriend G. The EMBRACE web service collection. *Nucleic Acids Res.* 2010 May 12. [Epub ahead of print] PMID: 20462862.
2. Markus Klint , Mikael Tholleson , Erik Bongcam-Rudloff , Svend Birkelund , Anders Nilsson and Bjorn Herrmann. Mosaic structure of intragenic repetitive elements in histone H1-like protein Hc2 varies within serovars of *Chlamydia trachomatis*. *BMC Microbiology* 2010, 10:81doi: 10.1186/1471-2180-10-81 17 March 2010
3. Danika Bannasch, Amy Young, Jeffrey Myers, Katarina Truvé, Peter Dickinson, Jeffrey Gregg, Ryan Davis, Erik Bongcam-Rudloff, Matthew T. Webster, Kerstin Lindblad-Toh, and Niels Pedersen. Localization of Canine Brachycephaly Using an Across Breed Mapping Approach. *PLoS One.* 2010; 5(3): e9632. Published online 2010 March 10. doi: 10.1371/journal.pone.0009632
4. Heli Salminen-Mankonen, Jan-Eric Litton, Erik Bongcam-Rudloff, Kurt Zatloukal and Eero Vuorio. The Pan-European research infrastructure for Biobanking and Biomolecular Resources: managing resources for the future of biomedical research. *EMBnet.News.* 15.2. pp. 3-8. July 2009.
5. Álvaro Martínez Barrio, Erik Lagercrantz, Göran O Sperber, Jonas Blomberg, Erik Bongcam-Rudloff. Annotation and visualization of endogenous retroviral sequences using the Distributed Annotation System (DAS) and eBioX. *BMC Bioinformatics.* BMC Bioinformatics 2009, 10(Suppl 6):S18doi:10.1186/1471-2105-10-S6-S18
6. Domenica D'Elia , Andreas Gisel , Nils-Einar Eriksson , Sophia Kossida , Kimmo Mattila , Lubos Klucar and Erik Bongcam-Rudloff. The 20th anniversary of EMBnet: 20 years of bioinformatics for the Life Sciences community. *BMC Bioinformatics.* BMC Bioinformatics 2009, 10 (Suppl 6):S1doi:10.1186/1471-2105-10-S6-S1
7. R. P. Joosten, J. Salzemann, V. Bloch, H. Stockinger, A.-C. Berglund, C. Blanchet, E. Bongcam-Rudloff, C. Combet, A. L. Da Costa, G. Deleage, M. Diarena, R. Fabbretti, G. Fettahi, V. Flegel, A. Gisel, V. Kasam, T. Kervinen, E. Korpelainen, K. Mattila, M. Pagni, M. Reichstadt, V. Breton, I. J. Tickle and G. Vriend. PDB_REDO: automated re-refinement of X-ray structure models in the PDB. *J. Appl. Cryst.* (2009). 42 [doi:10.1107/S0021889809008784]
8. Sperber G, Lövgren A, Eriksson NE, Benachenhou F, Blomberg J. RetroTector online, a rational tool for analysis of retroviral elements in small and medium size vertebrate genomic sequences. *BMC Bioinformatics.* 2009 Jun 16;10 Suppl 6:S4
9. Álvaro Martínez Barrio, Marie Ekjerlund, Göran O Sperber, Jonas Blomberg, Erik Bongcam-Rudloff and Göran Andersson. In silico analysis of the dog genome identifies Canine Endogenous Retroviruses (CfERVs). *Frontiers of Retrovirology: Complex retroviruses, retroelements and their hosts* Montpellier, France. 21-23 September 2009. *Retrovirology* 2009, 6(Suppl 2):P7doi:10.1186/1742-4690-6-S2-P7
10. David E. Gloriam, Sandra Orchard, Daniela Bertinetti, Erik Bjorling, Erik Bongcam-Rudloff, Julie Bourbeillon, Andrew R. Bradbury, Antoine

de Daruvar, Stefan Dubel, Roanld Frank, Toby J. Gibson, Niall Haslam, Friedrich W. Herberg, Tara Hiltke, Jorg D. Hoheisel, Samuel Kerrien, Manfred Koegl, Zoltan Konthur, Bernhard Korn, Ulf Landegren, Silvere van der Maarel, Luisa Montecchi-Palazzi, Sandrine Palcy, Henry Rodriguez, Sonja Schweinsberg, Volker Sievert, Oda Stoevesandt, Michael J. Taussig, Mathias Uhlen, and Christer Wingren. A community standard format for the representation of protein affinity reagents. *Molecular and Cellular Proteomics*. August 2009.

The contribution of the eBioKit to Bioinformatics Education in Southern Africa



Yasmina Jaufeerally-Fakim¹, Hans-Henrik Fuxelius², Erik Bongcam²

¹Faculty of Agriculture, University of Mauritius, Mauritius, ²Swedish University for Agricultural Sciences, Uppsala, Sweden

As bioinformatics has been making major progress and contributing to the development in the rest of the world, it has still not yet fully integrated the tertiary education and research sector in the countries of Southern Africa. In this context SANBio (Southern African Network for Biosciences)



Figure 1. Students of the University of Mauritius attending the eBioKit Education.

launched a project in 2009, on capacity building in this area so as to address the immediate needs of the scientific community within the region. This project is funded under the BIOFISA program from Finland. The main challenge is to bring to the scientists the knowledge and skills required to fully tap the resources available in the public domain. With an already established network of researchers, a series of workshops were initiated so as to train those in specific fields to become familiar with some of the commonly

used tools for DNA/protein analyses. Participants are from Malawi, Zambia, Namibia, Botswana, South Africa, Zimbabwe as well as Mauritius. Not all these countries however have easy access to the Internet or the connection can be very erratic. Although the SAFE cable has been in use for sometime, the SEACOM is relatively recent and many parts of Southern Africa do not yet enjoy good connection.

In order to address this shortcoming, the first action of the SANBio project was to put in place a training on the utilisation of the eBioKit. The latter allows a number of users to directly access to the inbuilt databases and work with the tools for analyses via their intranet. This makes teaching of a large class relatively easy with students being able to get their results very quickly. It has some of the most popular tools for alignment, phylogeny, as well as genome browsers etc. The databases include microbial, human and plants genomes. In itself it contributes to the practical sessions of a whole module in bioinformatics. The first workshop was held in Mauritius in February 2010 with Erik Bongcam-Rudloff and Hans-Henrik Fuxelius from the Swedish University for Agricultural Sciences (SLU), Uppsala, Sweden.

The material covered in this course was alignment and phylogeny software on real world data as well as homology searching on major protein databases locally stored on the eBioKit. EMBOSS and Phylip packages was primarily used for building the phylogenies and everything was accessed through the web interface to the eBioKit server. Once the process has been mastered on how to use these tools the participants can teach their colleagues on the same platform with the same material.



Figure 2. Dr. Hans-Henrik Fuxelius lecturing on homology searches using the eBioKit.



Figure 3. Dr. Erik Bongcam-Rudloff lecturing on using ensemble with the eBioKit.

Most of the participants were at their first hands-on training in using such tools and by the end they were eagerly waiting to have the kit at their own institution. Two kits have been delivered to Zimbabwe, one at the Tobacco Research Board and the other at the University of Zimbabwe. There are now two eBioKits at the University of Mauritius. The rest have been shipped to the other countries. Those who got their training are preparing to run their own local workshop; Mauritius and Malawi will hold one each for their local participants in July.

