


EMBnet.news

Volume 15 Nr. 3
October 2009

- 
- **Bioclipse 2: towards integrated biocheminformatics**
 - **Ensembl: A New View of Genome Browsing**
 - **Linux distributions for bioinformatics: an update and more ...**

Editorial

With the introduction of Next Generation (NextGen) Sequencing technologies to the Life Sciences arena, the volume of sequence data being created is growing at an astonishing rate. Sharing data, hardware and human resources is essential for the effective use of these data. EMBnet, as one of the largest and still growing network of bioinformatics organizations, is aware of the user community needs and is actively working on this issue with the collaboration of other bioinformatics networks worldwide. A fruitful cooperation exists with the Iberoamerican (RIBIO) and the Asia Pacific (APBioNet) bioinformatics networks as well as an affiliation to the US based International Society for Computational Biology (ISCB). Close contacts have also been established with the African Society for Bioinformatics and Computational Biology (ASBCB) and with the Southern African Network for Biosciences (SANbio). EMBnet's role in teaching and as a provider of knowledge of bioinformatics tools, solutions and on how to set up and maintain local bioinformatics databases is also an increasing activity. All these activities are reflected in the contents of several articles in this issue.

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



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

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

We provide the EMBnet community with a printed version of issue 108. Please let us know if you like this inclusion.

Cover picture: Peacock Butterfly, *Inachis io*. July 2009, Uppsala, Sweden [© Erik Bongcam-Rudloff]

Contents

Editorial	2
Letters to the Editor	
Status of Bioinformatics in Southern Africa: Challenges and Opportunities	3
News and Announcements	7
Reports	
SYNAPSES: Bridging the gap between Biologists and Bioinformaticians	8
Personal Account: Training in Grid Computing for Bioinformatics	11
Technical Notes	
PairsDB protein alignment atlas - interface and database tables	13
A More elaborative way to check codon quality: an open source program	18
Designing Primer Pairs and Oligos with OligoFaktorySE	22
Bioclipse 2: towards integrated biocheminformatics	25
Ensembl: A New View of Genome Browsing	28
Reviews	
Linux distributions for bioinformatics: an upd. ..	35
Protein spotlight 108	42
Node information	44

Editorial Board:

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

Email: erik.bongcam@bmc.uu.se

Tel: +46-18-4716696

Fax: +46-18-4714525

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

Email: domenica.delia@ba.itb.cnr.it

Tel: +39-80-5929674

Fax: +39-80-5929690

Pedro Fernandes, Instituto Gulbenkian. PT

Email: pfern@igc.gulbenkian.pt

Tel: +315-214407912

Fax: +315-214407970

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

Email: klucar@embnet.sk

Tel: +421-2-59307413

Fax: +421-2-59307416

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

Email: andreas.gisel@ba.itb.cnr.it

Tel: +39-80-5929662

Fax: +39-80-5929690

Status of Bioinformatics in Southern Africa: Challenges and Opportunities



Yasmina Jaufeerally-Fakim¹, Daneshwar Puchoo¹ and Luke Mumba²

¹ University of Mauritius, SANBio Bioinformatics Node, Mauritius

² Director Southern African Network for Biosciences (SANBio), New Partnership for Africa's Development (NEPAD), South Africa

Background

The Southern African Network for Biosciences (SANBio) includes twelve countries in Southern Africa namely Angola, Botswana, Malawi, Mauritius, Mozambique, Namibia, Lesotho, Swaziland, Seychelles, South Africa, Zambia and Zimbabwe. In 2005 SANBio started its function to promote Science and Technology for the benefit of the people of the region. SANBio draws its programme of work from the Africa's Science & Technology Consolidated Plan of Action (CPA) [1]. The CPA was adopted in 2005 by the African Ministerial Conference on Science and Technology (AMCOST) at its second conference in Dakar, Senegal. The same document was endorsed a year later by the AU Summit in Khartoum, Sudan. The CPA articulates Africa's common objective of socio-economic transformation and full integration into the world economy. It reaffirms the continent's collective action for using S&T for meeting the developmental goals of Africa with key pillars being capacity building, knowledge production and technological innovation. The CPA recognizes that S&T in Africa is plagued by such factors as weak or no links between industry and S&T institutions, a mismatch between R&D activities and national industrial development strategies and goals. The consequence of these weaknesses is that research findings in public institutions, including universities, do not get accessed and used by local industries es-

pecially small and medium enterprises. The CPA comprises of three key areas: research and development programmes; improvement in policy conditions and building innovation mechanisms; and implementation, funding and governance strategies.

