EMBnet.news

Volume 15 Nr. 4
March 2010

- EMBnet AGM 2009
- High Throughput Sequencing and the IT architecture, Part 1
- Utopia Documents and The Semantic Biochemical Journal Experiment

and more ...

Editorial

Next-generation sequencing technologies continue to revolutionise the Life Science community. Hand-in-hand with these developments, high-throughput genomics and proteomics technologies are both becoming more common worldwide and are evolving at a rapid pace. Sequencing platform facilities are being established in research institutions in all continents, making these new techniques accessible to researchers in a broad range of fields.

The new techniques open up the possibility of completely new forms of collaboration, as scientists in all continents perceive and embrace the imperative to share expertise and resources. Articles in this issue of EMBnet.news bear witness to these new trends.

EMBNet itself is also evolving to reflect these opportunities, and is now a more international network than ever before, covering all continents. To really emphasise EMBnet´s commitment to include and support non-European nodes, the organisation's latest AGM took place in Cancun, Mexico, the first time it has ever been held outside of Europe. A short AGM report is to be found in this issue, and a more extensive conference supplement is under preparation.

Year 2010 will be an exciting one, with new nodes from other continents joining, and new collaborative projects between nodes, with other networks (*e.g.*, ISCB, SanBio) and with scientists worldwide.

A very exciting and perhaps natural consequence of EMBnet's evolution is the announcement that its long-standing, successful publication, EMBnet.news, is soon to include a new peer-reviewed section. We encourage everyone involved with EMBnet and its communities to actively engage with this future, new-look journal, to build on its past success and tradition of being *your* journal, *for* you, *about* you, and providing a *showcase* for *your* work! Join us in this endeavour!

EMBnet.news Editorial Board

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

Contents

Editorial 2
Letters to the Editor
Utopia Documents and The Semantic Biochemical Journal experiment 3
News and Announcements
wEMBOSS and wrappers4EMBOSS 7
BITS 2009 8
Presentation of EMBnet nodes associated in 2009 ..10
Reports
The EMBnet Annual General Meeting 2009 and EMBnet-RIBIO joint conference 14
Argentinian EMBnet node: progress report 24
Brazilian EMBnet Node: progress Report 25
Chilean EMBnet node: progress report 31
Colombian EMBnet node: progress report 33
Greek EMBnet node: progress report 34
Pakistan EMBnet node: progress report 35
Spanish EMBnet node: progress report 36
Sri Lankan EMBnet node: progress report 38
ILRI-BecA, EMBnet specialist node: report 39
UMBER, EMBnet specialist node: report 41
Jornadas de Bioinfomática JB2009 43
Next Generation Sequencing Workshop 44
Technical Notes
High Throughput Sequencing and the IT architecture, Part 1 51
Protein Spotlight 56
Node information 58

Editorial Board:

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

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

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

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

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

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

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

Cover picture: Fatoumata Sissoko, student at 'The Malaria Research and Training Center (MRTC)', Mali distributing the EMBnet.News at the Joint ISCB Africa ASBCB Conference, Bamako, Mali 30 Nov. - 3 Dec. 2009.[© Erik Bongcam-Rudloff]

Utopia Documents and The Semantic Biochemical Journal experiment

Teresa K.Attwood*†, Douglas B.Kell‡§, Philip McDermott*†

James Marsh‡, Steve R.Pettifer* and David Thorne‡

* School of Computer Science
† Faculty of Life Sciences
‡ School of Chemistry, The University of Manchester, UK
§ Manchester Interdisciplinary Biocentre, 131 Princess Street, Manchester, M1 7DN.

Introduction

Recent technological advances have led to the accumulation of data on an unprecedented scale. Adding to this information overload is the advent of desk-top sequencing, with machines capable of delivering terabytes of data per hour. The problem is, the rush to increase the amounts of information we collect does not in itself bestow a miraculous increase in knowledge. For information to be usable, it needs to be stored and organised in ways that allow us to access it, to analyse it, to annotate it and to relate it to other information. Unfortunately, to date, we have failed to store and organise much of the rapidly accumulating information (whether in databases or documents) in rigorous, principled ways, so that finding what we want and understanding what's already known become increasingly exhausting, frustrating and costly experiences.

Scientists for whom these problems have become especially acute are database curators. Today, the largest protein sequence database in the world is UniProtKB [1]. UniProtKB currently contains >9 million entries, of which ~500,000 have been contributed by its manually-annotated component, Swiss-Prot [2]. By inspecting thousands of articles and hundreds of other database entries, it has taken 23 years for the Swiss-Prot curators to annotate about half of these sequences – indeed, Bairoch estimates that this gargantuan task has involved 600 person years of effort [3]! The difficulties faced by the curators are enormous: with ~25,000 peer-reviewed journals publishing ~2.5 million articles per year, this equates to something like two new papers appearing in Medline every minute [4]. Consequently, it is impossible for curators to keep abreast of developments, and more and more difficult for them to find relevant papers, or to locate relevant facts within them. It isn't really surprising, then, that Bairoch should opine, "It is quite depressive to think that we are spending millions in grants for people to perform experiments, produce new knowledge, hide this knowledge in a often badly written text and then spend some more millions trying to second guess what the authors really did and found" [3].

The life of curators would be a lot easier if articles could become better conduits to their underlying research data. In fact, it has already been argued that the distinction between an online paper and a database is already diminishing [5]. Nevertheless, it is clear that much more needs to be done to make the data contained in research articles more accessible.

In this Letter, we briefly outline a new development with Portland Press Ltd., the so-called Semantic *Biochemical Journal* experiment [6]. Behind this 'experiment' is a new software tool, Utopia Documents, which builds on the Utopia suite described in previous EMBnet.news articles, and elsewhere [e.g., 7-9]. Here, we provide a sketch of these new developments, in order to provide a taster of what can be achieved through academic-journal-publisher collaboration.

The "experiment"

Utopia is a software suite that semantically integrates visualisation and data-analysis tools: its most recent component, Utopia Documents, brings document-reading and document-

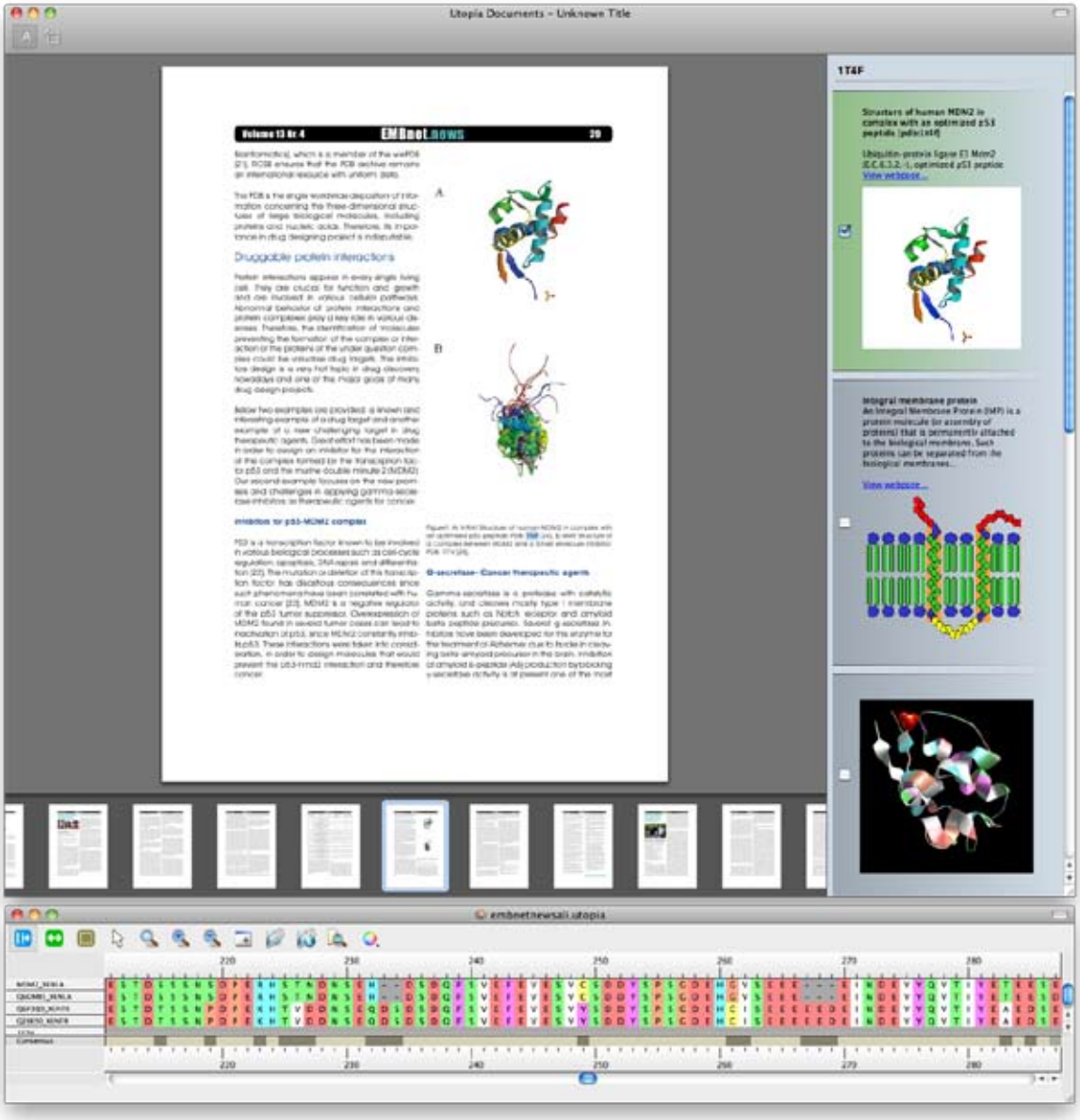

Figure 1. Composite screen-shot illustrating some of the features of Utopia Documents. Dominating the Figure is a page from EMBnet.news (volume 13, issue 4, page 29). Clicking on terms of interest in the text (e.g., 1T4F - the highlighted term half-way down the second column of text) provides definitions from various online databases, dictionaries or thesauri, and uses RDF-linked data to infer and retrieve related information. Here the reader has accumulated information about the 1T4F molecule from the PDB, a definition of 'membrane protein' from Dbpedia, and an interactive visualisation of 1T4F using Utopia's Ambrosia molecular viewer, as well as a related protein alignment, viewed using the CINEMA alignment editor. Hence, from a single page in an article, access is gained to information from databases, from online dictionaries and encyclopaedias and to interactive analysis tools, without having to leave the context of the PDF document.

management utilities to the suite. The aim of the *Semantic Biochemical* Journal (BJ) experiment was to use Utopia Documents to make the content of BJ electronic publications and supplemental data richer and more accessible. To achieve this, Utopia was integrated with in-house editorial and document-management workflows, allowing the BJ editors to mark up article content prior to publication.

The Utopia Document PDF-reader creates unique fingerprints of document contents as they are rendered onscreen, identifying key typographical and bibliometric features (authors, references, etc.). Its innovation lies in being able to turn static features of a document into objects that can be linked, annotated, visualised and analysed interactively. In so doing, the document is transformed from a digital facsimile of its printed counterpart into a gateway to related knowledge, providing readers with focused interactive access to analysis tools, external resources and the wider literature.

As part of the experiment, the journal editors have marked up papers in the December 2009 issue of the BJ using Utopia Documents. Aspects of these articles relating to protein sequence and structure analysis have been the main targets for mark-up, in the first instance, because this was the functionality built into the original Utopia toolkit. The kinds of additional mark-up currently provided by the software include: links from the text to external Websites (e.g., to databases like UniProtKB, PDB [10]) and InterPro [11]; term definitions from ontologies and controlled vocabularies; extra embedded data and materials (images, videos and the like); and links to interactive tools for sequence alignment and 3D molecular visualisation. Utopia does not itself provide any domain-specific functionality for processing or analysing data, but relies on external Web services – these are accessed via plug-ins whose appearance in the software interface is mediated by a 'semantic core' (which can be customised to any subject area by incorporating the relevant discipline-specific ontologies).

Reliance on external Web services is both a strength and weakness of the system: whereas it allows greater flexibility for customising the functionality of the suite, it also depends on the reliability of the external services it exploits – if these become unavailable (e.g., owing to routine maintenance or some kind of faulty operation), their functionality becomes unavailable to Utopia. These issues afflict *all* systems that rely on Web services, but are mitigated to some extent by the establishment of a Web-service registry, which systematically monitors and provides status reports on its registered services [12].

Future work

Utopia Documents is still at an early stage of development and there is much more work to be done. As the system is readily customisable, we plan to extend its scope, especially to encompass chemical biology – here, in particular, we plan to explore collaborations with the Royal Society of Chemistry, who have done pioneering work with their Prospect software (http://www.rsc.org/Publishing/Journals/ProjectProspect/). We are also embarking on exploratory discussions with Nature and Elsevier, and various pharmaceutical companies, in order to deliver bespoke mark-up and document-management solutions for these companies.

Another possibility that we're keen to explore is the use of Utopia Documents to mark up issues of EMBnet.news, which could add value at a critical time, as EMBnet.news becomes a *bona fide* peer reviewed publication – see Figure 1.

Find out more

The Semantic Biochemical Journal was formally launched on 10 December 2009. To gain further insights into the status of the project, and to better appreciate how this might benefit EMBnet.news in future, we encourage readers to view articles in volume 424(3) of the BJ (http://www.biochemj.org/bj/424/3/default.htm?S=0), and especially to read the launch article, Calling International Rescue: knowledge lost in literature and data landslide! (http://www.biochemj.org/bj/424/0317/4240317.pdf) - a preview of this interactive paper is given in the following video: http://www.youtube.com/watch?v=rIOgARI1a3E. Utopia Documents itself is available for download from http://getutopia.com/.

Funding

Utopia Documents has been funded by the European Union (EMBRACE, grant LHSG-CT-2004-512092), the Engineering and Physical Sciences Research Council (Doctoral Training Account), the Biotechnology and Biological Sciences Research Council (Target practice, grant BBE0160651), and

Portland Press Limited (The Semantic Biochemical Journal project).

References

1. The UniProt Consortium (2009) The Universal Protein Resource (UniProt). Nucleic Acids Res 37: D169-D174.
2. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31: 365-370.
3. Bairoch A (2009) The future of annotation/biocuration. Nature Precedings doi:10.1038/npre.2009.3092.1.
4. Hull D, Pettifer SR, Kell DB (2008) Defrosting the Digital Library: Bibliographic Tools for the Next Generation Web. PLoS Comput Biol 4: e1000204.
5. Shotton D, Portwin K, Klyne G, Miles A (2009) Adventures in semantic publishing: exemplar semantic enhancements of a research article. PLoS Comput Biol 5: e1000361. Doi:pcbi.1000361.
6. Attwood TK, Kell DB, McDermott P, Marsh J, Pettifer SR, Thorne D (2009) Calling International Rescue – knowledge lost in literature and data landslide! Biochem J 242: 317-333.
7. Pettifer, S., Attwood, T.K., McDermott, P., Sinnott, J. and Thorne, D. (2007) UTOPIA: User-friendly Tools for OPerating Informatics Applications. EMBnet.news 13: 19-24.
8. Sinnott JR, Pettifer SR, Attwood TK (2004) Introduction to the CINEMA5 sequence alignment editor. EMBnet.news 10(3).
9. Pettifer, S., Thorne, D., McDermott, P., Marsh, J., Villeger, A., Kell, D.B. & Attwood, T.K. (2009) Visualising biological data: a semantic approach to tool and database integration. BMC Bioinformatics 10: S19.
10. Kouranov A, Xie L, de la Cruz J, Chen L, Westbrook J, Bourne PE, Berman HM (2006) The RCSB PDB information portal for structural genomics. Nucleic Acids Res 34: D302–D305.
11. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn R, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJA, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2009) InterPro: the integrative protein signature database. Nucleic Acids Res 37: D211-D215.
12. Pettifer S, Thorne D, McDermott P, Attwood T, Baran J, Bryne JC, Hupponen T, Mowbray D, Vriend G (2009) An active registry for bioinformatics web services. Bioinformatics 25: 2090 - 2091.t

wEMBOSS and wrappers4EMBOSS

Volunteers asked to join the maintenance and development team

wEMBOSS is a Web interface for EMBOSS that provides the user with a personal data space server-side, where he can manage his data and analysis results in projects [1]. The companion package wrappers4EMBOSS allows to integrate under EMBOSS 3th party software like BLAST, fastA and CLUSTAL and to use MRS as sequence databank access method.

he wEMBOSS+wrappers4EMBOSS development team consisted of Martin Sarachu from the Argentinian EMBnet Node, and Marc Colet and Guy Bottu from the Belgian EMBnet Node. Unfortunately, Martin Sarachu tragically died from cancer on 9 September 2007. At the Belgian side Marc Colet retired as professor at the University of Brussels on 1 October 2008 and Guy Bottu is not working for BEN anymore since 1 May 2009, because the Belgian Federal Science Policy Office decided to cut the financing of BEN. Marc is not willing to continue working on the project and Guy is unsure whether he will continue to have time to spend. Marc and Guy however do not want to see this useful software lost for the bioinformatics community.

Therefore they are searching volunteers to take over the development. The idea is that the new member(s) should join the project at SourceForge and start working on wEMBOSS. To make the takeover easy Marc has recently released a version of wEMBOSS maintained as project under Eclipse and with an as much as possible cleaned up and debugged code. During the first six months Marc Colet will remain available for advice, but will ultimately leave the team. Guy will keep working on wrappers4EMBOSS as long as possible, but in the medium term the new member(s) should also collaborate on/take over the development of wrappers4EMBOSS.

For any further details please contact Marc Colet (marcrcolet@gmail.com).

References

Martín Sarachu and Marc Colet. wEMBOSS: a web interface for EMBOSS. Bioinformatics 2005, 21(4):540-541.

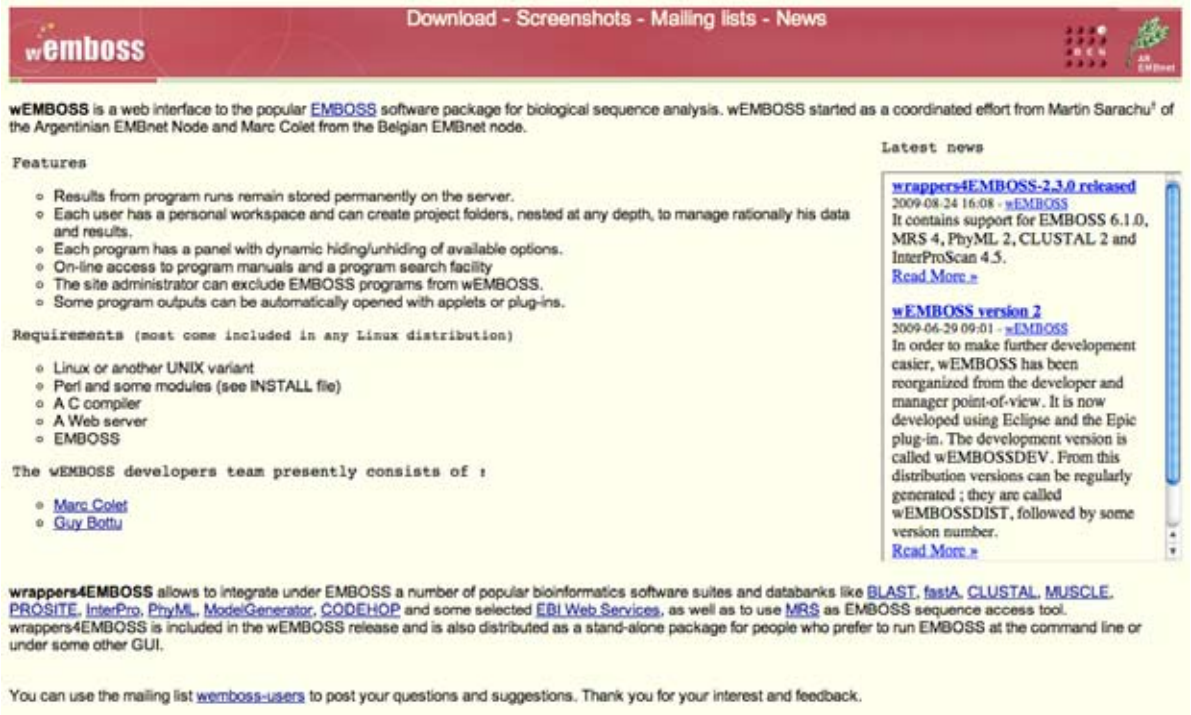

Figure 1. http://wemboss.sourceforge.net/.

BITS 2009

"Bioinformatics and Computational Biology for Life Sciences research"

Seventh Annual General Meeting of the Italian Bioinformatics Society

14-16 April 2010, Bari (IT)

http://bits2010.ba.itb.cnr.it/

Announcement and Call for Abstracts

The Italian EMBnet node in collaboration with the Italian Bioinformatics Society (BITS) and the Department of Biochemistry and Molecular Biology of the Bari University are pleased to announce the "*Seventh Annual General Meeting of the Italian Bioinformatics Society* (BITS 2010)" to be held April 14-16, 2010 in Bari, Italy.