This kit has been received with great excitement and above all it is a means to allow the rapid dissemination of knowledge in Bioinformatics. Both faculty members and students are utilising the resources. It is a catalyst to further upgrade teaching and research standards of the universities in the region. The utilisation of the eBioKit will set the pace for further development in the application of Bioinformatics among scientists in Southern Africa. It is indeed a great way of providing the means for enhancing the quality of research output and above all give students the opportunity to know better the exciting world of genomes and gene expression.

Efficient functional bioinformatics tools: towards understanding biological processes



Alberto Pascual-Montano^{1,2,*}, Jose M. Carazo¹

¹National Center for Biotechnology, CNB-CSIC, Madrid Spain

²Institute for Molecular Medicine Príncipe de Asturias, IMMPA-CSIC, Madrid Spain

*Corresponding author: pascual@cnb.csic.es

Abstract

Experimental high-throughput techniques in biology are generating large amounts of data related to genes and proteins at different levels. The functional analysis of such datasets is a necessary key step in their interpretation. In this contribution we review different methodologies and applications for functional bioinformatics in the context of automated data and text analysis in biology.

Introduction

The post-genomics era has been characterized by the emergence of several high-throughput techniques such as DNA microarrays, protein chips, chip-on-chip, and more recently, the new generation sequencers [1-4], which are allowing the scientific community to study biological processes from a global perspective to understand the basis of development and diseases. Nevertheless, the great potential of these technologies is resulting in a problem when researchers have to deal with the vast amounts of data that are generated by these technologies.

Life sciences in general are generating huge amounts of data with the accelerated development of high-throughput technologies. Therefore, in parallel to the accumulation of the experimental information, there is an obvious need to analyse this huge amount of data. The challenge lies not only in the data-processing area, in which the bioinformatics community has made significant

advances, but also in the interpretation of the experimental data. An essential task in this analysis is to translate gene signatures into information that can assist in understanding the underlying biological mechanisms.

During the last few years, the Functional Bioinformatics Group from the National Center for Biotechnology in Madrid (<http://bioinfo.cnb.csic.es>) has focused on the development of several methods and tools to assist researchers in the analysis and interpretation of genomics and proteomics data. One of our main goals has been to provide the research community with easy-to-access and easy-to-use tools, implementing complex data-analysis methodologies that allow a more complete understanding of cellular processes and diseases. In this way, we have generated a set of web-based applications that addresses three main fields: gene-expression data analysis, functional interpretation of large lists of genes or proteins and automatic knowledge extraction from the biomedical literature.

In our effort to develop applications for a broad number of users able to deal with computing-intensive problems, we have designed these tools incorporating efficient implementations of the algorithms and their variants, most of them using parallel and grid computing. In addition, most of these tools can be accessed via web-services, which allow users to integrate them in their analysis workflows.

Applications

The mission of our research group is the development and implementation of algorithms that facilitate the understanding of biological processes through the application of statistical and data-mining techniques. Figure 1 shows an overview of the applications and databases that we have developed, can be divided into three different groups:

Data analysis

This comprises applications for data clustering and biclustering, as well as data accommodation, normalization and pre-processing. Tools like Engene (<http://engene.cnb.csic.es>) [5] were designed for microarray data analysis, from pre-processing to clustering and classification. biONMF (<http://bionmf.cnb.csic.es>) [6], on the other hand, performs more complex analysis (clustering and biclustering) using specific matrix factorization. Finally, MARQ was designed to retrieve ex-



Figure 1. General overview of the applications available at the Functional Bioinformatics group from the National Center for Biotechnology. A brief explanation of the content of each web-based tool is provided.

perimental datasets from the GEO database [7] that have expression patterns similar or opposite to a given query. This is useful for meta-analysis studies.

Functional analysis

This group is composed of different analysis tools that aim at providing functional interpretation to a list of genes or proteins. Usually, this is a secondary step, after primary analysis of gene or protein expression. Applications for functional analysis in gene or protein sets use annotations from different data sources and calculate statistics to determine their significant presence or absence in the list. The GeneCodis tool searches for significant concurrent annotations in a list of genes (<http://genecodis.cnb.csic.es> [8]). Another application, ChipCodis (<http://chipcodis.cnb.csic.es> [9]),

mines regulatory systems in yeast using chip-on-chip data using a similar approach. Text mining tools, like SENT (<http://sent.cnb.csic.es> [10]) and Moara (<http://moara.cnb.csic.es> [11]), also belongs to this category, as the final goal is to provide functional interpretation of gene sets using the scientific literature as the main source of information.

Databases

Databases that collect and centralize interesting unique information about biological events are also a central part of our research interest. Centrosome:db (<http://centrosomedb.cnb.csic.es> [12]) is a database that contains a set of human genes encoding proteins that are localized in the centrosome. It offers different perspectives to study the human centrosome, including evolu-

tion, function, and structure. The ProteoPathogen [13] and Complayeast databases were designed to handle and manage proteomic studies.

In this report, we will focus our attention on four representative tools taken from data-analysis and functional-analysis categories. These tools are widely used by the scientific community, handling several thousand submissions every month:

GeneCodis

Gene Annotation Co-occurrence Discovery

An essential task in functional analysis of genes and proteins is the translation of gene signatures into information that can assist in understanding the underlying biological mechanisms. Several methods and tools have been developed to interpret large lists of genes or proteins using information available in biological databases. The common idea used in most of these methods is to find functional descriptors that are significantly enriched in the gene signature.

The first type of techniques that emerged in this field were focused on evaluating the frequency of individual annotations and apply a statistical test to determine which annotations are significantly enriched in a input list with respect to a reference list, usually the whole genome or all genes in a microarray. Annotations from different sources like the Gene Ontology (GO) [14] or KEGG [15] are commonly used in this context. Several tools have been developed following this approach [16,17]. A fresh line of research appeared with the observation that the use of thresholds to select the significant genes could underestimate the effect of significant biological effects during the functional analysis. This idea derived in a new and different analytical concept in which the distribution of annotations is evaluated over the whole list of genes, sorted by their correlation with the phenotype: the gene set enrichment analysis (GSEA) [18].

Nevertheless, both approaches, the standard enrichment analysis and GSEA methods, evaluate each annotation independently from the others without taking into account the potential relationships among them. However, most of the annotations in biological databases are interconnected because they are associated to common genes. Patterns that contain these relationships can provide invaluable information and extend our understanding of biological

events associated with the experimental system. It is therefore highly desirable to evaluate these associations in the functional analysis of gene lists [19].

In 2007, we introduced GeneCodis [20], a tool for modular enrichment analysis oriented to integrate information from different sources and find enriched combinations of annotations in large lists of genes or proteins. Modular enrichment represents a step forward in the functional analysis because of its capacity to integrate heterogeneous annotations and to discover significant combinations among them. A new version of this tool has recently been reported [8].

The application is simple in its concept: it takes a list of genes as input, and determines biological annotations or combinations of annotations that are over-represented with respect to a reference list. The novelty of this tool relies on the fact that, before computing the statistical test, it incorporates new functionality to extract all combinations of annotations that appear in at least x genes, x being a user-defined threshold [20]. To extract combinations of gene annotations, GeneCodis uses a modification of the methodology reported in [21], which implements an efficient variant of the apriori algorithm to extract associations among gene annotations and expression patterns.

Once all combinations of annotations that appear in at least x genes have been extracted, the method counts the occurrence of each set of annotations in the list of genes and in a reference list. Then, a statistical test is applied to identify categories, and their combinations, that are significantly enriched in the list of genes. Therefore, the result of the analysis performed by GeneCodis consists of a list of annotations or combinations of annotations with their corresponding p-values. Annotations showing p-values below a certain threshold can be considered significantly associated with the list of genes under study and can be used to discern the biological mechanisms relevant to the experimental system.

GeneCodis works with most of the model organisms in biological research. The whole list includes *Arabidopsis thaliana*, *Bos taurus*, *Caenorhabditis elegans*, *Candida albicans*, *Danio rerio*, *Drosophila melanogaster*, *Escherichia coli*, *Gallus gallus*, *Homo sapiens*, *Leishmania major*, *Mus musculus*, *Rattus norvegicus*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* and

Vibrio cholerae. Alternatively, to facilitate the usability of the application we have extended the type of gene identifiers that are supported, including proprietary identifiers from commercial platforms such as Affymetrix, Agilent, Codelink and Illumina. Regarding annotations, GeneCodis integrates functional and regulatory information by supporting Gene Ontology (GO), KEGG Pathways, micro RNA, transcription factors and InterPro motifs.

This tool can be freely accessed through its main site at <http://genecodis.cnb.csic.es>. A mirror has also been recently set up at the China Node for EMBnet at the Center for Bioinformatics in Peking University, available at <http://genecodis.cbi.pku.edu.cn>.

bioNMF

Non-negative Matrix Factorization in Biology

Matrix factorization techniques have become well established methods for the analysis of numerical data and signals. These methods can be applied to the analysis of multidimensional datasets in order to reduce the dimensionality, discover patterns and aid in the interpretation of the data. Among the most popular, Principal Component Analysis (PCA), Singular Value Decomposition (SVD) or Independent Component Analysis (ICA) have been successfully used in a broad range of contexts.

In 1999, Lee and Seung, developed a new matrix factorization technique named Non-Negative Matrix Factorization (NMF) [22] to decompose images into recognizable features. The main difference between NMF and other classical factorization methods relies on the non-negativity constraints imposed by the model. These constraints tend to lead to a parts-based representation of the data, because they allow only additive, not subtractive, combinations of data items. In this way, the factors produced by this method can be interpreted as parts of the data or, in other words, as subsets of elements that tend to occur together in sub-portions of the dataset. By contrast, classical factorization techniques decompose the data matrix into a new set of matrices of any sign, involving complex cancellations of positive and negative elements to reconstruct the original dataset. Therefore, the interpretation of the factors becomes non-intuitive and difficult. The comprehensible properties of the NMF method and the intuitivism of the results

it provides have centered the attention of many researches in different fields of science and, in particular, in the bioinformatics field, where NMF has been applied in different contexts.

Formally, the non-negative matrix decomposition can be described as follow:

$$V \approx WH$$

where $V \in \mathbb{R}^{m \times n}$ is a positive data matrix with n variables and m objects, $W \in \mathbb{R}^{m \times k}$ are the reduced k basis vectors or factors, and $H \in \mathbb{R}^{k \times n}$ contains the coefficients of the linear combinations of the basis vectors needed to reconstruct the original data (also known as encoding vectors). The main difference between NMF and other classical factorization models relies on the non-negativity constraints imposed on both the basis (W) and encoding vectors (H). In this way, only additive combinations are possible:

$$(V)_{mn} \approx (WH)_{mn} = \sum_{l=1}^k W_{ml} H_{ln}$$

Increasing interest in this factorization technique is due to the intuitive nature of the method, which has the ability to extract additive parts of data sets that are highly interpretable, while reducing the dimensionality of the data at the same time. In the biomedical field, NMF-related methods have recently gained a lot of popularity. For example, gene expression analysis [23-31], analysis of protein sequences [32], functional categorization of genes [33] or biological text mining [34].

We have developed a user-friendly, web-based tool that implements this factorization technique and its application in data biclustering (clustering by rows and columns simultaneously) and sample classification [6,35]. In addition, this web tool offers new improvements to process big data-sets in a distributed computing environment without the usage complexity present on this kind of systems. It also provides an automated access to external applications through a web-services interface.

This online tool provides a user-friendly interface, combined with a computationally efficient parallel implementation of the NMF methods to explore the data in different analysis scenarios. In addition to the online access, bioNMF also provides the same functionality included in the web

site as a public web-services interface, enabling users with more computer expertise to launch jobs on bioNMF server from their own scripts and workflows. The bioNMF application is freely available at <http://bionmf.cnb.csic.es>.

SENT

Semantic Features in Texts

Applications like GeneCodis described above are examples of initiatives to help in the biological interpretation of lists of genes and proteins in the context of certain experimental conditions. Annotations-based approaches provide a fast, easy, and statistically sound interpretation of a list of genes or their products. Although this information is extremely useful, its scope is limited by structured vocabularies and curated annotations.

Literature mining offers an interesting alternative to annotation-based methods. The rationale behind it is that it contains much richer information about the function of genes that can be captured in structured vocabularies. Biomedical literature covers almost all aspects of biology and biochemistry, and with no limit to the types of information that may be recovered through careful and exhaustive mining [36]. Many researchers have focused their attention on the use of text mining, with methodologies that go from determining protein-protein interactions from biomedical texts [37-40], to providing summary descriptions for genes or determining their similarities [41-45]. Even though lots of work in this area has been reported, its practical use by the scientific community is hindered by the lack of efficient and easy-to-use software.

SENT is a usable implementation of a methodology based on [34]: for a given set of genes or proteins, the associated abstracts from PubMed are extracted and processed. Then the NMF technique (like the one implemented in the bioNMF application described above) is used to cluster genes together with relevant terms. In this way, a set of semantic features (topics or collection of semantically related words) are associated to a set of genes. In other words, genes are clustered according to the content of their associated literature and, at the same time, those semantic topics provide insights into their functional implications.

The input of the system is a set of gene or protein identifiers and the number of semantic

features (factors or topics) to extract in the analysis. Titles and abstracts from articles associated with each gene are used to produce a meta-document and, from all gene meta-documents, a term frequency matrix is created. This matrix is then analyzed by means of the NMF algorithm, yielding a set of semantic features and a way to associate genes to these semantic features.

The collection of articles used in the analysis is built into an index. This index can be queried to retrieve articles that mention certain terms. In particular, it can be used to find the articles that are most relevant to each semantic feature, and by extension, most relevant to understand the list of genes. Coupled with the Gene Ontology enrichment analysis (using GeneCodis), also provided in the web site, SENT serves as a guide in the examination of the literature.

This tool offers several advantages in the area of biomedical text mining: first, SENT is oriented to give researchers a global functional picture of their genes of interest by summarizing the associated literature content on a small set of semantic topics. Second, SENT is able to categorize the list of genes or proteins according to these topics, and also to associate them with Biological Process terms in the Gene Ontology. Finally this functionality, and the way it is implemented using web-service technology, allows researchers to easily include this analysis into their workflows, providing their research with one more piece of information to be taken into account. SENT is available at <http://sent.cnb.csic.es>.

MARQ

A tool to mine GEO by content

DNA microarrays have become an extensively used technique to analyze global gene-expression profiles. Its widespread usage quickly promoted the creation of gene-expression data repositories such as the NCBI Gene Expression Omnibus (GEO) [7], which currently holds more than 10,000 experiments for over 500 organisms. Despite the huge amounts of information available in GEO, researchers usually focus their analysis on individual experiments without exploiting such valuable resources. Gene-expression analyses are usually performed from a gene-centered perspective, with the goal of identifying relevant sets of genes that are then used to determine the biological mechanisms relevant in that experimental setup. However, besides this gene-

to-phenotype approach, gene signatures can also be used to establish connections among different physiological conditions [46]. For example, we can discover connections among diseases, drugs or pathways by similarities in their gene-expression signatures; conditions inducing similar signatures might be modulating the same pathways, while conditions with opposite signatures might be involved in reversing the original phenotype.