The BITS 2010 event will highlight cutting-edge advances in all major topics of Bioinformatics and Computational Biology. Major aim of the conference is to provide an overview of Italian bioinformatics research and an international forum for in-depth assessment of the challenges involved in the fast moving field of Bioinformatics for molecular biology research. Scientific topics covered will include many areas of interest such as metagenomics, system biology, structural and functional genomics, trascriptomics and proteomics. A particular focus will be on tools and strategies for the storage, management and analysis of data produced by next generation sequencing platforms and other high-throughput techniques.

The program includes keynote lectures, oral presentations, poster sessions and two tutorials: 'Phylogenetic analysis using LIBI HPC facilities using mixed models' and 'Metagenomics and metadata analysis'. Tutorials will include introductory seminars and hands-on sessions. Participation is free of charge and limited to 25 persons amongst the first one registering to the conference.

Invited speakers

Prof. Erik Bongcam-Rudloff,	University of Uppsala, Uppsala, Sweden
Prof. Cecilia Saccone,	University of Bari and CNR,Bari, Italy
Prof. Edward N. Trifonov,	University of Haifa, Israel and Masaryk University, Czech Republic
Prof. Eske Willerslev,	University of Copenhagen, Copenhagen, Denmark

Abstracts submission

Deadline: March 15, 2010

Submitted contributions should address novel bioinformatics and computational biology methods, algorithms, databases, tools and applications for research and development in one or more of the following domains:

- Genomics
- Molecular Evolution and Comparative Genomics
- Protein structure and function
- Proteomics
- Transcriptomics
- Metagenomics
- New tools for NGS
- Systems Biology
- Biological Databases and Biobanks

This list is by no means exclusive of any further topics in the field of Bioinformatics.

Travel grants available for young researchers

Young researchers (up to 35 years old) without a permanent position and submitting an abstract for participating to the conference can apply for a travel grant. Deadline for application is March 15, 2010.

For more information about the conference, registration, sponsorship, exhibition opportunities and travel grants, please, visit the conference web site at http://bits2010.ba.itb.cnr.it

Presentation of EMBnet nodes associated in 2009

Bioinformatics at Nile University: Perspectives for Egypt and the whole Region

Mohamed Abouelhoda, Moustafa Ghanem

Center for Informatics Sciences, Nile University, Smart Village, Giza, Egypt

Nile University is a non-profit research university established in 2007 as part of the Egyptian Government's Strategic Plan aiming at building-up an IT society in Egypt. The Center for Informatics Sciences (CIS) at Nile University (www.cis.nileu.edu.eg) was established in January 2008 as a research center dedicated to the development and application of informatics methods for the management and interpretation of scientific information. The center's research agenda focuses on the areas of health, agriculture, and environment. Currently, CIS has 40 researchers working in six groups including the Bioinformatics one. The research direction and coordination of the groups is led by Dr. Moustafa Ghanem.

Bioinformatics at NU

The Bioinformatics Group at Nile University is leading an effort to promote the field of Bioinformatics in Egypt and the Arab region. The group was established in early 2008 under the leadership of Dr. Mohamed Abouelhoda. International scientific advisors include Professors Robert Giegerich (Bielefeld University, Germany) and Hani Gabra (Imperial College London). The group is committed to the following objectives: 1) building a national bioinformatics computational infrastructure, 2) running fundamental and applied research with focus on food and healthcare applications, and 3) conducting an ambitious human capacity building program in bioinformatics in response to the growing molecular biology research base in Egypt and the region.

In 2009, the pilot phase of the bioinformatics infrastructure plan at Nile University was completed with installation of a cluster of 21 servers with 160 cores and total 1TB RAM with 24 TB total Storage. Our infrastructure is well connected to the internet and expandable by design to include other computational resources in collaborating sites through Grid technology, which renders it ideal for local needs and for international collaboration.

NUBIOS (Nile University Bioinformatics Server) is a web-based portal running over the existing computational infrastructure. The server was developed as part of an internally-funded research project conducted in collaboration with the Department of Computing at Imperial College London in the UK and the Practical Computer Science Department at Bielefeld University in Germany. NUBIOS hosts public and local biological databases. It also provides access through web and programmatic interfaces to a wide variety of popular bioinformatics tools and novel tools developed by Nile University. Beneficiaries

Figure 1. Nile University.

of the infrastructure and NUBIOS services include, among others, Animal Health Institute, Agriculture Genetic Engineering Institute, and National Cancer Institute in Egypt as well as researchers in the Medical School at Imperial College London in the UK.

The group is currently engaged in a number of research projects. Over the past eighteen months, a major focus has been on building a cancer bioinformatics facility for the management, integration, and interpretation of Bladder and Ovarian Cancer data. This effort is conducted in collaboration with the Department of Oncology, Imperial College London, the Egyptian National Cancer Institute, and InforSense Ltd. Other projects for the application of advanced data management and analysis methods to pathogen detection and virus research are in preparation with partners from Europe, USA and Australia.

The group has well established research collaborations with international software companies. In 2008, it successfully completed a pilot project with Microsoft for building bioinformatics tools running on Microsoft Windows Cluster and released WinBioinfTools, an open source package for sequence analysis running under Windows HPC Server 2008. The group is also collaborating with IBM for developing tools designed to run on massively parallel architectures and the IBM Blue Gene technology.

Developing a strong base of competent bioinformaticians in Egypt and the region is a self-commitment of the group. Its efforts in this area include a dedicated bioinformatics stream in the Nile University Masters program and regular engagements in a number of regional and international activities. For example, the head of the group, Abouelhoda, is currently mentoring the ISCB Regional Student Group of North Africa.

The bioinformatics group is optimistic that EMBnet membership will help it in achieving its goals. Through EMBnet, wider international and regional collaboration is guaranteed and wider dissemination of the group work will be accelerated. Nile University will act as a dissemination node for all research and educational information of EMBnet to the Egyptian and regional community by means of its local and regional connections and by means of NUBIOS, which will host a local mirror of the EMBnet site.

The New South Wales Systems Biology Initiative

Bruno A Gaëta[1], Marc R Wilkins[2]

The University of New South Wales, Sydney NSW, Australia

[1] School of Computer Science and Engineering

[2] School of Biotechnology and Biomolecular Sciences

The New South Wales Systems Biology Initiative (SBI) was established in 2008, funded by the New South Wales Office for Science and Medical Research (OSMR) and the University of New South Wales (UNSW). The mission of SBI is to become Australia's foremost centre for Systems Biology, undertaking basic and applied research in the development and application of bioinformatics for genomics and proteomics, and providing collaborative bioinformatics services in these areas.

The SBI is based in the school of Biotechnology and Biomolecular Sciences at the University of

Figure 1. Screenshot of GEOMI, a network visualisation tool developed by Dr. Seokhee Hong (National ICT Australia). This version of GEOMI has been specifically tailored by the SBI to allow co-visualisation of protein- protein interaction networks or complexome networks with multiple protein parameters (e.g. abundance, half-life, function, localization) or gene expression data.

Figure 2. Screenshot of the Interactorium, a program developed by the SBI for visualisation of interaction networks. The Interactorium is based on Skyrails, a platform for 3D visualisation of complex networks developed by Yose Widjaja and Tim Lambert (school of computer science and engineering, UNSW). This screenshot shows the yeast Pure complex (a mitochondrial protein complex).

New South Wales in Sydney, Australia. The initiative is closely associated with the nearby Ramaciotti Centre for Gene Function Analysis, which provides microarray and sequencing services in New South Wales, and the Biomolecular Mass Spectroscopy Analysis. The SBI is also a member of Bioplatforms Australia, the national consortium of "omics"-related facilities.

The director of the SBI is Professor Marc Wilkins, who holds the chair of Systems Biology at UNSW. The centre currently employs one post-doctoral fellow and two bioinformatics research scientists, in addition to a number of postgraduate students and affiliated researchers in several departments of UNSW and in other universities.

The SBI provides collaborative bioinformatics services and expertise to users of genomic and proteomics facilities in Australia, in the context of the National Collaborative Research Infrastructure Scheme (NCRIS). The SBI has strong expertise in the areas of experimental design and infrastructure for genomics, transcriptomics and proteomics, and in integrative and network biology, particularly with regard to visualisation and analysis of biomolecular interaction networks. These services are provided free of charge to non-profit research organisations in the state of New South Wales. An ongoing project is the application of GEOMI, a platform for network visualisation, to the analysis of protein interaction networks in various contexts including gene expression, disease and post-translational modifications.

The SBI is also collaborating with Intersect Pty Ltd, a not-for-profit e-research support organisation, to implement computational infrastructure for the management, storage and processing of next generation sequencing data which will be made available to users in NSW.

Education and training are part of the mission of the SBI, and the centre has already run a number of workshops and symposia since its foundation, including a workshop on the future of bioinformatics featuring Dr. Ewan Birney from EBI, an advanced proteomics data analysis workshop, and a symposium on next generation sequencing applications held jointly with Sydney Bioinformatics. UNSW offers an undergraduate Bachelor of Engineering in Bioinformatics program (the largest bioinformatics undergraduate program in Australia) that several staff and affiliates of the centre contribute to. The director of the program, Bruno Gaëta, is affiliated with the SBI, and the centre has hosted a number of students for their final year research thesis project.

Information about the NSW Systems Biology Initiative can be accessed on its website at http://www.systemsbiology.org.au/, together with reports, publication lists and software developed by the centre.

Centre for Proteomic & Genomic Research: World Class Biotech Made in Africa

Reinhard Hiller and Judit Kumuthini
Institute for Infectious Diseases and Molecular Medicine, Faculty of Health Science, UCT
Anzio Road, Observatory, South Africa

The Centre for Proteomic and Genomic Research (CPGR) was created in 2006 as part of a government initiative to provide scientists in South Africa with state-of-the-art analytical services, technical expertise, project support and collaborative research capabilities in the genomics and proteomics arena. The organization has a particular interest in translational research and advancing scientific findings from the bench to the market, including a focus on tackling pressing health needs in Africa as well as on improving crops and livestock. The integrated core technology facility was founded as a not-for-profit organization through a grant provided by the Department of Science and Technology (DST) by way of its investment vehicles the Cape Biotech Trust (CBT) and PlantBio (PB).

The CPGR started operations in October 2006 with a vision of establishing a modern, world-class research facility that serves the needs of the scientific community in South Africa by providing state-of-the-art services, technical expertise and collaborative research capabilities in the high throughput genomics and proteomics sectors. More specifically, the CPGR's mission is:

- To facilitate high quality science in the fields of genomics and proteomics in South Africa through collaborative research initiatives with academia & industry;
- To enable growth of existing southern African biotech companies into new areas of commercial opportunity through provision of high-quality value-adding services;
- To stimulate new biotech activity through conversion of cutting-edge research into novel intellectual property and new products;
- To create new commercial ventures by translating available opportunity, expertise and activity into spin-offs or stand-alone companies in the areas of molecular diagnostics, biomarker discovery, bio-prospecting, animal health and plant genetics, amongst others; and
- To increase the knowledge-base and the number of suitably trained graduate, post-graduate and postdoctoral scientists in the biotech sector.

Becoming an associated node partnership status at EMBnet falls neatly under our mandate as a platform that provides specialist support and services to the academic community in (South) Africa. With the recent addition of a strong Bioinformatics unit to our state-of-the-art Genomics & Proteomics organization, we have created a leading one-stop-shop core facility that has the capacity to support numerous scientific projects and partners, locally and internationally.

The EMBnet Annual General Meeting 2009 and EMBnet-RIBIO joint conference

José R. Valverde

EMBnet/CNB, Centro Nacional de Biotecnología, CSIC. Madrid, Spain

Abstract

The 2009 Annual General Meeting (AGM) of EMBnet took place in the Mayan Riviera, Mexico, during the last week of October 2009 (http://www.EMBnet.org/en/EMBnet-RIBio2009/).This meeting was held together with the AGM of the Iberoamerican Bioinformatics Network, both sharing and participating equally in the associated Scientific meeting, which this year was centred on High Throughput Technologies and Systems Biology. Besides producing a sound scientific event, the 2009 AGMs of both Organisations have generated a wealth of exciting new initiatives that we all hope shall produce a major impact in Bioinformatics worldwide, such as the foundation of the new Iberoamerican Society of Bioinformatics (SOIBIO) and the agreement among major networks, societies and organizations from all over the world to launch an unprecedented coordination initiative.

Introduction

The Annual General Meeting of EMBnet for year 2009 (AGM09) took place at the Hotel Paradisus Riviera Cancun Resort, in the Mayan Riviera, Puerto Morelos, Quintana Roo, México, from October 26th to October 29th, 2009 and was organized in cooperation with the Iberoamerican Bioinformatics Network (RIBIO) which was also holding its AGM. ISCB provided sponsoring for the event in the form of travel funds for students/postdocs.

This has been the first time that EMBnet AGM has taken place outside Europe and marks a landmark in EMBnet activities, underscoring the compromise of EMBnet with addressing global Bioinformatics needs and helping promote and develop the discipline all over the world. This commitment was further supported by EMBnet invitation of representatives of major organizations, such as APBionet, ASCB, EMBL, ISCB, RIBIO and SANBio to attend the meeting.

Figure 1. Hotel Paradisus Riviera Cancun Resort: the reception area.

The local organizers, and chairs of the meeting, Dr. Julio Collado-Vides, Cesar Bonavides-Martínez, both from the Mexican EMBnet node at CCG, Cuernavaca, Mexico, went to considerable efforts to make this a most comfortable and welcoming environment. They not only ensured that attendants had all their needs met, but also went to great pains to provide excellent support for the scientific activities planned around the AGMs of both networks, including a modernly equipped meeting room, poster areas, and the supply of both type I and type II coffee breaks, and an attractive non-scientific program – this included a visit to Xcaret and an unforgettable ethnic party around Deads' Day, where we enjoyed Mexican delicacies, partied and honoured our beloved deceased ones (Martin Sarachu and Julián Esquivel) in the most emotional Mexican style.

Organization of the event was managed through a joint international organizing committee composed of members of both convocating networks, and including Erik Bongcam-Rudloff, Domenica D'Elia, Oscar Grau, Lubos Klucar, Enrique Morett, Lucía López Bojórquez, Oswaldo Trelles, José R. Valverde and Ana-Tereza Vasconcelos.

As is usual, the AGMs of both Organisations were associated with relevant scientific events. This year, the scientific program opened with a RIBIO conference and an Integrative Workshop

Figure 2. The conference room.

on Bioinformatics and Systems Biology, chaired by M. Defrance, A. Medina and S. Sandoval.

The driving line for this year's scientific conference was "Bioinformatics for High Throughput Technologies and the Interface of Bioinformatics and Systems Biology". It was organized through the joint effort of a large international scientific committee composed of people from both EMBnet and RIBIO and spanning all the continents: T.K.Attwood, E. Barreto, E. Bongcam-Rudloff, S. Chohan, D. D'Elia, J. de las Rivas, N.-E. Erikson, A. Gisel, D. Holmes, L. Klucar, G. Magklaras, A. Orozco, E. Pérez Rueda, O. Trelles, E. Vallejo, J. R. Valverde and A. T. Vasconcelos.

The scientific committee took care of issuing the calls for papers, inviting guest speakers and selecting papers for their presentation, either orally or during the poster sessions, following the recommendations received from a stringent peer review process that made heavy use of electronic support tools offered by EMBnet.

In retrospect, the overall quality of the conference was excellent and the coverage of the topics addressed was -if not fully comprehensive- wide enough to give all participants a broad overview of the current state of the art, of emerging trends and future directions, and most importantly, spurred us to consider many new ideas, and to foster warm expectations and suggestions for cooperation at the subsequent EMBnet-RIBIO joint discussion session.

In addition to these events, this year's meeting was intended to address major issues in the current organization of Bioinformatics around the world. To this end, we invited representatives of major organizations to a joint strategic meeting, where we agreed on common strategies to be developed in the coming years that will hopefully help bring closer all the major communities in Bioinformatics and foster further cooperation. In addition to the mentioned events, this year meeting was intended to address major issues in the current organization of Bioinformatics around the world. To this end, we invited representatives of major organizations to a joint strategic meeting were we agreed on common strategies forward to be developed in the coming years that will hopefully help bring closer all the major communities in Bioinformatics and foster further cooperation.

A special EMBnet.news issue dedicated to the conference will be released soon. Here we provide just some pics which express the good feeling of all us enjoining this "expectacular" meeting!

Coordination activities

The fortunate outcome of the meeting of many excellence scientists for the business and coordination activities of EMBnet and RIBIO AGMs is an opportunity for promoting scientific exchange and cooperation that cannot be missed.

This year, we took advantage of this exceptional opportunity to take our collaboration one step further: we all felt that this was a great opportunity to increase our ties, not only among EMBnet and RIBIO, but also with other existing organizations (like ISCB or APBionet) and emerging networks (like African bioinformatics initiatives). To this end, we invited key people from these organizations to attend our meeting and participate in a special joint discussion session: here, we addressed the opportunities for collaboration and mutual support among all the organizations, possibly leading to a future, joint enterprises that may foster communications and cooperation at an unprecedented level, spanning the various geographic locations, professional interests and organizational approaches to coordination of Bioinformatics.

We are happy to report that this session was a success, where all participants agreed on pursuing this common coordination interest and to promote the enterprise within their respective organizations; we look forward eagerly to future developments in this area.

EMBnet Annual General Meeting

The proper business meeting of EMBnet was carried out at the end of the scientific events. This meeting was consciously kept short, achieving efficiency by only addressing topics that required direct member participation, leaving all other coordination and planning work for the monthly virtual general meetings we have been maintaining through videoconferencing for several years now.

Below, we share some news of general interest that emerged from this business meeting.

The AGM opened with an introduction by the chairman, Erik Bongcam-Rudloff and proceeded to the Welcome words by the local organisers, Cesar Bonavides and Julio Collado-Vides.

Figure 3. Cesar showing Erik how to manage difficult tasks.

The EMBnet constituency

The secretary, Laurent Falquet, called all member nodes to signify their presence. The assembly effectively consisted of 15 national nodes, 2 special nodes and 9 proxies (26 valid votes).

Present [voting right]:
Argentina, Brazil, Colombia, Finland, Greece, Italy, Mexico, Norway, Pakistan, Slovakia, Spain, Sweden, Switzerland, IRLI-BECA – Kenya, IHCP/GMO.

Represented:
Australia (proxy to Nils Einar-Eriksson), China (proxy to Spain), Costa Rica (proxy to Spain), Hungary (proxy to Sweden), ICGEB (proxy to Argentina), Portugal (proxy to Italy), Russia (proxy to Switzerland), South Africa (proxy to Kenya), UMBER (proxy to Sweden).

Absent with apologies:
Belgium, Chile, France

Absent:
Canada, Cuba, India, Netherlands, Poland, MIPS, ETI, EBI

Observers or members of EMBnet committees, with no voting right:
Andreas Gisel (Italy), Nils-Einar Eriksson (TMPC chairman – Sweden), Jaufeerally-Fakim, Yasmina (Mauritius representative), Ezekiel Adebiyi (Covenant University, Nigeria)

Designation of the Election Committee

Andreas (IT) and Cesar (MX) were designated for the election committee

Approval of the Minutes of previous Business Meeting in Martina-Franca - Italy, September 2008

Approved without any objection.

Financial statement

The financial situation of EMBnet remains similar to previous years, with a limited budget for promoting activities. To this end, the treasurer made some recommendations that might help reduce expenses in benefit of the overall budget, and the point was raised that we need to be more proactive in ensuring appropriate processing of membership fees.

New applications for EMBnet membership

This year we received a formal application for a new EMBnet node from Nile University, Egypt, which was introduced by Erik Bongcam-Rudloff; an expression of interest to move the Australian node to South Wales University, Australia, introduced by Bruno Gaeta; and an expression of interest to apply for membership from CGPR, South Africa, which was also introduced by Erik Bongcam-Rudloff. All of these applications were accepted, and we want to express our appreciation for their interest and support, and our most warm welcome to them all. There is also an expression of interest in becoming an associated node from Covenant University, Nigeria until they obtain approval to become the national node, but formal paperwork was not complete before the AGM.

The constituency was reminded that the Associated Node status is a solution to become a member between AGMs and without official government support. It is also less expensive for those entities that have funding issues.

Confirmation of continued activity from temporarily inactive partners

The issue of active participation and how to deal with inactive partners in EMBnet was discussed next, bringing to the table the fact that there are several nodes that are very silent (ETI, EBI, NL, CA, MIPS) but still pay their fees, and that furthermore, there are some other nodes have not paid for the last 3 years… (ICGEB, CU). These points are accompanied by bad news from Switzerland, Belgium and Australia, which are experiencing serious trouble to remain in EMBnet.