MARQ compiles a gene-expression signature database from all datasets in GEO for five model organisms (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana*). For each dataset, a number of signatures are extracted, each representing a potentially relevant comparison between the dataset samples. The query consists of two sets of genes, up-regulated and down-regulated (optionally one may be empty). A rank statistic is used to give a score and significance p-value to the relationship of each signature to the query based on how the input genes rank in differential expression for that signature. Signatures with a direct relationship receive a positive score and signatures with an inverse relationship receive a negative score. In this way, datasets, similar in direct or opposite relationship with the content of the query gene set are successfully extracted by MARQ.

There are two types of possible queries: platform based, and organism based (cross-platform) depending on the preference for searching. Additionally, each signature is annotated with additional meta-information, such as words in the dataset description or GO terms over-represented in the list of genes deregulated in the comparison. An annotation-mining tool can be used to find patterns in these annotations to highlight relevant features common to the most significant signatures.

Finally, a set of signatures of interest can be analyzed using a meta-analysis methodology based in Rank Products [47] which highlight genes commonly deregulated in those signatures. The application is freely accessible at <http://marq.cnb.csic.es>.

Conclusions

Biology is becoming more and more a science that necessarily needs to include information technology as a basic discipline in order to

understand the extremely complex biological processes of living organisms. Tools and methodologies able to efficiently process the huge amounts of data that are generated in this field are still demanded by the molecular biology community. In this report, we have described a set of bioinformatics tools publicly available online through web browsers or web-service interfaces. The applications are oriented to the efficient and robust analysis of gene-expression information and the functional characterization of lists of genes or proteins from different perspectives. Integrative analysis of all available data and biological information, as the ones described here, is probably one of the best ways to move towards a complete study of biology from a global perspective.

Acknowledgments

This work has been partially funded by the Spanish grants BIO2007-67150-C03-02, S-Gen-0166/2006, PS-010000-2008-1 and European Union Grant FP7-HEALTH-F4-2008-202047.

References

1. Mardis, E.R. (2009) New strategies and emerging technologies for massively parallel sequencing: applications in medical research. *Genome Med*, 1, 40.
2. Ansorge, W.J. (2009) Next-generation DNA sequencing techniques. *N Biotechnol*, 25, 195-203.
3. Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K. and Wang, J. (2009) SNP detection for massively parallel whole-genome resequencing. *Genome Res*, 19, 1124-32.
4. Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., Guan, J., Fan, D., Weng, Q., Huang, T., Dong, G., Sang, T. and Han, B. (2009) High-throughput genotyping by whole-genome resequencing. *Genome Res*, 19, 1068-76.
5. Garcia de la Nava, J., Santaella, D.F., Cuenca Alba, J., Maria Carazo, J., Trelles, O. and Pascual-Montano, A. (2003) Engene: the processing and exploratory analysis of gene expression data. *Bioinformatics*, 19, 657-8.
6. Mejia-Roa, E., Carmona-Saez, P., Nogales, R., Vicente, C., Vazquez, M., Yang, X.Y., Garcia, C., Tirado, F. and Pascual-Montano, A. (2008) bioNMF: a web-based tool for nonnegative matrix factorization in biology. *Nucleic Acids Res*, 36, W523-8.

7. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Muetter, R.N. and Edgar, R. (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res*, 37, D885-90.
8. Nogales-Cadenas, R., Carmona-Saez, P., Vazquez, M., Vicente, C., Yang, X., Tirado, F., Carazo, J.M. and Pascual-Montano, A. (2009) GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Res*, 37, W317-22.
9. Abascal, F., Carmona-Saez, P., Carazo, J.M. and Pascual-Montano, A. (2008) ChIPCodis: mining complex regulatory systems in yeast by concurrent enrichment analysis of chip-on-chip data. *Bioinformatics*, 24, 1208-9.
10. Vazquez, M., Carmona-Saez, P., Nogales-Cadenas, R., Chagoyen, M., Tirado, F., Carazo, J.M. and Pascual-Montano, A. (2009) SENT: semantic features in text. *Nucleic Acids Res*, 37, W153-9.
11. Neves, M.L., Carazo, J.M. and Pascual-Montano, A. (2010) Moara: a Java library for extracting and normalizing gene and protein mentions. *BMC Bioinformatics*, 11, 157.
12. Nogales-Cadenas, R., Abascal, F., Diez-Perez, J., Carazo, J.M. and Pascual-Montano, A. (2009) CentrosomeDB: a human centrosomal proteins database. *Nucleic Acids Res*, 37, D175-80.
13. Vialas, V., Nogales-Cadenas, R., Nombela, C., Pascual-Montano, A. and Gil, C. (2009) Proteopathogen, a protein database for studying *Candida albicans*-host interaction. *Proteomics*, 9, 4664-8.
14. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25, 25-9.
15. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. and Yamanishi, Y. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res*, 36, D480-4.
16. Khatri, P. and Draghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21, 3587-95.
17. Dopazo, J. (2006) Functional interpretation of microarray experiments. *OMICS*, 10, 398-410.
18. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102, 15545-50.
19. Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, 37, 1-13.
20. Carmona-Saez, P., Chagoyen, M., Tirado, F., Carazo, J.M. and Pascual-Montano, A. (2007) GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol*, 8, R3.
21. Carmona-Saez, P., Chagoyen, M., Rodriguez, A., Trelles, O., Carazo, J.M. and Pascual-Montano, A. (2006) Integrated analysis of gene expression by Association Rules Discovery. *BMC Bioinformatics*, 7, 54.
22. Lee, D.D. and Seung, H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, 401, 788-91.
23. Brunet, J.P., Tamayo, P., Golub, T.R. and Mesirov, J.P. (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*, 101, 4164-9.
24. Carmona-Saez, P., Pascual-Marqui, R.D., Tirado, F., Carazo, J.M. and Pascual-Montano, A. (2006) Biclustering of gene expression data by non-smooth non-negative matrix factorization. *BMC Bioinformatics*, 7, 78.
25. Carrasco, D.R., Tonon, G., Huang, Y., Zhang, Y., Sinha, R., Feng, B., Stewart, J.P., Zhan, F., Khatri, D., Protopopova, M., Protopopov, A., Sukhdeo, K., Hanamura, I., Stephens, O., Barlogie, B., Anderson, K.C., Chin, L., Shaughnessy, J.D., Jr., Brennan, C. and Depinho, R.A. (2006) High-resolution genomic profiles define distinct clinico-pathogenetic subgroups of multiple myeloma patients. *Cancer Cell*, 9, 313-25.
26. Wang, G., Kossenkov, A.V. and Ochs, M.F. (2006) LS-NMF: a modified non-negative matrix factorization algorithm utilizing uncertainty estimates. *BMC Bioinformatics*, 7, 175.
27. Kim, P.M. and Tidor, B. (2003) Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res*, 13, 1706-18.

28. Gao, Y. and Church, G. (2005) Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics*, 21, 3970-5.
29. Inamura, K., Fujiwara, T., Hoshida, Y., Isagawa, T., Jones, M.H., Virtanen, C., Shimane, M., Satoh, Y., Okumura, S., Nakagawa, K., Tsuchiya, E., Ishikawa, S., Aburatani, H., Nomura, H. and Ishikawa, Y. (2005) Two subclasses of lung squamous cell carcinoma with different gene expression profiles and prognosis identified by hierarchical clustering and non-negative matrix factorization. *Oncogene*, 24, 7105-13.
30. Kim, H. and Park, H. (2007) Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23, 1495-502.
31. Han, X. (2007) Cancer molecular pattern discovery by subspace consensus kernel classification. *Comput Syst Bioinformatics Conf*, 6, 55-65.
32. Heger, A. and Holm, L. (2003) Sensitive pattern discovery with 'fuzzy' alignments of distantly related proteins. *Bioinformatics*, 19 Suppl 1, i130-7.
33. Pehkonen, P., Wong, G. and Toronen, P. (2005) Theme discovery from gene lists for identification and viewing of multiple functional groups. *BMC Bioinformatics*, 6, 162.
34. Chagoyen, M., Carmona-Saez, P., Shatkay, H., Carazo, J.M. and Pascual-Montano, A. (2006) Discovering semantic features in the literature: a foundation for building functional associations. *BMC Bioinformatics*, 7, 41.
35. Pascual-Montano, A., Carazo, J.M., Kochi, K., Lehmann, D. and Pascual-Marqui, R.D. (2006) Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE Trans Pattern Anal Mach Intell*, 28, 403-15.
36. Shatkay, H. and Feldman, R. (2003) Mining the biomedical literature in the genomic era: An overview. *Journal of Computational Biology*, 10, 821-855.
37. Blaschke, C., Andrade, M.A., Ouzounis, C. and Valencia, A. (1999) Automatic extraction of biological information from scientific text: protein-protein interactions. *Proc Int Conf Intell Syst Mol Biol*, 1999, 60-67.
38. Jenssen, T.K., Lægreid, A., Komorowski, J. and Hovig, E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28, 21-28.
39. Wren, J.D. and Garner, H.R. (2004) Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. *Bioinformatics*, 20, 191-198.
40. Hoffmann, R. and Valencia, A. (2005) Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, 21.
41. Chaussabel, D. and Sher, A. (2002) Mining microarray expression data by literature profiling. *Genome Biology*, 3, 1-0055.
42. Jelier, R., Jenster, G., Dorssers, L.C.J., van der Eijk, C.C., van Mulligen, E.M., Mons, B. and Kors, J.A. (2005) Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. *Bioinformatics*, 21, 2049-2058.
43. Raychaudhuri, S., Schütze, H. and Altman, R.B. (2002) Using Text Analysis to Identify Functionally Coherent Gene Groups. *Genome Research*, 12, 1582-1590.
44. Huang, W. and Marth, G. (2008) EagleView: a genome assembly viewer for next-generation sequencing technologies. *Genome Res*, 18, 1538-43.
45. Frijters, R., Heupers, B., van Beek, P., Bouwhuis, M., van Schaik, R., de Vlieg, J., Polman, J. and Alkema, W. (2008) CoPub: a literature-based keyword enrichment tool for microarray data analysis. *Nucleic Acids Research*, 36, W406.
46. Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S.A., Haggarty, S.J., Clemons, P.A., Wei, R., Carr, S.A., Lander, E.S. and Golub, T.R. (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313, 1929-35.
47. Girolami, M. and Breitling, R. (2004) Biologically valid linear factor models of gene expression. *Bioinformatics*, 20, 3021-33.

Expanding BioPAX format by integrating Gene Regulation



Irma Martínez-Flores, Verónica Jiménez-Jacinto, Alejandra C. López-Fuentes and Julio Colado-Vides

Programa de Genómica Computacional, Centro de Ciencias Genómicas-UNAM. Cuernavaca, Morelos, México

Biological information

Pathways

Knowing the complete sequencing of the genomes of organisms makes it possible to identify their genetic participants. However, the activity in an organism depends not only on the functionality of genes but also on relations and regulatory mechanisms established among them. The cell functions have a series of interactions, which often form pathways that can be grouped together for either organizational or functional reasons. These pathways assemble multiple biological data types, including metabolic pathways, signal transduction pathways, protein-protein interaction networks, gene regulatory pathways and genetic interactions.

Transcription regulation in *Escherichia coli* (genetic pathway)

Transcription is the first step leading to gene expression; therefore, transcriptional regulation is one of the most important regulatory mechanisms. This pathway describes how genes could be expressed, and transcription factors (TFs) as one of their main participants. The role of these

factors is to coordinate and regulate the expression of the genes of an organism. The processes occur with the participation and interaction of different regulatory elements.

A chromosome contains several regions, which include the following:

- genes involved in producing a polypeptide chain or stable RNA;
- promoters that are comprised by a transcription start site (TSS) and short conserved sequences upstream of the TSS, which mediate the binding of RNA polymerase. Each promoter is specific to a particular sigma factor;
- binding sites (BS) for TFs; these sites are recognized by the transcription factors within a genome;
- terminators, which are required regions to finish the transcription process.

In addition to these DNA segments, there are other types of molecules acting in genetic pathways, including:

- proteins, which perform a large number of regulatory functions (RNA polymerase, sigma factors (σ), TFs, etc.);
- small molecules (such as phosphorus, cAMP, iron, etc.);
- RNAs, such as mRNA, tRNA, rRNA and sRNA.

The interactions among all the elements above mentioned participate in genetic regulation pathways.

The organization of genes in a genomic context allows making regulation process more efficient [1]; for instance, operons express coordinately a set of genes of a single promoter in bacteria [2]. However, some cases are complex because they may contain several promoters, out of which some can be internal. Each promoter related to an operon produces a single mRNA, this structure is known as transcriptional unit (TU) [3].

Regulon DB

RegulonDB contains information regarding regulatory elements and the events involving them (<http://regulondb.ccg.unam.mx>). This curated information is high quality, complete and updated [4]. The last version includes a much more precise definition of biological concepts; all the elements we have already mentioned are curated in this database.

The events involving these elements include the complexes formed between TFs and effectors generating specific conformations, which can be either active or inactive. The function of the TFs depends on the binding to DNA-BS under the control of a specific promoter. The action of integrating all these events is called regulatory interaction (RI).

Also, RegulonDB is linked directly to a number of bioinformatic tools that facilitate the analysis of data sets and tools for microarray, as well as direct access to full papers supporting all this knowledge. In summary, RegulonDB represents a "gold standard" in the bioinformatics of gene regulation design bacterial genomics [4].

Biological Databases and Bioinformatics tools

Needs and difficulties

The number of biological databases is growing fast. Nevertheless, the information contained in these databases has distinct data models, access methods, file formats and semantics. This diversity of implementation makes it extremely difficult to collect data from multiple sources, and therefore slows down the scientific research that involves pathways [5, 6]. This generates the need to develop a format for biological pathway data exchange.

Several standard exchange formats have been developed in order to make heterogeneous data sources easier to use and share. The Systems Biology Markup Language (SBML) [7] and CellML [8] represent mathematical models of pathways designed for quantitative simulation of concentration profiles of components over time. The Proteomics Standards Initiative's Molecular Interaction (PSI-MI) format enables to exchange molecular interaction data sets [5]. Finally, the Biological Pathway Exchange (BioPAX) format enables the exchange of biological pathways in general [9].