Switzerland will temporarily remain in a dormant status. In Belgium, the current node is being phased out and there is activity to define an appropriate alternative to assume its functions. Australia is also phasing out its current node and there is a proposal to switch it over to South Wales University.

Figure 4. EMBnet at the ethno party.

Reports

Following this discussion, nodes up for re-election were asked to produce activity reports introducing the work done at each of them during the period since the last review. Below, a summary is reported for each node.

Complete reports by Argentina, Brazil, Chile, Colombia, Greece, Pakistan, Spain, Sri Lanka, ILRI-BECA, UMBER and the SBI at South Wales University, are included in this issue as separate articles.

Argentina

Since the last review, a LiveCD with bioinformatic tools and wEMBOSS was produced by the late Martin Sarachu. The node's hardware has been upgraded to hold new services like MRS or Gold Sting. Argentina has acted as coordinator for the Iberoamerican Bioinformatics Network, organizing conferences and meetings and contributing to organize the previous joint conference EMBnet-RIBIO in Spain, 2007. The services available for users were described and span a wide

Figure 5. EMBnet-RIBIO ladies community!

array of tools for sequence analysis and assembly and macromolecular structure analysis, as well as related databases.

Australia

The University of Sydney has decided to close down the national bioinformatics services they have been delivering in the last 19 years for a reorganization, and plan on resurrecting the Centre in three years time. A proposal was brought to move the node to the Systems Biology Initiative at South Wales University, led by Bruno Gaeta and its inclusion as specialist node in the interim. Bruno is also assuming relevant roles in APBionet and ISCB.

The constituency expressed its sadness for missing Sonia Cattley and our most warm farewell and best wishes in her new assignments.

Brazil

The facilities available at the Brazilian node were described at large, with detailed descriptions of hardware and software services. The Brazilian node is distributed among various cooperating institutions.

The Computational Genomics Unit Darcy Fontoura de Almeida, led by Ana Tereza Ribeiro de Vasconcelos is associated to LNCC and has the purpose of integrating high throughput sequencing and bioinformatics analysis in a single centre. This unit is a national centre of excellence and reference in HTS with access to advanced laboratory equipment. The node has participated in numerous financed projects, launching new Bioinformatics networks and initiatives in the region. They have been producing copious publications and several courses on Genomics and Bioinformatics in addition to service delivery.

Fundaçao Oswaldo Cruz, FIOCRUZ, described its support, staff, services and activities, centered on proteomics and molecular modelling, and including hands on courses, online training and post-graduate programs, as well as research, development and dissemination activities with potential economic impact.

EMBRAPA, with its laboratory for Computational Biology has a long record doing research, training, development and delivering advanced protein sequence and structure analysis tools like Sting. They are active in running courses on computational biology and structural bioinformatics, developing databases and software, participating in projects and producing publications through extensive cooperation with other groups in the region.

Chile

In the last years, their activity has strengthened in Systems Biology and Bioinformatics with an impact in the local community, organizing Mathematical Modelling courses, obtaining major grants and engaging on a large number of collaborations bridging Mathematics, Life and Health Sciences, Proteomics and Biotechnology.

Colombia

CNIN provides bioinformatics tools, genomics databases, training and support for the Colombian research community. They have been developing in-house some tools like SINCO, BLA.id, ENKI or BLEE addressing topics like molecular identification, pharmaceutics or taxonomy for biodiversity. Besides organizing 6 courses, they have also organized workshops on Sequence Analysis, General Bioinformatics and Microarrays.

Costa Rica

Already reported last year, although an update was submitted this year, during which they have consolidated the Master's course on Medical and Systems Biology Bioinformatics, which is estimated to be ready for launching on 2010 and would become the first academic certification in Bioinformatics in Central America. Besides this, they have been active in dissemination, research, collaborations and the constitution of the Iberoamerican Bioinformatics Society.

Cuba

No report was received.

Greece
BRFAA delivered a short address presenting the activities organizing courses, participating in EMBnet committees and organizing workshops on topics like Bioinformatics, Statistics, Proteomics, Data Mining, Sequence Analysis, doing R+D and getting actively involved in core EMBnet initiatives like project proposals.

Hungary
The Hungarian node is currently undergoing a major overhaul as it is being moved to a new location, which has precluded Endre Barta from submitting a more detailed report in time,

India
No report was available

Mexico
A short oral presentation addressed the enhancements and work developed in Mexico, including involvement in Grid initiatives, dissemination efforts to bring Bioinformatics closer to society, large research initiatives, etc... and, of course, organization of the current AGM.

Netherlands
No report was available

Norway
In a short oral presentation, the Norwegian EMBnet platform was described, with the services it delivers, both public and to registered users, including access to bioinformatics tools, statistical systems (R+bioconductor), DBMS, etc. They also reported about training activities (centered on EMBOSS, PERL and MRS), participation in the organization of the HTS workshop to be celebrated in November and on their active participation in the TMPC of EMBnet. The scientific output obtained by the node was also enumerated as well as their future plans on EMBOSS and MRS.

Pakistan
Their report was produced in a short oral presentation where they reviewed their strong efforts to bring up Bioinformatics in Pakistan, their facilities and major achievements like the institution of a large program for the exchange of students, which is now yielding excellent results thanks to the support of other EMBnet nodes, organization of regular workshops for CIIT Biosciences faculty, active participation in reforming existing curricula, developing online resources (BIREC, http://www.birec.org) and organizing meetings.

Figure.6. EMBnet-RIBIO gentlemen community!

Poland
No report was available

Portugal
Pedro Fernandes sent a short note stating he is currently suffering health problems and apologizing for not being able to be more active at the moment, although he maintains his interest in being active on EMBnet committees.

South Africa (SANBI)
No report available.

Spain
The Spanish node has been active within EMBnet committees, organizing courses on Bioinformatics, Biostatistics, Molecular Dynamics, Quantum Biology and UNIX. It organized the 2007 AGM and has participated in many external courses and workshops both local and international on a variety of topics, and has been pursuing cooperation roads with other organizations, such as RIBIO and APBionet. It has also recently applied as coordinator for a new network on Free Software for the Life Sciences.

Sri Lanka
Activities were quickly summarized, including their strong efforts in organizing training and cooperation with other EMBnet nodes.

MIPS
No report was available.

ILRI-BECA
Their mission is to provide tools, database and services to east and central Africa as a centralized platform. They have been actively involved in research activities, workshop and conference organization, developing and testing the eBioUSB and eBioMacKit developed in Sweden, as well as

Figure 7. Mexican folk singers at the ethno party.

training and capacity building through courses on Bioinformatics, Perl, R, Proteomics, etc.., and dissemination and coordination initiatives to develop a Regional Student Group associated with ISCB.

EBI

No report was available.

UMBER

As a Specialist node its main role is in R+D, where they have been highly involved in many projects like EMBRACE, IMPACT, EuroKUP, etc... They continue providing support for PRINTS, CADRE the DbBrowser web server, UTOPIA and MINOTAUR data analysis tools. In addition they have joined the International Society for Biocuration, and have established a relationship with Portland Press.

Sweden

Although the Swedish node was not up for re-election, Erik reported their activity during this last year. The Swedish node is actively participating in many excellence projects like EMBRACE and ELIXIR, and has been highly collaborative with other nodes to promote joint activities, like Sri Lanka (bioinformatics clustering, conferences, exchanges), Kenya (workshops), China (WebLab), Pakistan (PhD student exchanges), etc.. In addition it has participated in EMBnet committees (TMPC, EB) and is organizing the upcoming NGS workshop in Rome with other institutions.

Voting

The reports were followed by a formal vote about their re-election or discharging from EMBnet. All nodes were re-elected except for Cuba, India, Netherlands, Poland, South Africa and MIPS, which will be up for re-election next year.

EMBnet activity reports

This year activity reports came packed with exciting achievements and news which we would like to summarize here.

Presentation of new candidate nodes

The delegates from candidate nodes were invited to present themselves, their projects and their current status (personnel, resources, expertise area, services, user base, etc.) in not more than 15 min.

Erik Bongcam gave the presentation for Egypt and CPGR (South Africa) and Bruno Gaeta for SW University. All of them were accepted and we want to express our appreciation for their interest and support and our most warm welcome to them all.

EMBnet activity reports

This year activity reports came packed with exciting achievements and news which we would like to summarize here.

Executive board report

Members and functions: Erik Bongcam-Rudloff (SE), chairman, Laurent Falquet (CH), secretary, Oscar Grau (AR), treasurer, Jose R. Valverde (ES), member.

We held 11 monthly VGMs since the last AGM; the usual schedule is the 2nd Tuesday of the month at 4pm CET. The number of participants can vary from 8 to more than 20. VGM minutes are sent by email to the admin list and deposited on the web site in PDF, usually within one week after the meeting. They can be found on the web site (using your private login): http://www.EMBnet.org/VGM-reports.

Actions by the EB (report by Erik Bongcam-Rudloff)

- EGI LS SSC participation in FP7 ROSCOE proposal
- HealthGrid membership
- Regular virtual meetings connected to the VGMs or when needed
- Support for Sri Lanka Master programme
- Support for the creation of the Mauritius Bioinformatics SanBio node
- Support for the ISCB Africa ASBCB Conference in Mali

- Financial support for the creation of an Intercontinental 'Umbrella' organization with the participation of RIBIO, SanBio, Beca, ASBCB, APBionet and EMBnet.
- Direct contact with the PCs when it was needed, EMBnet.news, etc
- Affiliated EMBnet to ISCB
- Support for PARADIGM COST proposal (which was sadly rejected)
- Help for creation of South American Bioinformatics society
- Support for submission of FreeBIT proposal (an Iberoamerican Free Software network for the Life Sciences).

E&T PC report (report by Matej Stano)

Members and functions: Vassilios Ioannidis (CH), chairman (leaving), Jingchu Luo (CN) treasurer, Sofia Kossida (GR), member, Matej Stano (SK), member.

Brief report by Matej, who indicated that after Vassilios left, all activities stopped. They are currently waiting for the constitution of the new committee, and need a new Chair to lead the work.

Matej mentioned his own developments of the their education system (a portal of practical bioinformatics in Slovakia) and EMBnet page on Slovak Wikipedia.

P&PR PC report (by Domenica D'Elia)

Members and functions: Pedro Fernandes (PT), chairman, Lubos Klukar (SK), secretary, Kimmo Mattila (FI), treasurer, Domenica D'Elia (IT), member. Brief report by Domenica D'Elia.

- Organization of 20th EMBnet conference
- BMC Bioinformatics supplement (2009, Vol. 10, Suppl 6) published as outcome of the Conference organized for the celebration of the 20th anniversary of EMBnet
- Publication of the article "The 20th anniversary of EMBnet: 20 years of bioinformatics for the Life Sciences community" in the BMC Bioinformatics supplement (v10, S6)
- OCS/OCJ systems from Public Knowledge Project are now in full production:
  - conference.EMBnet.org
  - journal.EMBnet.org
- Support in the organization of EMBnet/RIBio International conference & Meeting 2009
- EMBnet.news
  - Development of a new format
  - Addition of new sections
  - 5 issues producedDistribution of printed copies in many places and meetings.
  - Guidelines for authors are now available on the web page
- Others:
  - Addition of EMBnet in Google maps
  - Wikipedia pages
  - A booklet describing EMBnet is deemed necessary
  - New quick guides:
    - gLite 3 released
    - mySQL, and AWK, by Nazim are on the works.

TM PC report (report by Nils-Einar)

Members: Nils-Einar Eriksson (SE), chairman, George Magklaras (NO), secretary, César Bonavides-Martínez (MX), member, Guy Bottu(BE), member, Emil Lundberg (NO), member.

Figure 8: Xcaret. Cancun's most amazing eco park.

At present the TMPC membership is being reconsidered, as Emil Lundberg, has been on leave since 2009-03-01. His functions were those of resident DNS/mail/Mac/UNIX guru. Guy Bottu was the MRS expert, but due to the current refurbishing of the node, Guy writes: '...I am available for giving some advice about subjects I know well, but I cannot engage in doing heavy work'. Nils-Einar Eriksson, is reaching the end of the second 3-year period as a TMPC member (up for re-election), he manages the EMBnet mail lists, takes care of some DNS-issues and does some system surveillance. He is now focusing on Next Generation Sequencing applications. Cesar Bonavides, reaching the end of the first 3-year period as a TMPC member is in charge of the EMBnet.org website, offers technical support for website users, helps other TMPC members when required (updating software or taking care of security issues) and tries to give/put a smile on other EMBnetters whenever he can. George

Magklaras is on his 4th year of its TMPC membership, George is our DNS/UNIX/Linux expert, addressing data security and RDBMS issues with the EMBnet.org website and specializing on data storage systems and large infrastructure (HPC) system administration.

Activities:

- Maintenance of the web site using Drupal
- Update the e-Learning web site
- Development of the backup system (advice about having a mirror system is taken and will be considered further)
- NGS approaches and participation in the organization of the Workshop in Rome next November.
- Work on analysis and deployment of distributed file systems. A summary and recommendation will be produced for the NGS workshop.

EMBnet related projects

Proposal for a committee on external relations

A proposal was raised for a new special committee to be created, dedicated to the creation of the Umbrella organization.

This proposal was accepted with unanimity.

The new committee will take care of contacts with other institutions, and pursuing funding proposals and finding partners. Work with the new Umbrella to be created between the societies and networks present in this AGM: APBionet, SANBio, RIBIO, EMBnet, ASBCB, ISCB.

Future Directions

Electronic voting system: this is a proposal that has been raised several times regarding the adoption of a voting system that could enable us to reach decisions with greater agility. An initial proposal to use the Ballotbin.com system was approved for first try as an interim solution until other options (Drupal) are studied, proposed and implemented by the TMPC (Nazim)

Alternative ways to raise EMBnet funds: this issue was discussed in a parallel session that took place separately in conjunction with representatives of other societies. A decision was taken to explore join ventures for creating an umbrella organization and prepare joint proposals to non-EU organizations and funding agencies.

PC goals for next year

PC goals have already been partially defined during reports and will be refined by new constituencies of each committee and presented in next VGMs with requests for funds needing approval.

New project proposals

Currently several proposals are ongoing and pending submission (e.g. ROSCOE) or evaluation (e.g. FreeBIT).

Figure 9: Bruno Gaeta presenting South Wales University activities and resources

Elections to positions inside the EMBnet Stichting

Elections were preceded by a short introduction where the topic of PCs constitution was addressed. The tradition was to have 4 members per committee, the Statutes or Bylaws do not impose any number – this meeting may decide upon number of members for each committee.

After consideration of vacancies and proposal of new members, candidates were voted resulting in the following changes to the various committees:

- **Executive Board:** Erik Bongcam-Rudloff, Laurent Falquet and Oscar Grau have left the EB. The new EB is now constituted by Terri Attwood, Andreas Gisel, Jose R. Valverde and Etienne de Villiers.
- **Education and Training:** Vassilios Ioannidis and Sonia Catley stepped down, and new members were elected, the committee is composed now of Sophia Kossida, Matej Stano, Jingchu Luo and Shahid Chohan.
- **Publications and Public Relations:** Andreas Gisel steps down to join the EB, and Martin Norling was elected. The committee now is composed of Pedro Fernandes, Domenica D'Elia, Lubos Klucar and Martin Norling.
- **Technical Management (R+D):** Guy Bottu leaves the committee, Nazim Rahman and Harald Dahle were elected to a full committee of Nils-Einar Erikson, Cesar Bonavides, George

Magklaras, Emil Lundberg, Nazim Rahman, Harald Dahle.

- **Foreign Relations:** this new committee is initially composed of Erik Bongcam-Rudloff and will be augmented with other members in future virtual general meetings.

Planning of EMBnet Virtual General Meetings

The monthly schedule of "Virtual General Meetings" (VGM) using the Marratech e-conferencing software proved adequate over the last year. A decision upon the periodicity of these meetings for next year, as well as a schedule of virtual meetings of the PCs was adopted to continue with the usual schedule (every second Tuesday of the month at 4pm CET).

Date and place of next EMBnet AGM meeting

Candidate nodes were welcome to present offers with as many details as possible: venue, date, facilities, connected scientific meeting or collaborative workshop, financial aspects, etc.

The date and place should be determined as soon as possible to organize properly. J. R. Valverde presented a proposal, seconded by Laurent Falquet that we start considering from now on planning AGMs two years in advance.

Matej Stano mentioned that Lubos is willing to organize next AGM. Pedro Fernandes could also do it next year. We had therefore two volunteers, but since none of the managers of these were present, we could not obtain more details.

Both candidates were asked to bring a detailed proposal with budget for the next regular VGM to be held on November 24th so a decision may be made as soon as possible.

Concluding remarks

The chairman thanked everybody for the good work and closed the meeting at 13:3

The AGM closed with an emotive farewell presentation by Laurent Falquet for the exiting chairman, Erik Bongcam-Rudloff.

Acknowledgements

First of all, we want to thank the organizing Networks and the Sponsors for their support, especially ISCB who contributed student/postdoc travel funding.

Figure 10. The AGM closed with an emotive farewell presentation by Laurent Falquet for the exiting chairman, Erik Bongcam-Rudloff.

A big thank you goes to the local organizers, Julio Collado-Vides, Cesar Bonavides-Martínez and Cristina Bojórquez for the huge efforts they made before, during and after the meeting to make this a most successful event. Their warm hospitality and meticulous care for details made this an unforgettable reunion of friends rather than a formal scientific meeting.

Thanks must be given as well to all the members of the Organizing and Scientific Committees, for their strenuous effort to successfully put together a high quality conference, to the session chairs, all the speakers, specially the key note speaker Chris Sander, and all the presenters of posters, whose excellent work has made this a brilliant scientific event.

We also want to thank the members of other networks, societies and organizations that attended the meeting: Chris Sander, Bruno Gaeta, Ezekiel Adebiyi, Yasmina Jaufeerally-Fakim, and many others... as well as to all those who intended to come but could not make it to the meeting due to other commitments, and to all of our colleagues of the former South American Bioinformatics Network (RIBIO), and now Sociedad Iberoamericana de Bioinformática (SOIBIO).

Finally, of course, we want to give a big thank you to all the participants in this meeting who contributed with their work and friendliness to make it a successful event.

Argentinian EMBnet node: progress report

Oscar Grau, Diego Bellante

Instituto de Bioqimica y Biologia Molecular (IBBM)
Faculdad de Ciencias Exactas,
Universidad Nacional de La Plata, Argentina

Since last AGM a Live CD was released by the late Martin Sarachu that was intended both for demonstrations of wEMBOSS and for people that have had network connections and do not have access to a bio-server.

Diego Bellante was hired as administrator of the Node.

Node's hardware has been upgraded with four AMD Athlon XP 2.1Ghz, 1 Gb RAM, Linux (RedHat 7.3) connected through 100 Mbit to a Pentium 4 2.4 Ghz and to a file server dual xeon 3 Ghz, 4 Gb RAM, and a raid controller with 1.5 TBytes SATA disks.

One P4 server with 1 Gb RAM and 750 Gb of storage is holding MRS, EMBOSS and Gold Sting databases.

As Coordinator of the Iberoamerican Network of Bioinformatics from CYTED and with AR.EMBnet node I organized "The Second Iberoamerican meeting of Bioinformatics" that took place in Buenos Aires in December 2006. 27 oral presentations and 12 posters were presented by participants from Chile, Colombia, Méjico, Perú, Spain and Argentina. This meeting was attended by more than 240 people.

Also, in collaboration with José Valverde, the joint EMBnet RIB meeting was organized in Torremolinos, Spain in June 2007. This high level meeting allowed the interaction of 18 latinoamericans with more than 50 europeans.

Programs in AR.EMBnet

- EMBOSS
- wrappers4EMBOSS: Integrates EMBOSS with BLAST, fasta and Clustal
- MRS: sequence retrieval system
- STADEN, PHRED and CONSED
- GOLD STING
- Web services: Apache Web Server, Shell connection SSH, Cluster openMosix, Sun Grid Engine.

Data bases in AR.EMBnet

- Nucleotide (EMBL) = EMBL release + EMBL updates
- PDB
- OMIM
- Unigene
- KEGG Ligand Compound, Enzyme, Glycan, Reaction
- GOA
- GO
- Enzyme
- UniProt (Protein) = SwissProt + TrEMBL
- Taxonomy
- Unigene
- UniUnique
- Interpro

Figure 1. Buenos Aires.

Brazilian EMBnet Node: progress Report

Ana Tereza Vasconcelos[1], Wim M. Degrave[2], Goran Neshich[3]

[1]Laboratório Nacional de Computação, Científica Laboratório de Bioinformática, Quitandinha Petrópolis, Rio de Janeiro (Brazil)
[2]Oswaldo Cruz Institute (IOC), FIOCRUZ, Rio de Janeiro, Brazil
[3]Brazilian Agricultural Research Corporation (EMBRAPA) and UNICAMP's Department of Biology, Campinas, Brazil

Mission

The Brazilian EMBnet node conducts research and development in Bioinformatics and Computational Biology, with emphasis on creating and applying computational and mathematical methods and models for solving biological problems. The Brazilian node is formed by a network of three institutions: The National Laboratory for Scientific Computation (LNCC - Petropolis), the Oswaldo Cruz Foundation (Fiocruz – Rio de Janeiro) and Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA - Campinas). The network maintains and develops databases and tools in bioinformatics and computational biology to supply the needs of thematic networks and national and international collaborative projects, organizes training courses on several levels, and promotes technology and innovation.