The aim of the BioPAX project is to develop a standard data exchange of biological pathways based on XML. The project is structured in several levels. The first level provides an exchange language of metabolic pathways; the second level allows the access to molecular interaction databases. The development of the third level has been initiated and will include signal transduction and regulatory networks. We are currently

working in collaboration with the BioPAX team, specifically in the expansion of the ontology related to transcriptional regulation of bacteria.

The aim of this work is:

- to collaborate with the BioPAX team to expand the scope of BioPAX;
- to extend the format to support gene regulation information;
- to establish a relationship so that data contained in RegulonDB can be translated into the BioPAX format;
- to implement a process to generate a BioPAX file containing the full data of RegulonDB.

Methodology

The process of translating data from RegulonDB into the BioPAX format can be divided in four general activities:

- understanding the BioPAX format. To make a correct translation of the data is an important step to understand the BioPAX format in terms of structure, concepts and definitions; and by means of this, the representation of knowledge;
- indentifying entities in RegulonDB. As we mentioned earlier, RegulonDB contains the complete detailed information of the elements of the transcriptional regulatory network of *E. coli*, thus, the RegulonDB physical entities involved in this process need to be identified;
- expanding the format with transcriptional regulation and adaptation of RegulonDB entities into BioPAX. This activity consists on mapping RegulonDB into BioPAX. During the mapping process, it was found that there was not a right place or class to map some RegulonDB data; it was therefore necessary to create or expand a number of BioPAX classes;
- automation of the translation process. A computer program was developed and implemented in order to translate the RegulonDB data into the BioPAX format.

Implementation

a) Understanding the BioPAX ontology

Considering that BioPAX is an ontology based format, it is necessary to understand the generic components of ontologies and how they represent biological knowledge.

Ontology components

Individuals: also known as instances of a class, they represent objects in the domain we are interested in [10]. For example: the *araC* gene.

Properties: the characteristics of individuals that can also express relations with others individuals. For example: sequence, binds to.

Classes: they can be interpreted as sets that contain individuals. They are described in formal ways that state precisely the requirements for membership to the class. For example, the Protein class would contain all the individuals that are proteins within our domain of interest [10].

In the BioPAX format, the root class for all the interacting components in the ontology is called "Entity" and represents a discrete biological unit. The entities include pathways, interactions and physical entities. The class that serves to represent entities with a physical structure is called "Physical Entity" and contains several subclasses.

An example of a class, its properties and an individual belonging to that class is presented below.

A class:

Class	Protein
Properties	Name
	Cellular Location
	Comment
	References

Individual of this class:

Name	AraC
Cellular Location	Cytoplasm
Comment	Molecular Weight: 33.384 Isoelectric point: 6.95
References	Hendrickson W., et al., 1990.

b) Identifying entities in RegulonDB

We identified RegulonDB physical entities involved in the transcriptional regulation process to establish a relation (mapping) between BioPAX and RegulonDB (Fig.1).

c) Expanding the format that supports transcriptional regulation and adaptation of RegulonDB entities into BioPAX

The mapping process consisted on matching the relationships of each element of the regu-

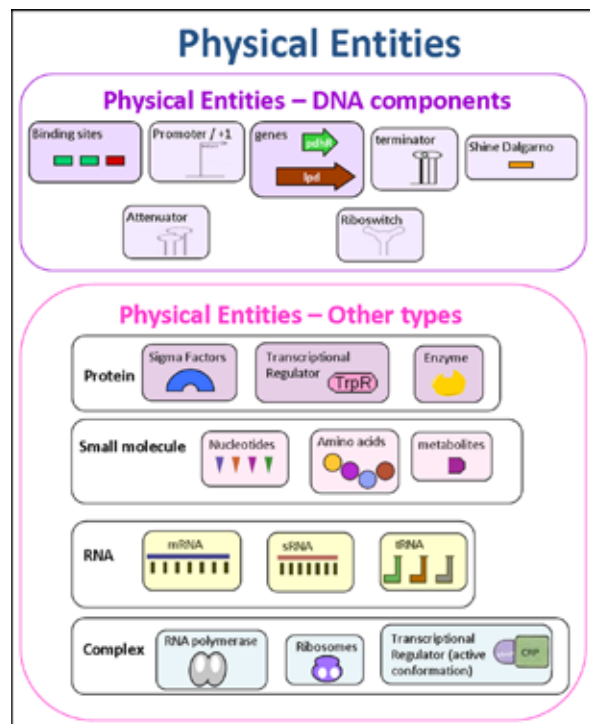


Figure 1. RegulonDB physical entities to mapping into the BioPAX format

latory network with the overall framework of the BioPAX ontology.

This was achieved by using the description of BioPAX classes to relate each RegulonDB entity with one or more classes. However, as BioPAX did not have enough support to represent genetic regulation, there was no place to map some data from RegulonDB. We collaborated with the BioPAX team to include details of the genetic regulation specifically; as a result, the format was improved, obtaining a better representation of it. This also allowed us to create rules to validate errors or warnings that will be useful in a future in terms of validating gene regulation behavior. Thus, classes, such as Template Reaction Regulation, Template Reaction, Dna Region, Dna Region Reference, Rna Region and Rna Region Reference were created and/or modified to have a suitable representation of the genetic information (Fig. 2).

The final relationship (mapping) between the RegulonDB and BioPAX is shown on Table 1.

This table contains an overview of the actual mapping; although, we also created a detailed document describing the specific properties of the involved elements. This document was

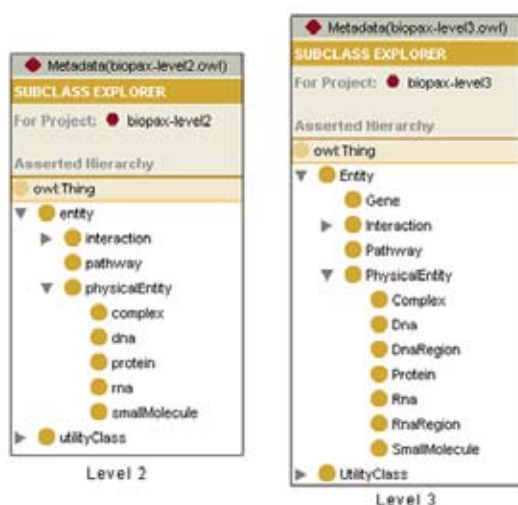


Figure.2. Example of changes on the BioPAX ontology.

used as a guide or template to translate all the RegulonDB data to BioPAX [11].

d) Automation of the translation process

RegulonDB contains a huge amount of data; therefore, translation must be an automated process. A computer program was implemented in Java; the algorithm used to translate the data is, in general terms, the following: load and open the BioPAX ontology; connect to RegulonDB; create individuals from each table. The class these individuals belong to in BioPAX is determined by a mapping that uses the document described above. For instance: there are 4000 genes stored in the RegulonDB table Gene, along with their characteristics. The program creates 4000 individuals and places them with all their properties in the BioPAX DnaRegion class, following the relation established in the mapping document.

Finally, when the program completes the translation, a file containing all the RegulonDB data represented in BioPAX format is generated (<http://regulondb.ccg.unam.mx/download/RegulonDBSignIn.jsp>); this file can be visualized using an ontology editor such as Protégé or a similar one.

Figure 3 shows the process described above.

Conclusions

We have collaborated with the BioPAX team to include details of genetic regulation to generate a complete and robust format. This expanded schema proved to be adequate to contain transcriptional regulation information, since we could successfully translate our data into the standard

RegulonDB	BioPAX
Object External DB Link	Unification Xref
Publication	Publication Xref
Evidence	Evidence and Evidence Code
Gene	Dna Region and Dna Region Reference
Product	Rna Region and Rna Region Reference
	Protein and Protein Reference
	Cellular Vocabulary
Gene Product Link	Template Reaction
Site	Dna Region and Dna Region Reference
Promoter	Dna Region and Dna Region Reference
Promoter Feature	Dna Region and Dna Region Reference
Terminator	Dna Region and Dna Region Reference
Effector	Small Molecule
Conformation	Complex
Transcription Unit	Dna Region and Dna Region Reference
Regulatory Interaction	Template Reaction Regulation
Attenuator	Dna Region and Dna Region Reference
	Dna Region and Dna Region Reference
Attenuator Terminator	Dna Region and Dna Region Reference
	Dna Region and Dna Region Reference
Shine Dalgarno	Dna Region and Dna Region Reference
Rfam	Dna Region and Dna Region Reference
Motif	Dna Region and Dna Region Reference
Operon	Pathway

Table 1. Relation between RegulonDB and BioPAX

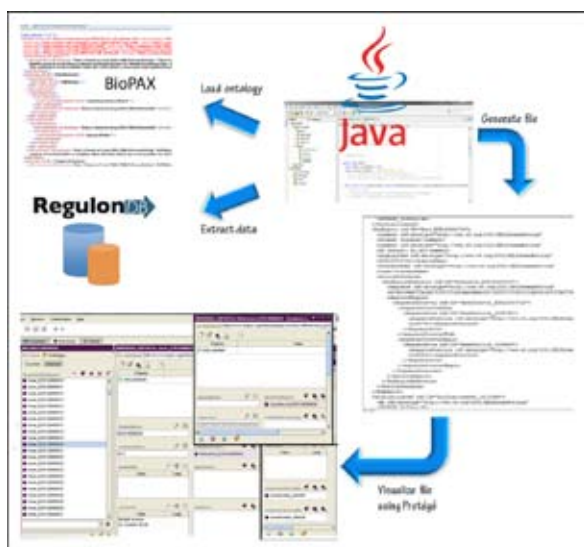


Figure 3. Diagram showing the process of mapping

format; the BioPAX file containing the full version of RegulonDB can be downloaded from the RegulonDB Web page at the main menu in Downloads/Full Version option.

RegulonDB data into the BioPAX format gives a consistent format to the database, which also strengthens this bioinformatics platform. Therefore, it will facilitate the process of sharing information with other databases, as well as making RegulonDB compatible with other standards and allow the addition of new types of data.

Acknowledgements

We would like to thank Gary D. Bader for his support in biological discussions; to Emek Demir and Luis José Muñiz for their invaluable computational assistance, and to Victor Del Moral and Romualdo Zayas for their technical support.

This work was supported by the National Institutes of Health grant GM071962.

References

1. Jacob F, Monod J: Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 1961, 3:318-356.
2. Patrick R. Murray MAP: *Microbiología Médica Genética bacteriana*, 5 edn. España: Elsevier; 2006.
3. Pierce BA: *Genetics*. In: *A conceptual approach*. Edited by Company WHFa, 2nd edition edn; 2005.
4. Collado-Vides J, Salgado H, Morett E, Gama-Castro S, Jimenez-Jacinto V, Martinez-Flores I,

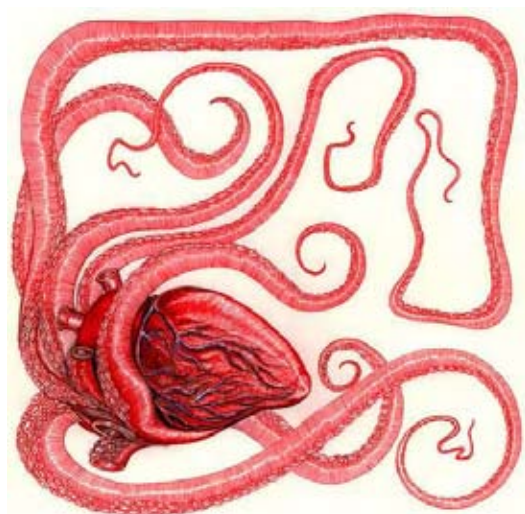
Medina-Rivera A, Muniz-Rascado L, Peralta-Gil M, Santos-Zavaleta A: Bioinformatics resources for the study of gene regulation in bacteria. *J Bacteriol* 2009, 191(1):23-31.

5. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C et al: The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. *Nat Biotechnol* 2004, 22(2):177-183.
6. Buetow KH: Cyberinfrastructure: empowering a "third way" in biomedical research. *Science* 2005, 308(5723):821-824.
7. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A et al: The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 2003, 19(4):524-531.
8. Lloyd CM, Halstead MD, Nielsen PF: CellML: its future, present and past. *Prog Biophys Mol Biol* 2004, 85(2-3):433-450.
9. BioPAX: Biological pathway exchange. [<http://www.biopax.org>]
10. The Protégé Ontology Editor and Knowledge Acquisition System [<http://protege.stanford.edu/>]
11. Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Penaloza-Spinola MI, Contreras-Moreira B, Segura-Salazar J, Muniz-Rascado L, Martinez-Flores I, Salgado H et al: RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res* 2008, 36(Database issue):D120-124.

love, love, love...

Vivienne Baillie Gerritsen

You need two humans for romantic love. That sounds straightforward enough. But you also need chemistry, as in chemical processes. It is an uncomfortable thought in a society where freewill is all the rage. Yet any of our feelings need a basis on which to work upon. And that is our brain with all its neuronal circuits and neurotransmitters that are being continuously fired from one neuron to another, sending messages of fright, anguish, enthusiasm, sadness, despair, love and surprise to name but a few. So what would be the chemistry at the heart of romantic love? Serotonin. Perhaps... With a notion as ungraspable as love, it is a very tricky business to try and pin it down to the makings of one molecule. Yet that is what a team of scientists tried to do. Their research hypothesis is particularly intriguing: they compared the infatuation we all experience in the early stages of love with a form of obsessive-compulsive behaviour.



Absence makes the heart..., by Ben Lawson

Courtesy of the artist

There are few feelings more beneficial to a human than those triggered off by love. But love was not given to us for therapeutic purposes. From a purely biological point of view, where poetry has little room, we fall in love for a reason. Indeed, falling in love means falling for a mate. Falling for a mate means – to put it bluntly – sexual intercourse. And sexual intercourse means perpetuation of the

sex than infidelity. Hence, according to such a theory, love would be the doings of evolution. You can agree with it. Or disagree with it. But it has its logic.

Once you start meddling with the notion of love and its chemistry, the curious mind wants to know which chemical entity could be actively involved in such a feeling. In this light, scientists compared a human's psychological state during the early stages of romantic love with obsessive-compulsive disorder (OCD). The neurotransmitter serotonin, or to be more precise the protein which carries it and is known as serotonin transporter, has a role in OCD in that its concentration is lower in patients suffering from the psychiatric condition than it is in healthy individuals. When individuals madly in love with someone were tested for the level of their serotonin transporters, the scientists found that – like OCD – their concentration was lower. Could that be where the term “lovesick” comes from?

Besides OCD and romantic love, serotonin and its transporter are known for their involvement in mood and behaviour. Neurons filled with serotonin are found in all parts of the brain – which goes to show their importance. The serotonin system seems to be critical in child brain development and the branching out of serotonergic projections. Later on in life, this particular branching or indeed the specific serotonin transporter polymorph that an individual has inherited can give rise to differences

type of psychiatric disorder following stressful situations, depending on the type of serotonin transporter he or she carries. Scientists even suggest here a basis for a difference in masculine or feminine moods or even psychiatric predisposition.