National laboratory for scientific computation - LNCC, Laboratory for Bioinformatics

Dr. Ana Tereza Vasconcelos.

Infrastructure

Computational

The LNCC, one of the National Institutes of the Ministry of Science and Technology, has at present the following available computational resources: Sunfire 6800 with 24 processors and 24 Gb of memory, SunFire 3800 with 4 processors and 32 Gb of memory, SGI Challenger with 8 processors and 2 Gb of memory and 2 Sun Enterprise 450, offering a set of tools with modern technology that is up-to-date and ready for the development of applications that demand high levels of computational and scientific resources. The computational resources of the LNCC also include 90 Unix workstations (IBM, Silicon Graphics, Sun and Linux), 350 PCs and 100 printers. The external network of the LNCC is made of two links, one of 34 Mbps (megabits/second) to the POP-Rio de Janeiro of the RNP which is operated by the LNCC in its old headquarters and another of 2 Mbps to the REDERIO. The rate of use of the two links (Rio de Janeiro-Petrópolis) is of approximately 50%. Expansion of the links with the REDERIO aiming at the interconnection with the REMAV-Rio de Janeiro (High-Speed Metropolitan Network) is under study. Two communication servers for dialed access, each having 30 digital lines (total 60 lines) should also be at our disposal and will be located in Petrópolis and in Rio de Janeiro (POP-Rio de Janeiro).

The platform of the internal network of the LNCC is composed at present of 2 CISCO Catalyst switches, model 6509, interconnected to 4 Gbps, interconnecting two clusters of Catalyst switches, model XL-2909, with FEC connections of 800 Mbps in each cluster. The master switches of each cluster have 2 expansion slots available, besides several 10/100 Mbps ports reserved for expansions. In total, the LNCC has approximately 500 10/100 Mbps ports in the clusters of switches interconnecting the workstations of its technical/scientific staff. The cabling is certified and warranted by Lucent Tecnologies for a period of 15 years. The switch connections are made of fiber optics and the links from these to the rooms (stations) are made in category 7 twisted-pair cables.

Genomic

The Computational Genomics Unit Darcy Fontoura de Almeida is associated to the Laboratory of Bioinformatics of the National Laboratory of Scientific Computation – LNCC. This Unit, coordinated by Ana Tereza Ribeiro de Vasconcelos, has the purpose of integrating the activities of high-throughput DNA sequencing and bioinformatics into a single center, thus allowing for the best possible use of the data

generated by the new 454 GS FLX sequencer of Roche. Inaugurated in September 19, 2008, the Computational Genomics Unit is a center of excellence of national reference in high-throughput sequencing. At present, the 454 GS FLX sequencer is the only one in South America that follows all the specifications of the Roche manufacturer. The laboratory can also count with an Agilent Bionalyzer 2100, a Nanodrop 3000 fluorometer, a Genomic Solutions HydroShear, a Qiagen TissueLyser, centrifuges, a Beckman Coulter Z1, Veriti thermocyclers and other support equipment.

Projects with financial support

2008 - Actual: Genômica Computacional e o Seqüenciamento Parcial do Genoma de Trypanosoma Cruzi

Financial support: Fundação Carlos Chagas Filho de Amparo à Pesq. do Estado do Rio de Janeiro-FAPERJ

2008 - Actual: apoio para a manutenção e instalação da unidade multiusuário de genômica computacional

Financial support: Fundação Carlos Chagas Filho de Amparo à Pesq. do Estado do Rio de Janeiro-FAPERJ

2008 - Actual: Rede Sul Americana e Iberoamericana de Bioinformática (Red SurAmericana e Iberoamericana de Bioinformatica)

Financial support: Nacional de Desenvolvimento Científico e Tecnológico-CNPq

2008 - Actual: Rede Nacional de Sequenciamento de DNA - Projeto Genoma Brasileiro: Determinação de Genomas Relevantes para a Saúde Humana

Financial support : Ministério da Ciência e Tecnologia-MCT, Ministério da Saúde-MS

2008 - Actual: Rede Brasileira de Pesquisas sobre o Câncer - RBPC

Financial support Ministério da Saúde-MS e Conselho Nacional de Desenvolvimento Científico e Tecnológico-CNPq

2007 - Actual: Biotecnologia - Insumos para Genômica e Proteômica

Financial support: Conselho Nacional de Desenvolvimento Científico e Tecnológico-CNPq

2007 - Actual: Prospecção de novos genes com potencial biotecnológico

Financial support: Conselho Nacional de Desenvolvimento Científico e Tecnológico-CNPq

2006 - Actual: Estudo multicêntrico para caracterização molecular das hemofilias A e B e determinação do estado de portador de hemofilia no Brasil

Financial support : Ministério da Saúde-MS

2006 - 2008: Brazilian Microbiological Resource Center (BMRC)

Financial support: Conselho Nacional de Desenvolvimento Científico e Tecnológico-CNPq, Empresa Brasileira de Pesquisa Agropecuária-Centro Nac. de Pesq. de Soja-EMBRAPA SOJA

2006 - Actual: CTpedia database

Financial support: Ludwig Institute for Câncer research

2004 - Actual: HAMAP BRAZIL - PAthogenic Proteins Annotation Project

Financial support: Swiss Institute for Bioinformatics

2004 - 2008: Projeto Genomica comparativa de Xylella fastidiosa

Financial support : Ministério da Ciência e Tecnologia-MCT, Universidade de São Paulo-USP

2004 - Actual: Fixadores de Nitrogenio

Financial support: Conselho Nacional de Desenvolvimento Científico e Tecnológico-CNPq, Empresa Brasileira de Pesquisa Agropecuária-Centro Nac. de Pesq. de Soja-EMBRAPA SOJA

Courses

- Genômica funcional de microrganismos patogênicos, 2009.
- Genômica e Bioinformática, 2008.
- Bioinformática I - Banco de dados do ponto de vista biológico , 2007.
- Tópicos Especiais em Genética II -Genômica Comparativa , 2007.
- Análise e Comparação de Genomas - Procariotos , 2006.
- Bioinformática I - Banco de Dados do Ponto de Vista Biológico , 2006.

Publications

1. PINEROGONZALEZ, J, CARRILLOFARNES, O, VASCONCELOS, A, GONZALEZPEREZ, A, VASCONCELOS, A. T. R. Conservation of key members in the course of the evolution of the insulin signaling pathway. Biosystems. , v.95, p.7 - 16, 2009.
2. Almeida, L. G., Sakabe, N. J., deOliveira, A. R., SILVA, M. C. C., MUNDSTEIN, A. S., Cohen, T., Chen, Y.-T., Chua, R., Gurung, S., Gnjatic, S., Jungbluth, A. A., Caballero, O. L., Bairoch, A., Kiesler, E., White, S. L., Simpson, A. J. G., Old, L. J., Camargo, A. A., VASCONCELOS, A. T. R. CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens. Nucleic Acids Research. , v.37, p.D816 - D819, 2009.

3. VEIGA, D. F., DEUS, H. F., C A, VASCONCELOS, A. T. R., ALMEIDA, J. S.DASMiner: discovering and integrating data from DAS sources. BMC Systems Biology. , v.1, p.1 - , 2009.
4. Gonzalez Perez, Abel, Espinosa Angarica, Vladimir, Collado-Vides, Julio, VASCONCELOS, A. T. R.From sequence to dynamics: the effects of transcription factor and polymerase concentration changes on activated and repressed promoters. BMC Molecular Biology. , v.10, p.92 - , 2009.
5. Cristiane C Thompson, Ana Carolina P Vicente, Rangel Souza, VASCONCELOS, A. T. R., Tammi Vesth, Nelson Alves Jr,, Tetsuya Iida, Fabiano L. Thompson Genomic taxonomy of vibrios. BMC Evolutionary Biology (Online). , v.1, p.1 - 10, 2009.
6. Pinto, Fabiana G. S., Chueire, Ligia M. O., Vasconcelos, Ana Tereza R., Nicolás, Marisa F., Almeida, Luiz G. P., Souza, Rangel C., Menna, Pâmela, Barcellos, Fernando G., Megías, Manuel, HUNGRIA, Mariângela Novel genes related to nodulation, secretion systems, and surface structures revealed by a genome draft of Rhizobium tropici strain PRF 81. Functional & Integrative Genomics. , p.1 - 8, 2009.
7. BARRETO, K. S., TORRES, A. R., BARRETO, M. R., VASCONCELOS, A. T. R., ASTOLFO-FILHO, Spartaco, HUNGRIA, Mariângela Diversity in antifungal activity of strains of Chromobacterium violaceum from the Brazilian Amazon.. Journal of Industrial Microbiology and Biotechnology. , v.1, p.10.1007/s10295 - , 2008.
8. Freire, Pablo, Vilela, Marco, Deus, Helena, Kim, Yong-Wan, Koul, Dimpy, Colman, Howard, Aldape, Kenneth D., Bogler, Oliver, Yung, W. K. Alfred, Coombes, Kevin, Mills, Gordon B., Vasconcelos, Ana T., Califano, Andrea, VASCONCELOS, A. T. R., Almeida, Jonas S. Exploratory Analysis of the Copy Number Alterations in Glioblastoma Multiforme. plos one. , v.3, p.e4076 - , 2008.
9. GODOY, L, VASCONCELOS, A. T. R., CHUEIRE, L, SOUZA, R, NICOLAS, M, BARCELLOS, F, HUNGRIA, M Genomic panorama of Bradyrhizobium japonicum CPAC 15, a commercial inoculant strain largely established in Brazilian soils and belonging to the same serogroup as USDA 123. Soil Biology & Biochemistry. , p.1 - 11, 2008.
10. GONZALEZ, A., Gonzalez-Gonzalez E., ESPINOSA, V., VASCONCELOS, A. T. R., COLLADO-VIDES, J. Impact of Transcription Units rearrangement on the evolution of the regulatory network of gamma-proteobacteria. BMC Genomics. , v.9, p.1 - 18, 2008.
11. de Mello Varani, Alessandro, SOUZA, Rangel Celso, Nakaya, Helder I., de Lima, Wanessa Cristina, Paula de Almeida, Luiz Gonzaga, Kitajima, Elliot Watanabe, Chen, Jianchi, Civerolo, Edwin, Vasconcelos, Ana Tereza Ribeiro, Van Sluys, Marie-Anne Origins of the Xylella fastidiosa Prophage-Like Regions and Their Impact in Genome Differentiation. plos one. , v.3, p.e4059 - , 2008.
12. VILELA, M. A., CHOU, I., VINGA, S., VASCONCELOS, A. T. R., Eberhard O. Voit, ALMEIDA, J. S. Parameter optimization in S-system models. BMC systems biology. , v.2, p.1752-0509-2-35 - , 2008.
13. Veiga, Diogo FT, Vicente, Fábio FR, Nicolás, Marisa F, Vasconcelos, Ana Tereza R Predicting transcriptional regulatory interactions with artificial neural networks applied to E. coli multidrug resistance efflux pumps. BMC Microbiology (Online). , v.8, p.101 - , 2008.
14. Espinosa Angarica, Vladimir, Gonzalez Perez, Abel, Vasconcelos, Ana T, Collado-Vides, Julio, Contreras-Moreira, Bruno, VASCONCELOS, A. T. R.Prediction of TF target sites based on atomistic models of protein-DNA complexes. BMC Bioinformatics. , v.9, p.436 - , 2008.
15. Bernardes, Juliana S, Fernandez, Jorge H, Vasconcelos, Ana Tereza R Structural descriptor database: a new tool for sequence based functional site prediction. BMC Bioinformatics. , v.9, p.492 - , 2008.
16. GONZALEZ, J. P., FARNES, O. C., VASCONCELOS, A. T. R., GONZALEZ, A.The impact of the emergence of IRS molecules on the evolution of the insulin signaling pathway. Biosystems. , v.prelo, p.01 - 10, 2008.
17. VILELA, M. A. M., BORGES, Carlos Cristiano, VINGA, S., VASCONCELOS, A. T. R., Santos Helena, Eberhard O. Voit, ALMEIDA, J. S. Automated Smoother for the Numerical Decoupling of Dynamics Models. BMC Bioinformatics. , v.8, p.1 - 8, 2007.
18. ALarCON, F., VASCONCELOS, A. T. R., lucia Yim, ZAHA, Arnaldo Genes involved in cell division in mycoplasmas. Genetics and Molecular Biology. , v.30, p.174 - 181, 2007.
19. VEIGA, D. F., VICENTE, F.F.R., FUENTE A, L., MAIA, M. A. G. M., VASCONCELOS, A. T. R. Genome-wide partialcorrelation analysis of Escherichia coli Microarray Data. Genetics and Molecular Research. , v.6, p.730 - 742, 2007.
20. BAREINBOIM, Elias, VASCONCELOS, A. T. R., SILVA, Jõao Carlos Pereira da Grammatical inference applied to linguistic modeling of biological. RECIIS. , v.2, p.329 - 333, 2007.
21. Souza, Rangel Celso, ALMEIDA, Darcy Fontoura de, ZAHA, Arnaldo, Morais, David Anderson de Lima, Vasconcelos, Ana Tereza Ribeiro de In search of essentiality: Mollicute-specific genes shared by twelve genomes. Genetics and Molecular Biology. , v.30, p.169 - 173, 2007.
22. THOMPSON, F., Bruno Gomez-Gil, VASCONCELOS, A. T. R., Tomoo Sawabe Multilocus sequence analysis reveals that Vibrio harveyi and V. camp-

bellii form distinct species.. Applied Enviromental Microbiology. , v.13, p.4279 - 4285, 2007.

23. Brocchi, Marcelo, Vasconcelos, Ana Tereza Ribeiro de, ZAHA, Arnaldo Restriction-modification systems in Mycoplasma spp. Genetics and Molecular Biology. , v.30, p.236 - 244, 2007.
24. GONZALEZ, A., ESPINOSA, V., VASCONCELOS, A. T. R., COLLADO-VIDES, J. Tractor _ DB (version 2.0): a database of regulatory interactions in. Nucleic Acids Research. , v.35, p.D132 - D136., 2007.

Books chapters

- ALMEIDA, Luiz Gonzaga Paula, de Vasconcelos, A. T. R., MAIA, M. A. G. M. A Simple and Fast Term Selection Procedure for Text Clustering In: Intelligent Text Categorization and Clustering ed.Heidelberg : Springer Berlin /, 2008, v.164, p. 47-64.
- WANDERLEY, M. F. B., SILVA, Jõao Carlos Pereira da, BORGES, Carlos Cristiano, de Vasconcelos, A. T. R. Application of Genetic Algorithms to the Genetic Regulation Problem In: Advances in Bioinformatics and Computational Biology ed.Heidelberg : Springer Berlin /, 2008, v.5167, p. 140-151.
- VARANI, A. M., LIMA, Wanessa Cristima de, MOREIRA, L., OLIVEIRA, M., SOUZA, Rangel Celso, CIVEROLO, E., VASCONCELOS, A. T. R., SLUYS, Marie Anne Van. Common Genes and Genomic Breaks: A Detailed Case Study of the Xylella fastidiosa Genome Backbone and Evolutionary Insights In: Plant Pathogenic Bacteria: Genomics and Molecular Biology ed.Reading : Inglaterra, 2008

Services

- EMBnet node
- Expasy Mirror
- CTdatabase
- Brazilian Microbiological Resource
- Mamibase
- Tractor DB
- Structural Descriptor DataBase
- SABIA – Software for automatic Bacterial Anottation

Fundação oswaldo cruz – FIOCRUZ, Platform for Bioinformatics, and Laboratory for Functional Genomics and Bioinformatics

Dr. Wim Degrave

The EMBnet node activities at Fiocruz are assured by the institutional Bioinformatics Platform, with support from the VPPLR-PDTIS program-RPT4A and by the IOC - Functional Genomics and Bioinformatics Unit and support from the Program for Scientific Computing, and the Fiocruz Network.

Team members:

- Wim M. Degrave
- Antonio Basilio de Miranda
- Thomas Dan Otto (currently at the Sanger Institute, UK)
- Fábio F. Mota - Technologist
- Mark Catanho - PhD student in Cellular and Molecular Biology
- Ana Carolina Guimarães - PhD student in Cellular and Molecular Biology
- Flávio Engelke - Master student in Biomedical Sciences

Activities

Bioinformatics services; support for genomics and proteomics platforms at Fiocruz, genome sequencing projects, software and application development; installation and upgrading of software; construction, implementation and updating of databases; design and maintenance of information services; organization of training courses and on-line training, research projects in comparative genomics, evolutionary biology and genome wide metabolic analysis, drug development in neglected diseases.
The node aims to:

- provide the environment and support in bioinformatics (biological data processing, access to genetic databases, creating and maintaining databases for proteomic analysis) and support for special applications such as molecular modeling, assembly and genome analysis, support for proteomics,
- organize hands-on training courses and on-line training to users, mostly within graduate and post-graduate programs;
- contribute to specific research projects through software and database development;
- disseminate bioinformatics as a tool and as a research and development discipline. The Bioinformatics node contributes to improvement of public health and the development of new technologies and tools;
- generate a potential economic impact, because it contributes to the patentability in research projects and innovations, and has captured external resources for this purpose.

Infrastructure

The main infrastructure of the unit is currently comprised of a dozen of smaller dedicated servers. Two larger servers are to be included in 2010. Fiocruz has an extensive network of fiber optics,

linking several thousands of PCs in the different Institutes that comprise the Foundation, and is connected to the RNP and REDERIO through high speed links. Fiocruz counts with several additional bioinformatics groups performing research and development in fields such as genomics, statistics and epidemiology, molecular modeling, georeferencing, systems biology etc., and counts with post-graduate courses in Computational and Systems Biology.

Special Services offered:

- Bioinformatics databases and applications
- Genome assembly
- Web servers
- General sequence analysis
- Proteome analysis
- Data processing

The most common software packages for sequence assembly and database are available.

Products developed by the team of the platform:

- BioParser * - Analyzer/parser for all varieties of BLAST and FASTA, with support for versions of BLAST with and without gaps.
- SQUID * - Friendly local grid environment for the use of BLAST and FASTA programs.
- GenoMycDB - Database for information related to the genome and proteome of mycobacteria
- REReP - Method to facilitate the assembly of genomes, based on the detection and filtering of seqs. repetitive (applicable to data obtained by the method of Sanger and probably pirosequenciamento)
- AnEnPi - Tool for clustering, similarity search, identification of cases of functional analogy and reconstruction of metabolic pathways.
- ProteinWorldDB - Database indexes of similarity between protein sequences of hundreds of genomes – http://www.proteinworlddb.org

Courses

- Computational analysis of sequence and protein (IOC 26051)
- Origin, Structure and Evolution of prokaryotic genomes (IOC 26052)

Recent new collaborations

- Analysis of the genome of Streptococcus pneumoniae, in collaboration with BioManguinhos (Dr. Marco Medeiros)
- Development of a multiplex PCR for distinguishing species of the genus Wolbachia, Ehrlichia, Rickettsia and Anaplasma, in collaboration with Dr. Agnes Rossi (top Mar/2009 - Ready for testing on bench)
- Analysis of genes of Vibrio mimicus, in collaboration with Dr. Ana Carolina Vicente (top Sep/2009)

Publications

1. Buschiazzo A, Goytia M, Schaeffer F, Degrave W, Shepard W, Grégoire C,Chamond N, Cosson A, Berneman A, Coatnoan N, Alzari PM, Minoprio P. Crystal structure, catalytic mechanism, and mitogenic properties of Trypanosoma cruzi proline racemase. Proc Natl Acad Sci U S A. 2006 Feb 7;103(6):1705-10.
2. Carvalho PC, Fischer JS, Chen EI, Domont GB, Carvalho MG, Degrave WM, Yates JR3rd, Barbosa VC. GO Explorer: A gene-ontology tool to aid in the interpretation of shotgun proteomics data. Proteome Sci. 2009 Feb 24;7:6.
3. Carvalho PC, Carvalho Mda G, Degrave W, Lilla S, De Nucci G, Fonseca R, Spector N, Musacchio J, Domont GB. Differential protein expression patterns obtained by mass spectrometry can aid in the diagnosis of Hodgkin's disease. J Exp Ther Oncol. 2007;6(2):137-45.
4. Carvalho PC, Freitas SS, Lima AB, Barros M, Bittencourt I, Degrave W, Cordovil I, Fonseca R, Carvalho MG, Moura Neto RS, Cabello PH. Personalized diagnosis by cached solutions with hypertension as a study model. Genet Mol Res. 2006 Dec 18;5(4):856-67.
5. Catanho M, Mascarenhas D, Degrave W, de Miranda AB (2006). "BioParser: a tool for processing of sequence similarity analysis reports". Appl Bioinformatics. 5(1):49-53.
6. Catanho M, Mascarenhas D, Degrave W, Miranda AB. (2006). "GenoMycDB: a database for comparative analysis of mycobacterial genes and genomes". Genet Mol Res. 2006 Mar 31;5(1):115-26.
7. Guimarães AC, Otto TD, Alves-Ferreira M, Miranda AB, Degrave WM. In silico reconstruction of the amino acid metabolic pathways of Trypanosoma cruzi. Genet Mol Res. 2008 Sep 23;7(3):872-82.
8. Noronha MF, Lifschitz S, de Miranda AB. (2008). "A Practical Evaluation of BioProvider". Lecture Notes in Computer Science 5167, p. 174-177.
9. Otto TD, Gomes LH, Alves-Ferreira M , Miranda AB, Degrave WM. (2008). "ReRep: Computational detection of repetitive sequences in genome survey sequences (GSS)". BMC Bioinformatics, 9:366. doi:10.1186/1471-2105-9-366.
10. Otto TD, Guimarães AC, Degrave WM, de Miranda AB. (2008a). "AnEnPi: identification and annotation of analogous enzymes". BMC Bioinformatics Dec 17;9:544.
11. Otto TD, Vasconcellos EA, Gomes LH, Moreira AS, Degrave WM, Mendonça-Lima L, Alves-Ferreira

M. (2008b). "ChromaPipe: a pipeline for analysis, quality control and management for a DNA sequencing facility". Genet Mol Res. Sep 23;7(3):861-71.