So the serotonin transporter, small as it is, has a far-reaching role in our lives. But how exactly does it work? Serotonin transporter is an integral membrane protein found in the cell membrane of neurons at the level of the presynaptic terminal, one end of which protrudes into the synaptic cleft. This is the part which grabs free serotonin and flings it back into the neuron ready for a new neurotransmitter cycle. But it needs ions to help it. First Na⁺ binds to the empty transporter in the synaptic cleft. This is the cue for serotonin to bind, followed closely by Cl⁻. The threesome then causes a conformational change in serotonin transporter which flips around bringing the part which is usually immersed in the synaptic cleft into the neuron cytoplasm. There it releases the serotonin molecule, thus replenishing the neuron with its neurotransmitter which is now available to spark off a variety of moods. The binding of intracellular K⁺ then causes the transporter to flip back into its original position with the receiving end in the synaptic cleft; K⁺ is released and serotonin transporter is ready to bind another ligand.

Consequently, it is not difficult to understand that serotonin transporter is crucial in regulating serotonin activity as well as homeostasis, and thus has a pivotal role in the regulation of moods and behaviours – with normality at one end and mental disorders at the other, notably when the level of

serotonin transporters is low. Besides romantic love, the serotonin transporter system is believed to be involved in many other types of behaviours such as appetite, sleep, sex, arousal, addiction, impulsiveness, anxiety, depression, OCD, alcoholism, autism...and even spiritual experiences.

The “romantic love versus OCD” hypothesis has met with scepticism. The individuals chosen for the study were certainly in love but none of them had had sexual intercourse with the ones they had fallen for. This was a prerequisite as the scientists defined romantic love as love where sex had not yet proved to be part of the bargain. This met with controversy. Were all these individuals not just suffering from stress caused by an unsatisfied desire due to repetitive procrastination? A behaviour not so far removed from OCD...

Unsurprisingly, the serotonin system is already a principal site of action of therapeutic antidepressants. Further knowledge of it will provide a greater understanding of the role of serotonin in brain development, the neurocircuits involved in emotional processing and, perhaps more importantly, the basis of a number of neuropsychiatric disorders which could then be the basis for the design of novel psychiatric drugs. So can romantic love really all be brought down to chemistry? And to this chemistry in particular? It will probably always remain a mystery. Is it not daunting, though, to take on board the fact that molecules can have such power over our emotions? And yet, without chemistry, we know that feelings would not exist. Talk about chemistry between people...

Cross-references to UniProt

Sodium-dependent serotonin transporter, *Homo sapiens* (Human) : P31645

References

1. Marazziti D., Akiskal H.S., Rossi A., Cassano G.B.
Alteration of the platelet serotonin transporter in romantic love
Psychological Medicine 29:741-745(1999)
PMID: 10405096
2. Tek C.
Correspondence : To the Editor.
Psychological Medicine 30 :241-242(2000)
PMID : 10722194
3. Nordquist N., Oreland L.
Serotonin, genetic variability, behaviour, and psychiatric disorders – a review
Upsala Journal of Medical Sciences 115:2-10(2010)
PMID: 20187845

We are now accepting manuscripts for review and publication and look forward to receiving your contributions and feedback.

Publication schedule for 2011:

17.1 - Next Generation Sequencing Data Analysis

Last submission date: 15 February 2011
Publishing date: 15 May 2011

17.2 - Metagenomics, metatranscriptomics and Biodiversity

Last submission date: 15 May 2011
Publishing date: 15 July 2011

17.3 - Networks and Connectomics

TBA in 17.1 issue

17.4 - Pharmacogenomics

TBA in 17.2 issue

National Nodes

Argentina

IBBM, Facultad de Cs. Exactas, Universidad Nacional de La Plata

Australia

RMC Gunn Building B19, University of Sydney, Sydney

Belgium

BEN ULB Campus Plaine CP 257, Brussels

Brazil

Lab. Nacional de Computação Científica, Lab. de Bioinformática, Petrópolis, Rio de Janeiro

Chile

Centre for Biochemical Engineering and Biotechnology (CIByB), University of Chile, Santiago

China

Centre of Bioinformatics, Peking University, Beijing

Colombia

Instituto de Biotecnología, Universidad Nacional de Colombia, Edificio Manuel Ancizar, Bogota

Costa Rica

University of Costa Rica (UCR), School of Medicine, Department of Pharmacology and Clinic Toxicology, San Jose

Cuba

Centro de Ingeniería Genética y Biotecnología, La Habana

Finland

CSC, Espoo

France

ReNaBi, French bioinformatics platforms network

Greece

Biomedical Research Foundation of the Academy of Athens, Athens

Hungary

Agricultural Biotechnology Center, Godollo

India

Centre for DNA Fingerprinting and Diagnostics (CDFD), Hyderabad

Italy

CNR - Institute for Biomedical Technologies, Bioinformatics and Genomic Group, Bari

Mexico

Nodo Nacional de Bioinformática, EMBnet México, Centro de Ciencias Genómicas, UNAM, Cuernavaca, Morelos

The Netherlands

Dept. of Genome Informatics, Wageningen UR

Norway

The Norwegian EMBnet Node, The Biotechnology Centre of Oslo

Pakistan

COMSATS Institute of Information Technology, Chak Shahzaad, Islamabad

Poland

Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warszawa

Portugal

Instituto Gulbenkian de Ciencia, Centro Portugues de Bioinformatica, Oeiras

Russia

Biocomputing Group, Belozersky Institute, Moscow

Slovakia

Institute of Molecular Biology, Slovak Academy of Science, Bratislava

South Africa

SANBI, University of the Western Cape, Bellville

Spain

EMBnet/CNB, Centro Nacional de Biotecnología, Madrid

Sri Lanka

Institute of Biochemistry, Molecular Biology and Biotechnology, University of Colombo, Colombo

Sweden

Uppsala Biomedical Centre, Computing Department, Uppsala

Switzerland

Swiss Institute of Bioinformatics, Lausanne

Specialist- and Assoc. Nodes

CASPUR

Rome, Italy

EBI

EBI Embl Outstation, Hinxton, Cambridge, UK

Nile University

Giza, Egypt

ETI

Amsterdam, The Netherlands

ICGEB

International Centre for Genetic Engineering and Biotechnology, Trieste, Italy

IHP

Institute of Health and Consumer Protection, Ispra, Italy

ILRI/BECA

International Livestock Research Institute, Nairobi, Kenya

MIPS

Muenchen, Germany

UMBER

Faculty of Life Sciences, The University of Manchester, UK

CPGR

Centre for Proteomic and Genomic Research, Cape Town, South Africa

The New South Wales Systems

Biology Initiative
Sydney, Australia

for more information visit our Web site
www.EMBnet.org

EMBnet.journal

ISSN 1023-4144

Dear reader,

If you have any comments or suggestions regarding this journal we would be very glad to hear from you. If you have a tip you feel we can publish then please let us know. Before submitting your contribution read the "Instructions for authors" at <http://journal.EMBnet.org/index.php/EMBnetnews/about> and send your manuscript and supplementary files using our on-line submission system at <http://journal.EMBnet.org/index.php/EMBnetnews/about/submissions#onlineSubmissions>.

Past issues are available as PDF files from the Web site:
<http://journal.EMBnet.org/index.php/EMBnetnews/issue/archive>

Publisher:

EMBnet Stichting
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 next issue:
15 february 2011