The empresa brasileira de pesquisa agropecuária - EMPRAPA, Laboratory for Computational Biology

Dr. Goran Neshich

Embrapa, through its laboratory for Computational Biology, has a long record of offering services to academic partners through the internet using its experience and knowledge in maintaining its own product STING. It mirrors also public databases such as PDB, Uniprot, Prosite, HSSP, DSSP, ProTherm etc. while maintaining STING mirrors at 5 continents. Embrapa´s activities include intensive service, education and development and involves students as well as experienced colleagues both from Brazil and from Latin America.

Infrastructure

The lab is located in an environment with plenty of space for students, researchers, computer space, dedicated server space, dedicated space for training. Currently, the hardware infrastructure is going through an extensive renewal and new machines are being installed, replacing old (2001) acquired servers and PC stations. Expanded storage space is being acquired to aid in the ever growing problem of expanding disc space for DB and their back-ups. We have SUN, SGI and Dell clusters, totaling at about 60 CPUs, while total storage space around 15 Tb in separate servers. Around 15 Linux/Windows dual system PCs are dedicated for student work and other 20 are dedicated for training only. Due to infrastructure updating, lab reconstruction and team renewal, the lab stopped temporarily to offer general EMBnet services until all pending issues are resolved, in order to provide for better services.

Resources

We are restoring and expanding at the same time the STING and its database, transforming it into a federative contribution platform. We would like to offer to the EMBnet not only the new STING but also our experience in upgrading it, maintaining it, mirroring it and using it for educational purposes.

Education

We run a course for two major universities and their program for bioinformatics – UFMG and Unicamp. Both are well attended and teach mostly structural computational biology, but also some tools and DBs from sequence – one dimensional world.

During the last three years we offered total of 3 courses for more than 50 students, mainly covering material from structural computational biology and structural bioinformatics.

Database construction

We constructed a first Latin American database that was registered in the NAR DB issue. Since then, we aggregated many parameters into that same STING _ DB, making it the largest of its kind available for access over the web. Currently this database contains more than $28.5 \bullet 10^9$ registers (61,000 PDB files, ~130,000 chains, ~300AA/chain, 731 descriptors/AA).

Software development

We also published and posted on the Web STING suite of software programs for educational and analytical purposes. Analysis STING protocols are designed for routine use and can generate advanced reports about structure, sequence, function, stability and binding of proteins and their ligands.

Projects

- Study of Macromolecular Communication in Homo and Hetero complexes through their interfaces Unicamp-IB+Embrapa/CNPTIA. Large scale protein function prediction tools" Genoscope, France + mbrapa/CNPTIA
- "TargetsDB - Base de dados de alvos terapêuticos validados" UFMG + Embrapa/CNPTIA
- Automatic prediction of protein-protein interfaces based on a novel hydrophobicity index studies Unicamp-IB + Embrapa/CNPTIA
- Free Bioinformatics Technology consolidation and application in Biomedicine (FreeBIT) Red Iberoamericana de Bioinformatica + Embrapa/CNPTIA
- "Druggable proteins: Identification of potential therapeutic targets for development of agrochemicals, veterinary and medical drugs and vaccines for treating plant and animal diseases important for agriculture and live stock" UFMG+USP+UNICAMP+UNIFEI+EMBRAPA. GenoProtPlus SUN Computers e EMBRAPA

- Molecular modeling and structural analysis of the protein twitching motility a product of XF1633 gene of Xylella fastidiosa. EMBRAPA - CNPTIA

Publications

1. Jorge H. Fernandez, Marcia O. Mello, Leticia Galgaro, Aparecida S. Tanaka, Marcio C. Silva-Filho, Goran Neshich "PROTEINASE INHIBITION USING SMALL BOWMAN-BIRK TYPE STRUCTURES." Genet. Mol. Res. 6 (4): 846-858 (2007)
2. R.C. Melo, C. Ribeiro, C.S. Murray, C.J.M. Veloso, C.H.. da Silveira, Goran Neshich, W. Meira Jr, R.L. Carceroni and M.M. Santoro:"Finding protein-protein interactions by contact-maps matching."Genet. Mol. Res. 6 (4): 946-963 (2007)
3. Walter Rocchia and Goran Neshich:"Electrostatic Potential Calculation for biomolecules - creating a database of pre-calculated values reported on a per residue basis for all PDB protein structures."Genet. Mol. Res. 6 (4): 923-936 (2007)
4. Stanley R. M. Oliveira, Gustavo V. Almeida, Kassyus R. R. Souza, Diego N. Rodrigues, Paula R. Kuser-Falcão, Michel E. B.Yamagishi, Edgard H. Santos, Fábio D. Vieira, José G. Jardine and Goran Neshich:"STING _ RDB: A relational database of structural parameters for protein analysis with support for Data Warehousing and Data Mining."Genet. Mol. Res. 6 (4): 911-922 (2007)
5. NESHICH, G:"Computational Biology in Brazil."PLoS Comput Biol. Oct; v3 (issue 10, e185), 2007.
6. MOUTRAN, Alexandre; BALAN, Andrea; PEREZ, Carolina Santacruz; FERREIRA, Rita Café; RAMOS, Carlos; FERREIRA, Luís Carlos Souza; NESHICH, Goran: "Crystallographic structure and substrate-binding interactions of the 3 molybdate-binding protein of the phytopathogen 4 Xanthomonas axonopodis pv. citri."
7. R. C. Togawa, C. Ribeiro, I. Mazoni, T. Pelligrinelli, and NESHICH, Goran: " The Table of Interface Forming Residues as the Specificity Indicator for Serine Proteases Bound to Different Inhibitors "Accepted: BIOCOMP 08, 2008.
8. Carlos H. da Silveira, Douglas E. V. Pires , Raquel C. Melo , Cristina Ribeiro , Caio J. M. Veloso , Julio C. D. Lopes , Wagner Meira Jr, Goran Neshich , Carlos H. I. Ramos , Raul Habesch, Marcelo M. Santoro: " Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. PROTEINS: Structure, Function, and Bioinformatics,2008 Aug 14;74(3):727-743. 2008
9. Janaina Gomide ; Raquel Melo Minardi ; Marcos Augusto dos Santos ; Wagner Meira Jr. ; Julio Cesar Dias Lopes ; SANTORO, Marcelo; NESHICH, Goran:" Using Linear Algebra for Protein Structural Comparison and Classification."Genet. Mol. Res., 2009.

Chilean EMBnet node: progress report

J. Cristian Salgado

Universidad de Chile, DCC - Escuela de Ingeniería, Dept. of Computer Science, Santiago, Chile

In the past three years we have held a large number of activities related to bioinformatics and Systems Biology, which have had a big impact on the local scientific community. We have organized two mathematical modeling courses (1-2 weeks) aimed at the national biological research community.;

- the first one, Mathematical Modelling of Biological Systems, was offered in 2008 by local instructors with a solid background on biology, mathematics and computer science and was attended by over 70 participants;
- the second was an advanced course on Mathematical Models in Biology, was given by Benoit Perthame from Laboratorie J.-L. Lions, UPMC/INRIA, and it was attended by over 20 participants;
- in addition, we are planning on extending these outreach activities and offer a regional Latin American Systems Biology course next year (2010).

During this period, we have been awarded a major national grant (1 M$ USD/yr) for the creation of the Institute for Cell Dynamics and Biotechnology: A Centre for Systems Biology (http://www.icdb.uchile.cl/icdb). This institute is composed by scientists whose background and area of expertise is very diverse. There has been a tremendous amount of cross-fertilization between the scientists participating in this institute, leading to a large number of collaborations, such as:

Figure 1. ICDB Institute - The Chilean node.

- mathematicians and bioengineers/biotechnologists (modelling of joint gene regulation and metabolic networks);
- cell biologists, bioengineers and mathematicians (modelling iron mediated oxidation and cell ageing at metabolic and genetic levels);
- mathematicians, bioengineers and microbiologists (modelling substrate and electron diffusional effects in biofilms with microbial populations in bioleaching);
- proteomics experts and cell biologists and biotechnologists, (applying modern computer tools in the analysis of proteomic and gene microarray expression data and in the use of a metabolic model to simulate and optimize virus production for gene therapy vector synthesis);
- molecular modellers/medicinal chemists and cell biologists (modulating interactions between HFE/transferrin receptor).

This scientific diversity has generated a very large number of scientific collaborations between mathematicians, biologists, bioengineers, computer scientists and chemists which have focused both on Systems Biology as well as Mathematical Biology, a young innovative discipline in our country and in Latin America. These interactions have generated several industrial patents and over 65 scientific publications in high impact factors journals. One of them was in fact published on the EMBnet's special issue of BMC bioinformatics (2009, V10, S6). A full list of publications is available upon request.

In 2008 the institute supported and trained of 71 Ph.D. students, postdocs and young scientists in programmes ranging from bioengineering, mathematical modelling and computer science to biochemistry, neuroscience, microbiology and chemistry. These students are receiving formal training in biology, computer science and mathematics and most of them are doing part of their research projects abroad or in collaboration with top laboratories and research centres in US and Europe (Laboraratoire Jacques-Louis Lions University of Paris VI, University of Cambridge, University of Kent, University of Delft, National Biotechnology Center of Spain, Functional Genomics Centre at the University of Manchester, the Bioengineering Laboratory at Northwestern University, the Metabolomics Laboratory at the University of Stuttgart and the Proteomics Laboratory at the University of Virginia and others).

Colombian EMBnet node: progress report

Emiliano Barreto Hernández
Bioinformatics Center, Biotechnology Institute, National University of Colombia.

The Bioinformatics Center of the Biotechnology Institute at the National University of Colombia (CBIB) provides Bioinformatics tools, databases, training and support to the Colombian research community.

CBIB maintains the public access to updated versions of all the major genomics databanks (EMBL, PDB, UNIREF, UNIPROT, TAXONOMY, PROSITE and PATHWAY), through our SRS (sequence Retrieval System) service.

CBIB provides access to several known Bioinformatics tools such as EMBOSS (through EMBOSS:GUI, wEMBOSS and jEMBOSS), BLAST, SMS, SRS, DOTLET, PHRAP, PHRED , Consed), and also provides access to some tools developed "in-house", listed in the next table:

Tool	Details
SINCO	Pharmaceutical colorant database
BLA.id	Information system that allows the molecular and clinical data cross, and ß-lactamases identification from resistant organisms at intra-hospital level
ENKI	Database of molecular and taxonomy information of Colombian biodiversity
BLEE	Extended spectrum ß-lactamases (ESBL) molecular identification information

In the training area, the CBIB runs 6 Bioinformatics courses to Colombian scientists (during the academic semester), which were attended by more than 120 students. Besides this activity, the CBIB runs others training courses and workshops on Sequence Analysis, General Bioinformatics and Microarrays.

We have also continued supporting and advising the researchers involved in the following projects:

- Phylogenetic approach for a Colombia's Andes amphibian distribution hypothesis;
- ENKIdb, molecular and taxonomic linking system for Colombian species;
- the functional biodiversity and the edaphic metabolism of agricultural soils associated with potato crop;
- development of a bioinformatics tool for the identification of BLEEs genes;
- machine learning approach for the identification of protein binding sites;
- correlation with diagnostic categories a successful approach for selecting SNPs predictors of chronic fatigue syndrome.

In 2008, the node's hardware was upgraded with a DELL HPC machine, with 10 nodes (dual processor Xeon quad core 2.83 Ghz, 8Mb RAM) and storage with 6 Tb space. These machines represent a considerable improvement in our computing capacity, and that had immediate impact on our users' perception of our work.

Figure 1. Biotechnology Institute of the National University of Colombia.

Greek EMBnet node: progress report

Sophia Kossida

Biomedical Research Foundation of the Academy of Athens (BRFAA), Athens, Greece
http://www.bioacademy.gr

The Bioinformatics and Medical Informatics Team of the Biomedical Research Foundation of the Academy of Athens (BRFAA) is the Greek EMBnet node. It is headed by Sophia Kossida and it is composed of 26 persons including Senior Researchers, PhD candidates, Technicians, Master's and undergraduate students.

During the past two years, the Greek EMBnet node organized various workshops and seminars in BRFAA as well as in other Institutes in Greece, promoting Bioinformatics both at an introductory level as well as at an advanced one with specialized topics. In collaboration with the Swedish and Swiss nodes, two of these seminars entitled "Introduction to Bioinformatics" and "Introduction to Sequence Analysis" were delivered. Moreover, a Master's program has been organized together with the University of Athens entitled "Biomedical Informatics Technologies".

In late 2008, through the Greek EMBnet node, the first Bioinformatics book written in Greek was published, entitled "Bioinformatics: Potentials and Perspectives". The 14 chapters of this book are assembled harmonically in order to offer to the reader the basic knowledge to understand the specialized issues of Bioinformatics such as Proteomics, Systems Biology, Biological Databases etc. This book could be useful to students who want to familiarize themselves with the basic concepts of Bioinformatics, educators who want to have their knowledge updated and researchers who work on the same or similar scientific field. Many Universities and Technological Institutions in Greece use it already in their teaching curriculum for Bioinformatics courses.

The members of the team are actively pursuing research within different fields of Bioinformatics. Several publications have come out the last two years. A list of them is accessible at: http://www.bioacademy.gr/bioinformatics/publications.htm

The immediate future plans of the team include first an upgrading of its hardware infrastructure in order to host implemented web bioinformatics tools specialized in protein data analysis. Then, we plan to develop an e-learning bioinformatics platform for Greek Universities and Technological Institutions, which will be funded exclusively from National funds. Finally, additional seminars will be organized in BRFAA, introducing Bioinformatics to graduate Greek students.

Figure 1. Mycenae:This mythical city of king Agamemnon was the background for many classical Greek tragedies and the beginning of the Trojan war of the Iliad.

Pakistan EMBnet node: progress report

Nazim Rahman, Shahid N. Chohan

COMSATS Institute of Information Technology (CIIT), Department of Biosciences, Chak Shahzad Campus, Islamabad, Pakistan

COMSATS Institute of Information Technology (CIIT), Department of Biosciences in Islamabad is the EMBnet Pakistan National Node. The node started in 2006. Following is a brief report of our activities and future goals.

In 2006, little was known about bioinformatics in Pakistan and there were very few resources available for bioinformatics education and training. We addressed this problem by:

- providing regular workshops for CIIT Biosciences faculty;
- reforming existing curricula and developing new bioinformatics curriculum;
- developing online learning resources;
- organizing and participating in workshops, seminars, conferences.

There is a severe shortage of skilled bioinformaticians in Pakistan and most bioinformatics courses are taught by computer scientists and biologists who teach their parts without context to the other. In the absence of bioinformaticians, the best option is to train biologists and computer scientists in bioinformatics. Regular workshops for bioinformatics instructors proved very successful to improve bioinformatics education.

We proposed reforms in existing curriculums and modules of courses which permit existing faculty members to teach bioinformatics concepts and skills. For example, the Genomics course at CIIT Islamabad now covers metagenome annotation using annotathon (http://annotathon.univ-mrs.fr/).

We are trying to convince decision makers to increase the number of institutions offering bioinformatics programs since the five Pakistani Universities offering bioinformatics education can educate only a fraction of the students wishing to pursue education in bioinformatics. This year we introduced a Bioinformatics Program in CIIT Sahiwal and MS Bioinformatics in CIIT Islamabad. The MS Bioinformatics program has been created with intent to overcome our severe shortage of qualified bioinformatics instructors, researchers, and developers.

Pakistan and the developing world in general won't be able to satisfy its bioinformatics education requirements in the near future due to shortage of qualified persons. Given this situation, online education is a very attractive option. We launched BIREC (Bioinformatics Information Resource and Elearning Center,(http://www.birec.org) on January 1, 2009. This resource is becoming quite popular within and outside Pakistan. Comprehensive online bioinformatics courses that would include PowerPoint presentations, videos, exercises, quizzes, and much more are also in the pipeline. Once live, these courses would be a free and easily accessible resource for bioinformatics educators and students worldwide.

Eventually, we would like to offer an online degree program in bioinformatics and we are actively working towards achieving this goal. However, we are still a few years away from realizing this goal. With the installation of PERN-2, we now have the necessary infrastructure to realize this goal. See http://pern.edu.pk/index.php for more information on PERN-2.

EMBnet Pakistan National Node lobbied hard and won approval of full scholarships for 50 Pakistani students to pursue PhD Bioinformatics overseas from HEC. See: http://www.hec.gov.pk/InsideHEC/Divisions/HRD/Scholarships/ForeignScholarships/HECOSBN/Pages/Default.aspx, for details. Eleven students are currently pursuing PhD in Bioinformatics at Uppsala University in Sweden.

To raise awareness of Bioinformatics and its applications, members of EMBnet Pakistan National Node regularly make oral and poster presentations in scientific events. In addition, we solely or partially organized several bioinformatics events:

- Role of Bioinformatics in Medical Research and Health Services
- Latest Trends in Bioinformatics
- Crash Course in Bioinformatics
- Use of Bioinformatics in Genomics Research
- Bioinformatics: Current Progress and Practical Applications (Baku, Azerbaijan)
- Workshop on Bioinformatics
- COMSTECH-CIIT National Workshop on Bioinformatics for Computer Scientists

Currently we have 2 Xeon Servers, eBioKit running, 2 bioinformatics labs, and direct link to PERN-2. We also provided Jemboss in 2006-2007. We are currently striving to improve and upgrade our computing facilities.

We have contributed 2 articles to EMBnet.news and 2 Quick Guides to EMBnet.org. Our contributions to EMBnet.news and EMBnet in general are expected to increase in future.

11 Bioinformatics undergraduate research projects were completely successfully under the supervision of members of our node.

Our node has come a long way since 2006. Most of the work was concentrated around raising awareness about bioinformatics and training the first generation of bioinformaticians.

Our goals for the next three years are:

- create avenues for utilizing the skilled manpower for scientific research and development;
- turning http://www.pk.EMBnet.org into bioinformatics portal as well as node website which would integrate tightly with http://www.EMBnet.org;
- setup bioinformatics infrastructure accessible outside the node;
- provide bioinformatics analysis services;
- increase and consolidate our online elearning activities;
- continue seminars and workshops to raise awareness on Bioinformatics and its applications in Life Science.

We are very thankful to EMBnet for their support and guidance since we joined the network in 2006. Without their support and guidance, we would have been much less successful in our effort to introduce and promote Bioinformatics in Pakistan.

Spanish EMBnet node: progress report

José R. Valverde

EMBnet/CNB, Centro Nacional de Biotecnología, CSIC. Madrid, Spain

During the last year the Spanish EMBnet node has been active in the Executive Board of EMBnet (see EB report in this issue).

Several EMBnet courses have been organised at CNB in Madrid during the period 2008-2009:

- International EMBnet introductory course on Molecular Dynamics using TINKER (extended and based on contents from Swiss EMBnet node course using CHARMM)
- National EMBnet introductory course on Quantum Biology (training delivered in Spanish)
- Biostatistics (an updated and extended version of previous introductory courses).
- Advanced UNIX (organized as a seminar series every Wednesday for three months).

All courses were organized as hands-on courses with theoretical lectures paired with practical sessions where students could try the tools first hand.

Additionally, the node manager participated in the II Workshop on Bioinformatics organized in Ota University, Canaanland, Nigeria on July. The course made heavy used on the in-development USB key from EMBnet/CNB. The key is now ready for use and we intend to update it periodically (see EMBnet.news vol. 15, nr. 3: pp. 8-11).

EMBnet/CNB has also actively participated on the organization of the Next Generation Sequencing Workshop held in Rome in November.

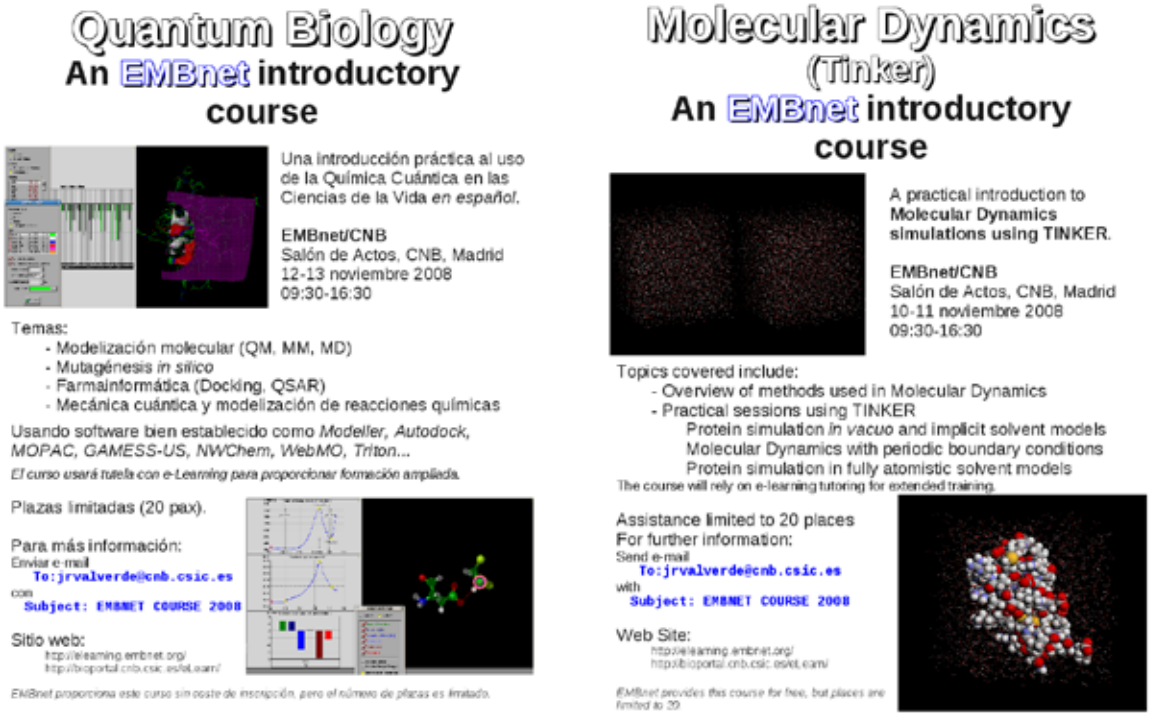

Figure 1. Announcements of the MD and QB courses held at CNB in November 2008.

In November last year, and in cooperation with the Austrian node, we ensured transfer of critical services to the Spanish node preserving URLs. This initiative ensures continued availability of services published in scholarly journals and hosted in EMBnet as per common editorial policies.

As a side work, EMBnet/CNB has been active in the promotion and creation of the Iberoamerican Bioinformatics Society, which is intended to take the lead from RIBIO, as well as in the preparation and submission of various proposals (of which FreeBIT proposal decision is still pending).

On the minus side, the current crisis is heavily affecting the development of node activities. A decision has been reached to offload computational work to related institutions in Spain (CSIC and CESGA), and to concentrate services in increasing local return for CNB. For the next year we project delivery of EMBnet courses in:

- Biostatistics
- Sequence Analysis
- Evolution
- Molecular Modeling
- Quantum Biology
- Analysis of High Throughput data

These courses will be offered first to CNB scientists and, if there is free space, will be open as International EMBnet courses. If possible, two editions (local and international) may be provided.

The Spanish node wants to stress its strong belief in the utmost importance of EMBnet delivering visible benefits to node hosting institutions for the safe continuity of its functions (notwithstanding community-at-large service provision) and would like to encourage the community to participate in discussions seeking the best way to achieve this goal.

Sri Lankan EMBnet node: progress report

Kanchana Senanayake
Institute of Biochemistry, Molecular Biology and Biotechnology (IBMBB), University of Colombo, Sri Lanka

Mission - IBMBB

"INITIATE, PROMOTE, FACILITATE Advance Research & Human Resource Development in Molecular Life Sciences to achieve National and International Development"

The Institute of Biochemistry, Molecular Biology and Biotechnology is an independent institute of the University of Colombo established in 2004. It is located in the main campus of the University of Colombo.

Its main activities are in human resource training and research and development in Molecular Life Sciences and allied fields. It has two masters programs (Master of Science in Molecular Life Sciences, Master of Science in Cellular and Molecular Life Sciences), MPhil/PhD programs and several research programs in genomics/proteomics. A Master of Science program in Bioinformatics is currently being developed.

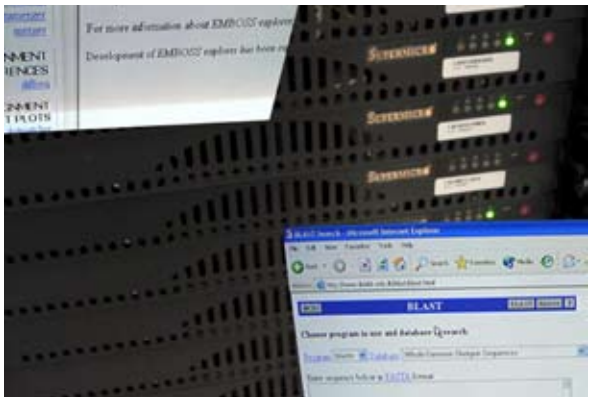

Figure 1. A computer cluster and screenshots from some of the applications it runs.

Figure 2. The IBMBB building.

Institute of Biochemistry, Molecular Biology and Biotechnology was first elected as the Sri Lankan national node in 2007. Since then we have managed to achieve several goals such as dissemination of knowledge through workshops and training programs and setting up resources in bioinformatics. Being the only institute of this kind in Sri Lanka we are aware of our responsibility to provide technical support and knowledge to Sri Lankan Molecular Biologists on using Bioinformatics tools for their research and to provide access to online databases and tools in bioinformatics.

Using the knowledge and hands on experience gained by the training received in Sweden we have managed to configure a bioinformatics cluster with a local installation of the BLAST tools and the mirror of the BLAST database and it also includes the EMBOSS explorer which is a GUI for the EMBOSS tools. This cluster is temporarily hosted at the URL www.ibmbb.edu.lk on the public domain.

- BLAST search tool; www.ibmbb.edu.lk/blast
- EMBOSS explorer; www.ibmbb.edu.lk/emboss

There is also a local installation of the ebiokit which includes Ensembl genome browser, Biomart, MRS, Galaxy and wEMBOSS toolkits for Bioinformatics. The "ebiokit", a collection of open source bioinformatics programs was installed at the IBMBB by Prof. Erik Bongcam-Rudloff's research team at the Swedish University of Agricultural Sciences, Sweden.

For further details regarding IBMBB please visit www.ibmbb.lk

ILRI-BecA , EMBnet specialist node: progress report

Etienne de Villiers

EMBnet BecA-ILRI, Nairobi, Kenya

Scientific achievements

The mission of ILRI-BecA node is to provide widely used bioinformatics tools, databases and data storage to the east and central African research community through a central Bioinformatics platform. For that reason we tested the feasibility of the created bioinformatics solutions with the following "test cases":

1. application of genomics and immunoinformatics to identification of genes related to virulence in *Mycoplasma mycoides* the causative agent of contagious bovine pleuropneumonia (CBPP);
2. application of genomics and proteomics to camel *Streptococcus agalactiae*: development of vaccines and diagnostics to support camel milk marketing through improved control of mastitis.

Work on both "test cases" is still ongoing but several preliminary results demonstrate that the work have strengthened the research capacity of the involved scientists and academic institutions exploiting the bioinformatics resources deployed and created at the node.

Software developed

eBioUSB: during the Introduction to Bioinformatics course in 2007 students used a Bioinformatics Live-CD with a large number of bioinformatics software, it was however not possible for them to easily save their work. In 2008 we developed a "Bioinformatics workbench on a USB memory stick", eBioUSB [1] in collaboration with the Sweden EMBnet node, which contains all the basic bioinformatics tools, taught during the course and enabled students to save their data on the same device.

eBioMacKit: the Swedish EMBnet node developed the eBioMacKit to serve as a portable bioinformatics workbench or a local area server for a mid-sized laboratory. The solution is based on external portable harddrives that contain all the databases. This enables one to update large databases by exchanging them for new disks with new data, circumventing the poor Internet bandwidth issues in East and Central Africa. The BecA nodes in East and Central Africa will use this server solution to serve as a general bioinformatics platform [2].

Figure 1. Etienne de Villiers lecturing.

Capacity building

The ILRI-BecA node hosted several introductory Bioinformatics courses. In May 2008 we presented a 9-day course "Introductory course in Bioinformatics" [3]. The course team consisted of four persons, Erik Lagercrantz, Maria Wilbe, Erik Bongcam-Rudloff, Alvaro Martinez Barrio, from Swedish University of Agricultural Sciences (SLU), Uppsala University (UU) and Linnaeus Centre for Bioinformatics (LCB) and two persons, Etienne de Villiers and Saidimu Apale, from the ILRI-BecA node. We received 74 applicants from eight Eastern and Central African countries. We could only accommodate 24 applicants that were selected from Uganda, Sudan, Tanzania, Burundi, Somalia, Cameroon, Ethiopia and Kenya. Four travel fellowships were awarded to participants from Sudan, Tanzania, Cameroon and Ethiopia. The lectures were recorded on video and subsequently used at the University of Buea, Cameroon and Maseno University in Kenya. Students re-

Figure 2. ILRI staff and visitors.

ceived an eBioUSB to take back home to continue with their work. The course was very well received.

A second 'Introductory course in Bioinformatics' was held in May 2009 at the ILRI-BecA node. The course team consisted of, Etienne de Villiers (ILRI-BecA), Erik Bongcam-Rudloff, Hans-Henrik Fuxelius and Katharina Truve from the Swedish University of Agricultural Sciences (SLU), Uppsala University (UU) and Linnaeus Centre for Bioinformatics (LCB) and George Githinji (KEMRI-Wellcome Trust program). In total 25 applicants from Kenya, Tanzania, Uganda and South Africa were selected from 152 applications. Eight travel fellowships were awarded to participants from Tanzania and Uganda. An Internet video conferencing system was used to enable 22 students from University of Khartoum, Faculty of Animal Production in Sudan to participate. A former student from the 2008 course, Dr. Abdelaziz Ahmed Fadlelmoula, facilitated this action. The response from the University of Khartoum and Sudanese higher education authorities was positive and they have shown interest to host the next course in Sudan. Each of the students received an eBioUSB.

Students from a course we presented in 2006, established a Regional Student Group (RSG) affiliated with The International Society for Computational Biology (ISCB) in 2007. The goal of the RSG's is to conduct events, which will be beneficial to the professional development of Bioinformatics and Computational Biology students on a local level.

RSG East-Africa was very active in promoting Bioinformatics in the region and organized a one-day Bioruby/Bioperl workshop in May 2008 hosted and funded by the ILRI-BecA node. A total of 20 participants participated in the workshop facilitated by George Githinji.

In July 2008, RSG East-Africa organized a two-day introductory course using R and Bioconductor with help from the ILRI-BecA node.

The RSG East-Africa group invited Prof. Anna Tramontano from the Department of Biochemical Sciences, University of Rome "La Sapienza", Rome, Italy, for a 3-day Introductory course in Proteomics in November 2008. The node helped with logistics, providing accommodation for Prof. Tramontano and awarded two travel fellowships to enable students from Uganda and Tanzania to attend the course.

The RSG East-Africa was actively involved in organizing the first "African Virtual Conference on Bioinformatics 2009 hosted by the ILRI-BecA node in February 2009 [4].

Publications

1. New tools for bioinformatics teaching. Wilbe M. and Bongcam-Rudloff, E., EMBnet.news, 14.2: 3-4.
2. eBioMacKit: a bioinformatics portable teaching kit. Alvaro Martinez Barrio and Erik Bongcam-Rudloff, EMBnet.news 13.4: 6-10.
3. Nairobi, Kenya course report. Bongcam-Rudloff E., de Villiers EP., EMBnet.news, 13.1: 3-4.
4. AFBIX09: bioinformatics virtual conference. Nelson Ndegwa, EMBnet.news, 15.1: 16-17.

Figure 3. The ILRI Facilities in Nairobi, Kenya.

UMBER - University of Manchester Bioinformatics Education & Research

EMBnet specialist node: progress report

Teresa K.Attwood

Faculty of Life Sciences, University of Manchester, UK

Introduction

As a Specialist Node, UMBER receives neither finance nor a mandate from government to serve a national community. Instead, working in an academic institution, our role is to perform research and teaching, the fruits of which labours we offer freely on the Web. Page restrictions forbid a full account of our work in the last 3 years, but here we give a flavour of some of our activities.

European Projects

We have participated in a number of European projects: e.g., we've contributed to several workpackages within EMBRACE (European Model for Bioinformatics Research and Community Education), where our principal outputs have been further development of Utopia (utopia.cs.man.ac.uk) and creation of the Web Service Registry (www.embraceregistry.net); within IMPACT (which maintains and enhances InterPro), we continue to support the development of the PRINTS protein fingerprint database, and are currently providing guidance on protein family classification; in the context of EuroKUP (European Network for Kidney and Urine Proteomics), we coordinate Working Group 4 (with Erik Bongcam-Rudloff) and liaise with eLICO (the e-Laboratory for Interdisciplinary Collaborative Research in Data Mining and Data-Intensive Science).

Figure 1. Interior of the Michael Smith Building, where UMBER is located.

Tools & Resources

We've added a range of new software tools to the DbBrowser Web server (www.bioinf.manchester.ac.uk/dbbrowser), primarily for protein sequence analysis and database annotation: the most prominent of these are MINOTAUR (www.bioinf.manchester.ac.uk/dbbrowser/minotaur/about.html) and Utopia (as mentioned above). We also continue to maintain, and contribute towards the development of, key bioinformatics databases, such as InterPro, PRINTS (www.bioinf.manchester.ac.uk/dbbrowser/PRINTS) and the Central Aspergillus Data REpository, CADRE (www.cadre-genomes.org.uk).

Affiliations with Societies and Journals

Much of our work concerns database maintenance and annotation, so I was pleased to see the formation of the new International Society for

Biocuration (ISB, www.biocurator.org). I was recently elected to the ISB's Executive Board, and I hope that EMBnet may have a productive relationship with this highly relevant Society. We have also formed a fruitful relationship with Portland Press, publishers of the Biochemical Journal. In a joint project, we have created Utopia Documents (getutopia.com/products/documents), which is underpinning a new venture in 'semantic' scholarly publishing. Again, I hope to extend this relationship to other publishers and journals, and especially to EMBnet.news.

Selected publications

Peer-reviewed journals

1. Attwood TK, zKell DB, McDermott P et al. (2009) Calling International Rescue – knowledge lost in literature and data landslide! Biochem J 242: 317-333.
2. Pettifer S, Thorne D, McDermott P, Marsh J, Villeger A, Kell DB, Attwood TK (2009) Visualising biological data: a semantic approach to tool and database integration. BMC Bioinformatics 10: S18.
3. Pettifer S, Thorne D, McDermott P, Attwood T et al. (2009) An active registry for bioinformatics Web services. Bioinformatics 25: 2090-2091.
4. Hunter S, Apweiler R, Attwood TK, Bairoch A et al. (2009) InterPro: the integrative protein signature database. Nucleic Acids Res 37: D211-D215.
5. Mabey Gilsenan J, Atherton G, Bartholomew J, Giles PF, Attwood TK et al. (2009) Aspergillus Genomes and The Aspergillus Cloud. Nucleic Acids Res 37: D509-D514.
6. Stockinger H, Attwood TK, Chohan SN, Cote R et al. (2008) Experience using Web services for biological sequence analysis. Brief Bioinform 9: 493-505.
7. Vlahou A, Schanstra J, Frokiaer J, El Nahas M, Spasovski G, Mischak H, Domon B, Allmaier G, Bongham-Rudloff E, Attwood TK (2008) Establishment of a European Network for Urine and Kidney Proteomics. J Proteomics 71: 490-2.
8. Nordle AK, Rios P, Gaulton A, Pulido R, Attwood TK, Tabernero L (2007) Functional assignment of MAPK phosphatase domains. Proteins 69: 19-31.
9. Kim J-H, Mitchell AL, Attwood TK, Hilario M (2007) Learning to extract relations for protein annotation. Bioinformatics 23: i256-i263.
10. Roma-Mateo C, Rios P, Tabernero L, Attwood TK, Pulido R (2007) A novel phosphatase family, structurally related to dual-specificity phosphatases, that displays unique amino acid sequence and substrate specificity. J Mol Biol 374: 899-909.

Figure 2. Exterior of the modern Michael Smith Building.

Newsletters & Magazines

1. Attwood TK, Kell DB, McDermott P, Marsh J, Pettifer SR, Thorne D (2009) Knowledge lost in literature and data landslide: Calling International Rescue! The Biochemist, December 2009, pp.23-38.
2. Pettifer S, Attwood TK, McDermott P et al. (2007) UTOPIA: User-friendly Tools for OPerating Informatics Applications. EMBnet.news 13: 19-24.

Books & book chapters

10. Higgs P, Attwood TK (2008) Bioinformatyka i ewolucja molekularna (Bioinformatics and Molecular Evolution). Murzyn K, Liguzinski P, Kurdziel M Warsaw translators. Wydawnictwo Naukowe PWN.
11. Attwood TK, Parry-Smith DJ, Phukan S (2007) Introduction to bioinformatics. India: Dorling Kindersley.
12. Attwood TK (2007) Genetic databases. In: Encyclopaedia of the Human Genome. London: Nature Publishing Group.
13. Cammack R, Attwood TK, Campbell PN, Parish JH, Smith AD, Stirling JL, Vella F editors. (2006) Oxford Dictionary of Biochemistry and Molecular Biology, Second Edition. Oxford: Oxford University Press.
14. Attwood TK, Mitchell A, Gaulton A, Moulton G, Tabernero L (2006) The PRINTS protein fingerprint database: functional and evolutionary applications. In: Jorde L, Little P, Dunn M, Subramaniam S editors. Encyclopaedia of Genetics, Genomics, Proteomics and Bioinformatics. New Jersey: John Wiley & Sons.

Jornadas de Bioinfomática JB2009 - November 3rd-6th 2009

Pedro Fernandes

Instituto Gulbenkian de Ciência (IGC), Oeiras, Portugal

More than a year ago, the portuguese and spanish bioinformatics communities decided that they should organize a common conference with the aim of reinforcing existing links and creating new ones. Geographical reasons justify it fully, as the costs of doing things together is reasonably low since they mostly concern rather short travels. The Instituto Gulbenkian de Ciência, with the longest record of bioinformatics activities in Portugal, and host to the Portuguese Node of EMBnet since 1990, organized the event. It itook place at the headquarters of the Fundação Calouste Gulbenkian in Lisboa.

The meeting was physically attended by about 130 participants, more or less half from Portugal and the other half from Spain. The underlying theme was 'Challenges in Bioinformatics' and that justified the choice of the four keynote speakers: Tanja Kortemme (Molecular Design

Figure 1. Pedro Fernandes introduces Pooja Jain to James Watson.

of Multi-Specific and Selective Protein-Protein Interactions), Can Alkan (Structural Variation Discovery and Characterization of Segmental Duplications with Next-Gen Sequencing

Figure 2. JB2009 Group Photo.

Technologies), Paul Scheet (Making sense out of 1000 Genome) and Erin Halperin (Computatioal Challenges in the Analysis of Genetic Variation). In total there were 23 oral presentations and 53 posters. Two discussion panels were set-up, one on 'New Sequencing Technologies: Opportunities and Challenges' and the other on 'Spanish / Portuguese cooperation in Bioinformatics'.

The conference ended on November 6th, just before the start of a talk by James Watson, that happened to be organized in the same place. He was photographed while he was being introduced by the local organizer to Pooja Jain of the University of Nottingham, presenter and co-author of the poster entitled Constrained Protein Topology Determination, that won the best poster award.

Several collaborations have started during this event, as expected. The value of organizing conference together was restressed, and plans have been laid to do it on a yearly basis.

Figure 3. Presentation in one of the parallel sessions, JB2009.

Next Generation Sequencing Workshop

November 18-20, 2009. Rome, Italy

José R. Valverde

EMBnet/CNB, Centro Nacional de Biotecnología, CSIC. Madrid, Spain

Abstract

In November 2009, EMBRACE (FP6 NoE), CASPUR, EMBnet, the Italian Society of Bioinformatics and UPPMAX allied to organize a workshop on Next Generation Sequencing technologies and data analysis tools focusing on "Building Next Generation Sequencing platforms and pipeline solutions". The workshop took place in Rome and was addressed to bioinformaticians interested in discussing issues related to the management of platforms and analysis of next-generation sequencing (NGS) data. In addition to lectures from leading scientists in the field, the workshop also included a hands-on session, a "hack-a-thon" where participants could try by themselves some of the tools to solve an unpublished, real world problem in genome sequence assembly. The workshop web site is located at www.nextgensequencing.org.

Introduction

Next Generation Sequencing technologies have drawn the attention of the Molecular Biology scientific community by its unprecedented power and relatively low cost. It is now feasible to perform sequencing runs covering billions of base pairs in fragments from 35 to 400 bp that can be put to traditional (e.g. sequencing a genome with 50x coverage) and innovative (e.g. sequencing all transcripts in an organism or metabolome) uses.

The rapid advance of technology has raised concerns in many leading edge scientists, which see amazing new possibilities in this progress but find traditional tools designed to cope with conventional sequencing tasks wanting and newly developed tools too complex, difficult or specific to be conveniently applied to their interests. Out of this interest in current technological and scientific progress, arose a demand for bringing together the experts in the field to gather an overall view of the current status of the art and define future directions for dealing with the analysis problems.

In order to address this demand, the EMBRACE (EU FP6 NoE), UPPMAX and CASPUR with the support of the Italian Society of Bioinformatics and EMBnet organised a workshop in Rome, from the 18th to the 20th of November on Next Generation Sequencing technologies, addressed to bioinformaticians interested in learning more and discussing about these new techniques and the approaches to analyze the data they generate building appropriate infrastructures and data analysis pipelines.

Actual interest in the workshop was reflected in the fact that the original plans had to be altered in order to accommodate the large demand of scientists wanting to participate and that required reassignment and location of meeting facilities in order to be able to expand available space and increase the number of participants.

The workshop was organized with some major goals in mind: first of all, it should help participants get an overview of the applied scientific problems being addressed and the tools used to tackle them; secondly, we wanted to show as well the daunting magnitude of the problem posed by these new technologies and the complex platforms needed to be able to manage the large amounts of data generated; and finally, we wanted to provide both specialized and user views on the problem and provide a nurturing environment for discussions that fostered the exchange of experiences and feedback to developers shedding new light towards future directions to be pursued.

To this end, the organizers assembled an international team with varied expertise, interests and geographical provenance to constitute the organizing committee and ensure appropriate coverage of all the topics addressed (Erik Bongcam-Rudloff, Tiziana Castrignano, Eija Korpelainen, Inge Jonassen, Graziano Pesole, Nils-Einar Eriksson, Etienne deVilliers, Andreas Gisel, Laurent Falquet, Jose R Valverde and Gert Vriend). The committee in turn assembled a pan-

Figure 1. Group picture of workshop attendees.

el of speakers that we are glad to report successfully met in our view and, most important, in the eyes of all participants, the goals defined at the onset.

Workshop contents:

Welcome address, Erik Bongcam-Rudloff

The workshop opened with a welcome address of the chairman of the Organizing Committee, Dr. Erik Bongcam, who introduced rational for the meeting, the main lines to be developed, and thanked all the people and organizations that contributed to make it possible: EMBRACE (thanks to Gert Vriend), CASPUR (thanks to Tiziana Castrigliano), the Engineering Dept. of University of Rome, the Uppsala HTC facility UPPMAX (thanks to Ingela Nyström) and the EMBnet.

After this keynote address, the workshop moved on to address the issue of building sensible analysis platforms for NGS data analysis and where the main approaches used to address this problem where reviewed first hand:

Hardware and Storage, Tony Cox, Welcome Trust Sanger Institute, United Kingdom

In his talk, Tony Cox described the major challenges posed by NGS to Information Technologies and Data Management: data generated is huge, easily resulting in 500.000 images, with a weight of 8MB each, per run. These will need to be analyzed and converted to sequence data prior to any actual practical use. Once sequences are available, various uses will be possible depending on the input samples and the intended use of the information, resulting in turn in large analysis result datasets as well. From here, we have only reached the first step in the scientific inference process and further downstream analysis will come as scientists use these results to build upon new analysis.

All in all, this results in a logical division of the process into a pipeline, where output from one step feeds the next in the analysis process, all of them generating vast amounts of data that needs to be efficiently and safely managed.

Tony Cox described the approach taken by a major sequencing oriented institution such as the Sanger Center, the challenges they found, the solutions the came up with and the experiences they could drive out from the process, giving a series of rules of thumb useful for anybody planning to deploy an NGS data analysis infrastructure, probably the most relevant of which was the advice to plan for change: technology is evolving quickly and users are continuously coming up with new applications.

Data storage for HTS platforms. George Magklaras. Biotechnology Centre, Oslo, Norway

George Magklaras addressed the problem posed by High Throughput Sequencing (HTS) technologies from a similar standpoint, but with a different twist: the case of a centralized computing facility that has to cope with the needs of a

distributed network of users spread at a national level.

This presentation, like all the others that followed, confirmed the soundness of the pipeline approach to data analysis and management and reviewed the requirements and needs of a large community of high throughput sequencing users, presenting estimates well in line with the experience of Sanger Center and all the other speakers.

In his talk, George Magkalas concentrated on the various solutions available to deal with the management and storage of the data generated and pointed out the pros and cons of each solution, giving a set of recommendations to help define the hardware and software requirements of a sound NGS data analysis platform.

A multidisciplinary platform of computing resources, large scale storage and know-how in Uppsala, Jonas Hagberg, Uppsala University, Sweden.

The next presentation addressed the same problem, dealing with data analysis from scientists at a national level using a different approach: deployment of a coordinated, distributed national e-infrastructure over various centres.

Again, Jonas Hagberg confirmed the requirement estimates and experiences described on previous talks, agreeing in the basic pipeline approximation, but his talk provided an altogether different point of view to solving user needs, based on extended consultation with users and experts and a distributed solution to platform implementation, presenting rationale for the decisions they made and the solutions chosen to provide adequate support to users.

Sequence read archives at EBI, Guy Cochrane, EBI, United Kingdom

Guy Cochrane introduced the European Nucleotide Archive (ENA) being developed at EBI and approached the problem of data storage from a new perspective: that of pure data management instead of concentrating on solving the infrastructure problem (already addressed by G. Magklaras).

Guy Cochrane described the mechanisms and services being implemented in ENA, the kind of users it is addressed to and the uses currently envisioned for this archive as well as the services being implemented to deliver the functionality planned, addressing issues like data formats, standardization and coordination with other major players like NCBI. He also described the current approaches they are considering for data submission, retrieval and distribution, and gave us a glimpse to what they already have in the works for the future.

Bioinformatics for NGS Giorgio Valle, CRIBI, University of Padua, Italy

Giorgio Valle gave us yet another view on how to deal with the NGS data nightmare: that of an international consortium addressing major sequencing undertakings. His talk also acted as a knee play to the next topics in the workshop, moving attention towards the problem of data analysis itself.

The experiences derived from major sequencing enterprises, like the tomato genome was reviewed and put in historical context, showing the advantages that new NGS technologies can provide to classical sequencing approaches and the useful combination of both.

In addition, Giorgio covered the various approaches used to deal with a major undertaking and reviewed the software tools used and the development work they had to do to assemble, build scaffold paths and manage redundancy in the data to reconstruct the base genetic information needed to procced to their next logical step: gene prediction using sequence alignments and ab-initio methods, and subsequent whole transcriptome analysis, thus highlighting the dramatic nature of the problem: genomic sequencing is but the first step in a series of increasingly more complex and useful analysis where NGS techniques can also be of great help.

Mapping short reads as numbers, Alberto Policriti , Applied Genomics Institute Univeristy of Udine, Italy

The problem of sequence alignment, as demonstrated in the previous talk, is central to most of the analysis we want to perform on NGS data. In this talk, Albeto Policriti reviewed the basic theory related to the problem as a basis to describe new methods in development and use that aim at overcoming traditional shortcomings of classical algorithms. He gave a cursory review of many of the current tools of the trade. Then he moved on to present some of their most novel algorithms being developed to accurately match short tags and their applications.

Figure 2. Fontana di Trevi.

NGS alignment and applications, Zeming Ning, Sanger Center, United Kingdom

Zeming Ning pursued the methodological analysis of the tools used for NGS data analysis by presenting the new SSAHA2 algorithm. Again, he reviewed existing methods, their advantages and shortcomings and used this context to describe recent advances and applications as well as the typical analysis workflow.

Project HOPE: The last piece of the puzzle, Hanka Venselaar, University of Nijmegen, The Netherlands

With the methodological basis for sequence alignment covered, the workshop moved on to downstream data analysis. Hanka Venselaar described project HOPE, a new tool aimed to make easier the life of end users by allowing them to automate the comprehensive (to the limits of available knowledge) analysis of a protein, using both sequence comparisons, structural predictions and structural modeling and comparison to deliver ready to use, understandable knowledge.

Hanka's talk marked a new turning point in the workshop, shifting the topic from raw sequence comparison to downstream applications.

De novo assembly, Laurent Falquet, Vital_IT, SIB, Switzerland

In this talk, Laurent Falquet addressed the issue of de novo sequence assembly, reviewing the problems and challenges, the basic methodological and algorithmic approaches and the most popular tools in use, without forgetting to drive attention to the upcoming problem of ultra-high throughput sequencing of thousands or millions of genomes.

Given the core interest of this problem, a good understanding of the tools and how well they compare against each other, how and when they should be used and the limits of the state of the art is a most valuable knowledge to derive from experience. Laurent Falquet drove of their ongoing projects and on test experiments they carried out to distill this knowledge and present it in a useful way.

Vertebrate sequencing, Jim Stalker, Sanger Institute, United Kingdom

The current interest in NGS has driven many scientists to start a wealth of specific projects, as described by previous talks. But well beyond these "petty" problems, major sequencing institutes are already considering vast problems like the "1000 genomes" project that will collect 1000 human genomes from around the world... and more projects aimed at other organisms.

In his talk, Jim Stalker described how Sanger Center and EBI as coordinators manage the huge load of data, defining standards for storage and workflows that help involved scientists efficiently address these large projects that, although seemingly huge today, may become standard experiments sooner than expected. Their experience in dealing with these is most helpful to organize and plan work at any institution interested in NGS, and was summarized as a series of recommendations of general application.

Alternative splicing using NGS data, Graziano Pesole, University of Bari, Italy

Moving ahead with the problem of practical data analysis, one of the obvious steps after genome sequencing is transcriptional analysis, identification of expressed genes and characterization of their transcripts. Graziano Pesole proceeded to present the recent breakthroughs in determining transcriptional maps and identifying alternative splice sites, highlighting the relevance of what we know is a major issue that is expanding up to 10 times the human transcriptome.

Here again, deep sequencing has obvious advantages: it is now feasible to identify large numbers of transcripts by sequencing the cDNAs and mapping these back to the genome, hence gaining first hand experimental evidence of alternative gene expression. Graziano Pesole reviewed the various alternative approaches available to analyze the data, identify transcription start and termination and splice sites and reviewed the existing software that can be used to this end.

NGS transcriptomics, Marc Sultan, Max Planck Institute Berlin, Germany

Marc Sultan pursued the quest into transcriptomics started by Graziano Pesole in the next talk: with NGS not only can we identify transcripts, we can also quantify gene expression by measuring the presence of each transcript in the recovered sequences. Marc Sultan reviewed his experience and gave useful advice on how best to plan and pursue a gene expression experiment using NGS, reviewing the roadblocks, biases and artifacts we need to be aware of and the methods to evaluate statistical significance of resulting data.

Marc Sultan also provided an end user point of view and raised the issue of usability of existing tools for facilitating and accelerating their adoption and exploitation in the real world.

Chipster for NGS, Aleksi Kallio, CSC Finland

As an answer to the complains on usability, Aleksi Kallio described recent advances being added into CHIPSTER, a tool originally developed for microarray data analysis that is now moving into NGS with the integration of analysis and visualization extensions. To top it all, CHIPSTER, developed at the Finish EMBnet node, is an open source tool that can be freely installed and extended.

Sequence clustering (SeedMap and mBED), Des Higgins, University College, Dublin, Ireland

As other speakers had mentioned by now, the analysis of an individual species genome is but the first step in a long ladder to understand the tower of life. From here, an obvious jump is to compare sequences from many genomes or, given availability of NGS, of tens or thousands of sequences from different species.

The classic program, and probably the most widely used still nowadays for sequence alignment with phylogenetic intent is Clustal. In his talk, Des Higgins described recent methods being developed to deal with the problem of aligning and clustering such large numbers of sequences increasing efficiency and accuracy. After reviewing existing methods and recent developments, Des Higgins described the new methods being developed currently to deal with full genomes or large numbers of sequences.

Metagenomics, Daniel Huson, University of Tubingen, Germany

Once we have a method to cluster sequences we can also classify their respective organisms and move on to the thriving field of metagenomics and the problemof taxonomical and functional analysis. Daniel Huson introduced the problem of studying DNA of uncultured organisms (metagenomics) and the metagenome as the cumulative genetic information of a community of organisms.

Metagenomics can address many new problems but poses additional challenges as well. Thanks to NGS it is now possible to characterize the distribution of species in a community and to compare metagenomes to analyze populations and their evolution and changes in response to environment. And beyond: we can not only identify species, we can also identify the genes present in a community and understand its behavior as a coherent whole. Thanks to tools like MEGAN it is now becoming increasingly easy to perform these studies once all the preparatory comparison work has been carried out.

EGI LS-SSC, Tristan Glatard, CNRS Clermont-Ferrand, France

Despite its obvious interest, most of the preceding talks relied heavily on similar preparatory work requiring treatment and analysis of huge numbers of sequences, a task usually requiring unprecedented computing power rarely available in common scientific institutions. It is here where all the work that has been invested in Grid computing can play a major role.

Tristan Glatard introduced next EGI LS-SSC, part of a bit project proposal to EU, ROSCOE, which will try to provide support for Life Sciences in the Grid. He gave examples of what can be done and what is being developed on the Grid illustrating its significant potential to bring scientists the computing power they need.

Where from here? Open discussion.

The formal workshop talks closed with an open discussion where developers, users and in general all participants were able to exchange opinions, put forward projects and ideas and provide feedback and interactions that we hope will prove useful in the near future to orient work. Most significantly, this session may become the seed for a new community with common interests that we hope will grow healthier and stronger in the future.

Figure 3. Hack-a-thon session: Laurent Falquet.

The hack-a-thon.

While the formal content of the workshop terminated with the presentations, the associated activities did not: for all participants interested, we also organized a practical hands on session where a sample problem in de novo sequencing was to be resolved using various different combinations of experimental datasets and methodological approaches.

The hack-a-thon, directed by Laurent Falquet, gave participants first hand experience on solving a real world problem using unpublished data, allowing them to compare their results with those of experts in the field and to demonstrate that each working group found another result dependent on which parameter settings were used - there are many of them - and to gather a direct feel of the challenges posed by NGS data analyis.

Figure 4. Participants at the hack-a-thon session.

Conclusion

The general opinion among all participants in the workshop was that it was a successful meeting, that covered in depth and breadth the main topics related with the status of the art in NGS data analysis, with an excellent balance in dealing with the many dimensions of the problem (hardware, data analysis, algorithms, tools, practical applications, real world problems and hands-on experience).

This has been a pioneering Workshop in this topic that has been able to achieve a broad coverage of the field with major scientists, allowing us to collect common points, advice and expertise. We intend to reflect all this knowledge in a white paper that will be available with all information we gathered during this workshop on the Next Generation Sequencing web site www.nextgenerationsequencing.org.

In addition, the organizers hope that it will have served as a meeting point for major workers and developers in the field, where they have been able to exchange opinions and experiences, find points in common and gather user feedback, helping them put their work in context and derive useful information to plan their work ahead. This process will be supported with a NGS forum soon available on our NGS web site.t

Finally, we all hope that this proves a useful seed for an emerging community that will grow from the original participants to a large group of professionals sharing a common interest and passion in science.

Acknowledgements

We want to thank first and foremost the institutions that have made this workshop possible through their support, specially, the European Commission through its funding for 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 LHSG-CT-2004-512092), UPPMAX (the Uppsala Multidisciplinary Center for Advanced Computational Science), CASPUR (Consorzio interuniversitario per le applicazioni di supercalcolo per università e ricerca), the Italian Society of Bioinformatics, and EMBnet (the European Molecular Biology Network).

The organizers want to express their deep gratitude to all the invited speakers who graciously accepted to participate and contributed to make this a most interesting and scientifically sound event, and, of course, all the participants who, through their enthusiasm and interest made this a successful smash hit.

Finally, we want to apologize to all the scientists that could not be admitted to the workshop due to lack of resources and available space despite their earnest and keen expressions of interest. We are already planning new workshops to adapt to the increasing demand.

High Throughput Sequencing and the IT architecture

Part 1 : Volume dimensioning and filesystems

George Magklaras

The Biotechnology Centre of Oslo, The University of Oslo (Biotek - UiO), Oslo, Norway

Improvements in DNA sequencing technology have reduced the cost and time of sequencing a new genome. The new generation of High Throughput Sequencing (HTS) devices has provided large impetus to the life science field, and genome sequencing is now a necessary first step in many complex research projects, with direct implications to the field of medical sequencing, cancer and pathogen vector genomics, epi- and meta-genomics.

However, despite the falling sequencing cost and time-lines, there are other associated costs and difficulties in the process of maintaining a functional data repository on large-scale research projects. The new generation of HTS technologies [1] has introduced the need for increased data-storage technologies whose capacity is well beyond the average local data-storage facilities [2]. In fact, the computing world has produced a new term for this paradigm, that of data-intensive computing [2a]. Data-storage costs are falling; however, a study of the functional specifications of popular HTS equipment, such as Roche's 454 pyrosequencers [3], Illumina's hardware [4] and ABI SOLiD technology [5], suggests that a single high-throughput experiment run creates several Tbytes of information. If one takes into account that genome sequencing is often performed repeatedly in order to study genetic variation [6], the capacity of a suitable data-archiving facility needs to scale to several Petabytes of information, which is well beyond the scale of most group, departmental or university computing facilities.

Storage of the data is only one of the technical problems. The distribution and post-processing of large data-sets is also an important issue. Initial raw data and resulting post-processing HTS files need to be accessed (and perhaps replicated), analyzed and annotated by various scientific communities at regional, national and international levels. This is purely a technological problem for which clear answers do not exist, despite the fact that large-scale cyber infrastructures exist in other scientific fields, such as particle physics [7]. However, genome sequence data have slightly different requirements from particle physics data and thus the process of distributing and making sense of large data-sets for Genome Assembly and annotation requires different technological approaches at the data-network and middleware/software layers. For instance, security concerns for clinical HTS settings are an issue, as genomic information concerning patients is really a patient record and thus needs to be addressed in concert with hospital IT and security procedures, subsequent to institutional security compliance procedures. Moreover, other field-relevant procedures, such as the de novo genome assembly of vertebrate-size genomes, require an unusually large amount of RAM per processor/core (more than 500 Gigs of RAM), which may be a challenge, depending on the budget size of the HTS facility and the expertise required for running large shared-memory computers.

This series of articles will discuss and attempt to address all of these challenges by means of flagging various existing and emerging IT technologies. In this first part, we will examine a strategy to plan for the amount of disk space you need to store and archive HTS data, and look at various choices for one of the most critical modules of an IT storage infrastructure: the file-system.

Handling HTS data volumes is a classic example of data-intensive computing [8]. One of the most important aspects of handling a data-intensive computing problem is to understand the amount of data you will be dealing with.

Table 1 provides an indicated maximum (sizes vary depending on the exact sequencing parameters and the experiment) data volume per device type on an annual basis. All HTS devices have a simple conceptual pipeline. Each stage

of the pipeline indicates a storage tier. Each storage tier represents different storage functional requirements, in terms of the amount and access pattern of storage needed:

- Tier 1: Includes the raw unprocessed data as they come out from the instrument (mostly images). For most HTS devices, Tier 1 data will generate several Tbytes per run (several thousands of Gigabytes), especially as the instrument's ability to become more precise gets better with time (firmware or device upgrades). This type of storage acts as a front stage area and needs maximum I/O performance, as concurrent disk write and read operations occur most of the time: write ops occur from the HTS devices; read ops are essentially copies of the raw data by the analysis nodes. Normally, the HTS workstations offer local high-performance disk drives to accommodate these requirements per instrument (DAS, Fiber Channel). When the initial sample run is complete, these data need to be moved to the Tier 1 area to clear the local hard drives for more space.
- Tier 2: Initial processing data stage: including base (or colour) calls, intensities and first pass quality scores. These data are currently in the order of several tenths of Gigabytes to 300 Gigabytes per run maximum for certain types of sequencers.
- Tier 3: Includes aligned and analyzed data (alignments of all the reads to a reference or de novo assembly, if required). This can be at least as big as the initial processing stage (Tier 2), as the initial reads themselves have to be preserved as part of the alignment output. At the end of each successful processing step, the raw data of Tier 1 are removed.
- Tier 4: The final, fourth tier includes data that should be backed up off site, in order to provide for disaster recovery, as well as a long-term archive. This includes a mirror of Tiers 2 and 3, plus the archive requirements. It is not financially feasible or technically practical to off-site backup Tier 1 data, at least not for every run, as the volume of data is huge. There is some data redundancy between Tiers 2 and 3, as in theory one could resort to Tier 3 reads according to the alignment output and then discard Tier 2 data. However, this might not be feasible/desirable in all analysis scenarios, and thus we assume it is good practice to backup and archive both Tier 2 and Tier 3 data.

Tier 1 could be implemented as a disk redundant (RAID 1, 6, other) data-storage area with capacity given by the following equation:

$$Tier1_{store} = \sum (N_{hts} \times G_{bpr} + (N_{hts} \times G_{bpr})/4)$$

$$N_{hts} = number\ of\ per\ type\ HTS\ devices,$$

$$G_{bpr} = Gigabytes\ per\ run$$

The (Nhts x Gbpr)/4 factor represents a small recommended buffer to accommodate unexpected stalls of the HTS pipeline (loss of computing nodes, problems in copying/referencing Tier 1 data, etc).

Tiers 2 and 3 can occupy a common set of disks to form the analysis staging area, according to the following equation:

$$Tier2,3_{store} = \sum (N_{runs} \times G_{analysis} + (N_{runs} \times G_{analysis})/3)$$

$$N_{runs} = expected\ number\ of\ runs\ per\ year,$$

$$G_{analysis} = Gigabytes\ per\ run\ for\ Tiers\ 2\ and\ 3\ (Table\ 1)$$

Finally, Tier 4 backup and storage requirements depend on the data retention policies. We as-

Table. 1: Associating HTS devices and data volumes

HTS Device	No. of runs per year	Tier 1 Data/run (Gbytes)	Tier 2 Data/ run (Gbytes)	Tier 3 Data/ run (Gbytes) (Analysis)	Tier 4 Data/run (Gbytes) (Backup and archive)	Produced data per year (Tbytes)
Illumina	100	9728	100	300	400	990
454	100	200	50	25	75	27
SOLiD	100	6144	100	100	200	80

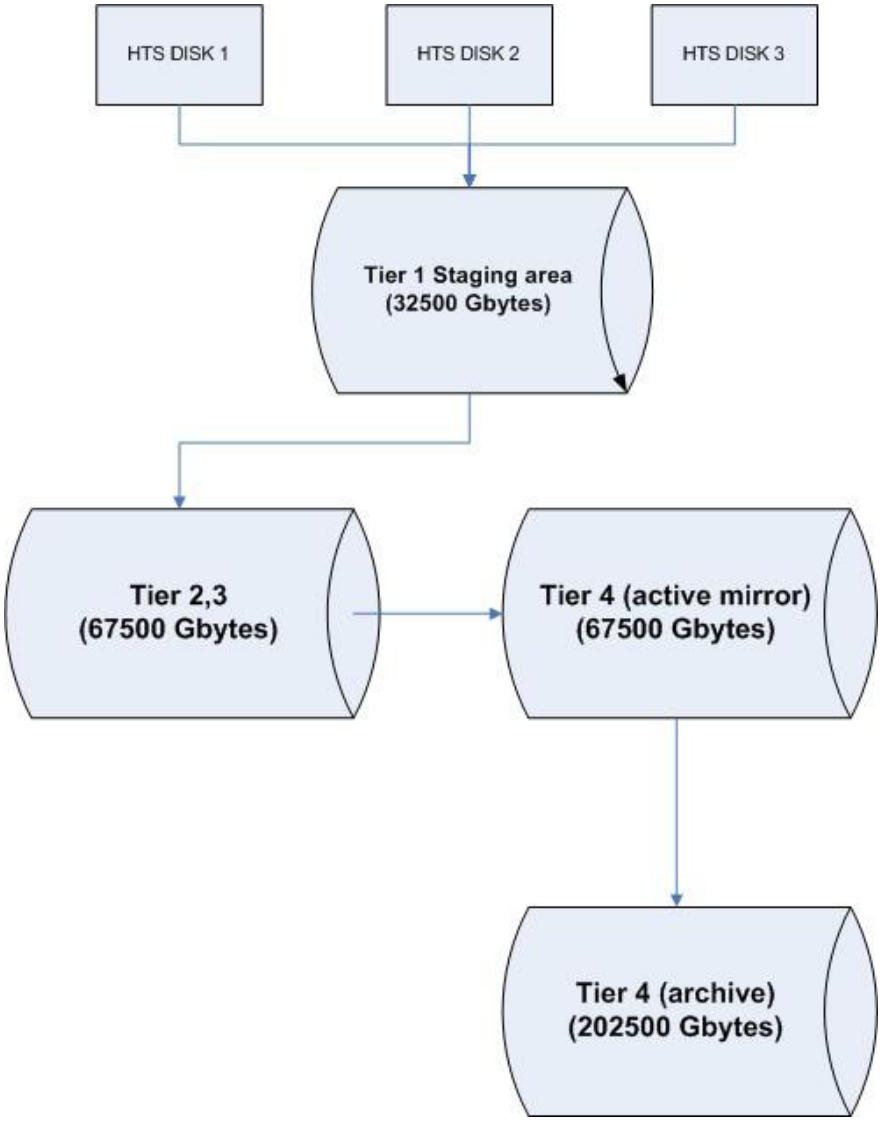

Figure 1. Breaking down the storage requirements per tier.

sume that processed experimental data should be kept for a number of years, thus:

$$Tier4_{store} = Tier2,3_{store} + R_{period} \, x \, Tier2,3_{store}$$

$$R_{period} = number\ of\ years\ to\ keep\ the\ data$$

Based on these equations, and Table 1 requirements, for an example that includes 2 Illumina sequencers, a couple of 454, a single SOLiD machine and a retention period of 3 years, we need a tiered infrastructure as shown in Figure 1. Note how Tier 4 is broken down to the active mirror (facilitates disaster recovery) and the archive, which could include slower Disk-to-Disk (D2D) or tape robot archive solutions.

This methodology shows how easily data volumes can scale to large data-sets. The tiered architecture allows one to scale different storage requirements independently.

Whilst the previous paragraphs quantified the amount of data produced by an HTS facility and the volumes, it is really important to consider how to build those volumes. It becomes clear that handling the size of the HTC data-sets can be an IT engineering challenge. Traditional High Performance Computing (HPC) has addressed the issues of tackling large data-sets by introducing a number of technologies (queue batch/grid

systems, specialized processor interconnects, parallel programming). Not all of them are adequate to address the challenges of large HTS data-sets. In particular, two IT areas need a significant engineering overhaul to handle the volume of HTS data efficiently.

A file-system is a key component of the Operating System that dictates how the files are stored and accessed. Weaknesses in the file-system ability to handle large data-sets can severely affect HTS operations.

Today's common file-system choices include options such as ext3 [9] on Linux platforms, NTFS [10] for Windows platforms, and HFS+ [11] for Apple-based computers. They perform well with common desktop storage loads and with certain multi TiB scenarios on a file server. However, they do not scale well up to hundreds of Terayte or Petabyte-scale scenarios, either because of specific file and volume limits (ext3) and limits in handling efficiently a large number of small files in deeply nested directory structures. Finally, there are issues with concurrent access to a set of files and directories, a common scenario in large HPC clusters.

For these reasons, HPC setups have employed different file-system varieties to satisfy large and concurrent access file scenarios. The first category is commonly referred to as 'clustered' or 'shared' disk file-systems. Some commonly employed examples of clustered file-systems include IBM's General Parallel File System (GPFS) [12], Apple's XSAN[13] and other commercial offerings, such as Isilon's OneFS solution [14]. There is a great variety of academic HPC setups that already run large GPFS installations, handling concurrency and high availability with varying degrees of success. Isilon's system solutions also had varying degrees of success acting as a Tier 1 storage solution for some large academic setups.

Most (if not all) clustered file-system implementations assume the existence of a block-device layer, in order to connect client devices and backend storage disks. So, they offer the illusion of a backend storage device appearing as local disk to a system, provided that a Fiber Channel (FC) or iSCSI based solution is employed. This is a highly desirable functionality on the Tier 1 storage area. For example, instead of employing common utilities such as FTP/Rsync to move data off the HTS device to the Tier 1 area, the same thing could be performed by a simple file level copy operation. Depending on the purity of the block-layer protocol (pure FC will not include TCP/IP overhead), copying multiple TiBs of data from the instrument disks to the Tier 1 staging area could be performed more efficiently. Latter paragraphs will touch on a promising technology to further simplify this process (FCoE).

The second (and less commonly employed) category of file-systems is referred to as distributed file-systems. A distributed file-system allows access to files located on another remote host as though working on the actual host computer. In contrast to clustered file systems, the file locality illusion now is facilitated by means of network-layer code and not by block-layer device emulation. NFS and the emerging pNFS standard [15] are commonly employed in many bioinformatics setups; however, the protocol suffers from scalability and data-redundancy problems. To address these problems, a new generation of distributed fault-tolerant and parallel file-systems exists. Two notable examples of such file-systems are the Sun's Lustre file-system [16] and the Hadoop Distributed File System (HDFS) [17].

It is sometimes difficult to distinguish between certain features of clustered and distributed file-system solutions, as the feature-set of new generation distributed file-systems is expanding. However, one notable difference between traditional cluster file-systems and new distributed file-systems is that the latter are explicitly designed to handle data-sets in the order of several Petabytes, something that might not be entirely true for most of the previously mentioned cluster file-systems. In that sense, both Lustre and HDFS are highly suited to facilitate the file-system storage of Tiers 2 and 3, and at least the active mirror of Tier 4. Which of the two is more suitable is a matter of choice and experimentation. Tailored solutions that lead to definite conclusions do not yet exist, and understanding the pros and cons of these two platforms is work in progress.

The most notable differences between Lustre and HDFS is that the second is less mature in generic production environments and requires a substantial effort to express the data-set using the map/reduce [18] concept. HDFS is part of a data-processing framework and as such most data-sets must be converted before they can take advantage of its power. There have been successful attempts to employ the Hadoop

framework on bioinformatics problems [19]. On the other hand, Lustre is a much more file-oriented storage framework. It also requires substantial effort to setup, especially when it comes to converting data from older Lustre versions, but it does not require any explicit data-set conversion.

I hope that I have given your storage architects plenty of food for thought. In the next article, I shall look at the data-network layer.

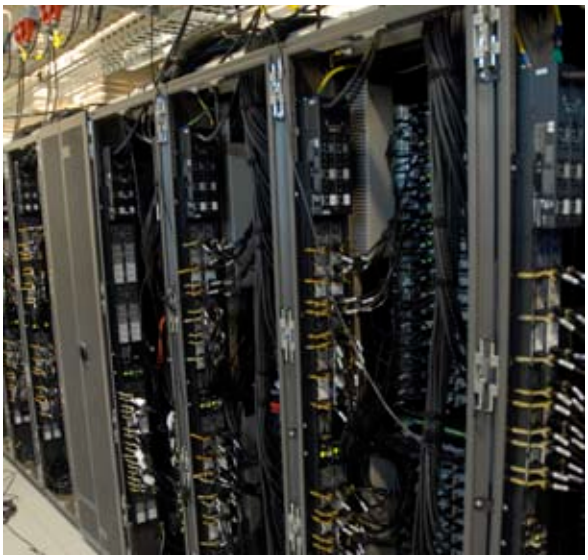

Figure 2. A computer cluster.

References

1. "Genome sequencing: the third generation", Published online 6 February 2009 | Nature | doi:10.1038/news.2009.86
2. "Big data: Welcome to the petacentre", Published online 3 September 2008 | Nature 455, 16-21 (2008) | doi:10.1038/455016a

2a. Gorton I., Greenfield P., Szaley A., Williams R. (2008), "Data Intensive Computing for the 21st Century, IEEE Computer, April 2008 issue.

3. The 454 sequencer product line specifications: http://www.454.com/products-solutions/product-list.asp
4. Illumina Genome sequencers product line specifications: http://www.illumina.com/pagesnrn.ilmn?ID=26
5. Applied Biosystems SOLiD system specifications: http://www3.appliedbiosystems.com
6. 1000 genomes project portal: http://1000genomes.org/page.php
7. The worldwide Large Hydron Collider (LHC) Computing Grid portal at CERN: http://lcg.web.cern.ch/LCG/
8. Kouzes et al (2009), "The changing paradigm of data-intensive computing", Computer, January 2009 issue, IEEE Computer Society, pages 26-34.
9. Stephen C. Tweedie (May 1998). "Journaling the Linux ext2fs Filesystem" (PDF). Proceedings of the 4th Annual LinuxExpo, Durham, NC. http://jamesthornton.com/hotlist/linux-filesystems/ext3-journal-design.pdf.
10. The NTFS wikipedia page: http://en.wikipedia.org/wiki/NTFS
11. "Technical Note TN1150: HFS Plus Volume Format". Apple Developer Connection. March 5, 2004. http://developer.apple.com/technotes/tn/tn1150.html.
12. Schmuck, Frank; Roger Haskin (January 2002). "GPFS: A Shared-Disk File System for Large Computing Clusters" (pdf). Proceedings of the FAST'02 Conference on File and Storage Technologies: 231-244, Monterey, California, USA: USENIX. t
13. "Apple Ships Xsan Storage Area Network File System". Apple Inc.. http://www.apple.com/pr/library/2005/jan/04xsan.html.
14. Isilon Systems OneFS product literature: http://www.isilon.com/products/index.php?page=software_OneFS
15. The wikipedia entry on NFS: http://en.wikipedia.org/wiki/Network_File_System_(protocol)
16. The Lustre filesystem community Wiki: http://wiki.lustre.org/index.php?title=Main_Page
17. The Apache Hadoop project: http://hadoop.apache.org/core/
18. J. Dean and S. Ghemawat, 'MapReduce: Simplified DataProcessing on Large Clusters,' in OSDI 2004: Sixth Symposium on Operating System Design and Implementation, 2004. [Online]. Available: http://labs.google.com/papers/mapreduce-osdi04.pdf
19. Simone Leo et al (2008), 'Parallelizing bioinformatics applications with MapReduce', poster presented at the Cloud Computing Applications 2008 Conference, available at: http://www.cca08.org/papers/Poster10-Simone-Leo.pdf

mint condition

Vivienne Baillie Gerritsen

It is very likely that mint – and its close cousin menthol – is one of the most popular flavours or sensations known worldwide. Is there any population left on Earth that hasn't sucked a mint sweet or chewed on mint gum? Mint is drunk in beverages, and brushed onto teeth. Added to sauces, and put into chocolates. Smeared onto chests and added to paper handkerchiefs. Why is it that mint and menthol are found, one way or another, almost everywhere on this planet? Transport would be an obvious answer. But there is more to it than that. Besides the numerous health benefits, mint – and menthol – have a quality that is readily appreciated by many: freshness. This sensation is the legacy of two kindred proteins – P450 cytochromes – found in mint plants.

Mentha

by Fir0002/Flagstaffotos

copyleft license
http://en.wikipedia.org/wiki/GNU_Free_Documentation_License

The virtues of mint plants have been appreciated for millennia, and like the great majority of medicinal herbs, the mint plant is named after a Greek mythological character: the nymph Minthe. Persephone was jealous of Pluto's love for Minthe, so she promptly transformed her into a plant. Unfortunately, Pluto was not able to restore Minthe to her former state but assured her that she would not be forgotten since her fragrance would be distinctive and pleasant. Especially when she was trod upon… Minthe became a Mediterranean weed whose benefits were widely acknowledged. Dried mint leaves have been found in Egyptian tombs. The Romans used it extensively and introduced the plant to Great Britain on one of their visits. The British then introduced it to many parts of the world as they colonized different parts of it.

The two most popular mint plants are spearmint (*Mentha spicata*) and peppermint (*Mentha piperata*). Known for literally thousands of years, their essential oils are used to treat numerous ailments, such as headaches, indigestion, diarrhea, motion sickness, colds, gallstones and infections to name a few. What is it that does us so much good? The answer is menthol and carvone. Spearmint hosts the enzyme limonene-6-hydroxylase which is involved in the production of carvone – the chemical entity which gives the well-known spearmint flavour. The peppermint plant, on the other hand, hosts limonene-3-hydroxylase, the enzyme involved in the production of menthol. Both limonene hydroxylases belong to the large P450 cytochrome family whose members all have a central role in producing thousands of natural plant products amongst which the hundreds of oxygenated monoterpenes – to which belong carvone and menthol – that are the source of the aromas and flavours so particular to specific essential oils.

Carvone and menthol are end products following the hydroxylation – by the spearmint and peppermint hydroxylases respectively – of one same chemical entity: limonene. Limonene-6-hydroxylase hydroxylates limonene on C6 thus producing trans-carveol which is subsequently modified to become carvone. Limonene-3-hydroxylase, however, hydroxylates limonene

on its C3 thus producing trans-isopiperitenol which – five steps later – is modified to become menthol. The two enzymes are very similar, and their substrate binding sites very restrictive – a discovery which came as a surprise to scientists. Indeed while, as a rule, in the P450 cytochrome family any change of activity usually requires a certain number of mutations, only one mutation is needed to modify the limonene hydroxylases' binding activity.

This particular mutation converts a phenyl-alanine into an isoleucine in the sequence of the spearmint hydroxylase. Originally a limonene-6-hydroxylase, this phenylalanine to isoleucine mutation causes the spearmint enzyme to become a limonene-3-hydroxylase! The spearmint enzyme is thus capable of synthesizing menthol like its cousin, the peppermint hydroxylase! Such a mutation points to the fact that these particular amino acids are not only essential but are most probably involved in the orientation of the substrate limonene within the binding pocket so that it is hydroxylated either at position C3, or position C6.

Single mutations which are capable of changing so drastically a protein's function are of great interest in the world of research. Not only do they point to very specific minute regions in a protein's sequence, but they provide valuable information for the understanding of instances such as substrate binding, substrate orientation, pocket binding structure, enzyme function and metabolic pathways. Needless to say, they are of high biotechnological interest. In the case of the limonene-3- and limonene-6- hydroxylases for instance, the study implies that their substrate binding pockets must be small and pretty tight, and one mutation is capable of influencing substrate orientation in a very subtle way. Naturally, such studies are of great importance within the world of commerce for the yield of peppermint oil, for example, by way of the genetic engineering of *E.coli* or yeast. Nothing many of us would complain about; it is so rare to be able to enjoy something with the knowledge that it is also good for you.

Cross-references to UniProt

Cytochrome P450 71D15, *Mentha piperita* (Peppermint) : Q9XHE6
Cytochrome P450 71D18, *Mentha spicata* (Spearmint) : Q9XHE8

References

1. Schalk M:, Croteau R.
A single amino acid substitution (F363I) converts the regiochemistry of the spearmint (–)-limonene hydroxylase from a C6- to a C3-hydroxylase
PNAS 97:11948-11953(2000)
PMID: 11050228

2. Wust M., Little D.B., Schalk M., Croteau R.
Hydroxylation of limonene enantiomers and analogs by recombinant (–)-limonene 3- and 6-hydroxylases from Mint (*Mentha*) species: evidence for catalysis within sterically constrained active sites
Archives of Biochemistry and Biophysics 387:125-136(2001)
PMID: 11368174

3. Lupien S., Karp F., Wildung M., Croteau R.
Regiospecific cytochrome P450 limonene hydroxylases from Mint (Mentha) species : cDNA isolation, of (–)-4S-limonene 3-hydroxylase and (–)-4S-limonene 6-hydroxylase
Archives of Biochemistry and Biophysics 368:181-192(1999)
PMID: 10415126

National Nodes

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

Australia
RMC Gunn Building B19, University of Sydney, Sydney

Belgium
BEN ULB Campus Plaine CP 257, Brussels

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

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

China
Centre of Bioinformatics, Peking University, Beijing

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

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

Cuba
Centro de Ingeniería Genética y Biotecnolgía, 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

IHCP
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.news

ISSN 1023-4144

Dear reader,

If you have any comments or suggestions regarding this newsletter we would be very glad to hear from you. If you have a tip you feel we can 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 of EMBnet.news are available as PDF files. You can get them from the EMBnet organization Web site:
http://journal.EMBnet.org/index.php/EMBnetnews/issue/archive

Publisher:
EMBnet Executive Board
c/o Erik Bongcam-Rudloff
Uppsala Biomedical Centre
The Linnaeus Centre for Bioinformatics, SLU/UU
Box 570 S-751 23 Uppsala, Sweden
Email: erik.bongcam@bmc.uu.se
Tel: +46-18-4716696

Submission deadline for the next issue:
February 20, 2010