CPA R&D Programmes and Implementation

The programmes contained in the CPA are implemented through regional networks of centres of excellence, consisting of hubs and nodes. The objectives of these networks are: to improve quality of and access to infrastructure and facilities; develop further institutional and political regulations; improve the human skill base; obtain political and civil society support; strengthen the capacity of regional institutions; integrate R&D into sectoral programmes; improve the applicability of S&T towards the Millennium Development Goals and Sustainable Development; and to develop innovative funding instruments and build international partnerships. Research and Development Programmes of the CPA consists of five clusters. Under each cluster there are several programmes. The clusters are: *Cluster 1: Biodiversity, Biotechnology and Indigenous Knowledge*. This cluster focuses on the conservation and sustainable use of biodiversity; safe development and application of biotechnology; securing and using Africa's indigenous knowledge base. *Cluster 2: Energy, Water and Desertification*. This includes building a sustainable energy base by increasing rural and urban access to environmentally-sound energy sources and technologies; securing and sustaining water to ensure sustainable access to safe and adequate clean water supply and sanitation; combating drought and desertification by improving scientific understanding and sharing of information on the causes of and extent of drought and desertification in Africa. *Cluster 3: Material Sciences, Manufacturing, Laser Technology and Post-Harvest Technology*. This includes the development of new and improvement of existing infrastructure by building new skills or expertise in material sciences, promoting the sharing of physical infrastructure and exchange of scientific information and the promotion of public sector partnerships in material sciences research and innovation. *Cluster 4: Information and Communication Technologies; and Space Science and Technologies*. This in-

cludes the creation of experts engaged in computer science, information systems as well as informatics; building skills in software research and development. It also includes the establishment of the African Institute of Space Science. *Cluster 5: Mathematical Sciences*. This includes the establishment of an African Mathematical Institutes aimed at strengthening the African Mathematical Institutes network that was constituted in 2005 with the sole purpose of building a new generation of African scientists and technologists with excellent quantitative problem-solving skills.

SANBio operates from its secretariat based at the Council for Scientific and Industrial Research (CSIR) campus in Pretoria, South Africa. It has been active in the setting up of regional projects whereby several countries participate. Among those are: The scientific validation of traditional medicines for HIV treatment, conservation and utilization of plant genetic resources in the SADC (Southern African Development Community) region for food security, Fish biodiversity of inland rivers of Southern Africa, Mushroom production, Livestock Development and Capacity building in bioinformatics. Each project is managed by a country node under the supervision of the SANBio Director.

Capacity Building in Bioinformatics

The University of Mauritius is the regional node for the project on capacity building in Bioinformatics. The main objective of the action is to provide training to scientists in the countries of Southern Africa for the development and utilization of bioinformatics in research institutions and universities. It will create a platform whereby research in biosciences will be strengthened through the incorporation of bioinformatics within the current activities. Such an initiative will have a regional impact and improve the perspectives of the research output.

Conventional approaches for research projects that are under way in the region in biological sciences are based on methods of molecular biology, biochemistry and biotechnology. Most of the universities have well-developed departments for these subjects. They also have their computer science departments and this action will bridge them with the aim of setting up facilities for bioinformatics. The justification for such a project is the fact that Southern Africa has yet to tap genomics and expression

data that have become available in the recent past. Most other regions of the world are actively utilising and applying such information for bringing solutions to their priority problems. Southern African countries face massive difficulties in tackling infectious diseases, food production, animal health, environmental degradation and population growth. Malawi has several scientific set ups where research is being undertaken in molecular epidemiology of HIV, malaria and other human diseases [2,3]. The School of Veterinary Medicine of the University of Zambia and the Livestock and Pest Research Centre of Zambia are conducting a major work on the tick-borne diseases such as Trypanosomiasis. More than 75 % of the livestock are found in areas where the tsetse fly, which is the vector for this disease, can be found. Current work there include epidemiology of human trypanosomiasis, comparative studies of cysteine proteases of *Trypanosoma brucei* and *T. congolense*, characterisation of Theileria parasites and immunogenicity of Theileria sporozoites.

The Theileria genome information is available [4] and the scientists in Zambia will benefit from training in using genome browsers and other databases for extracting useful information on this organism. Works on sequencing other Theileria species are also under way at ILRI (International Livestock Research Institute) in Kenya. Tapping genomic and expression data, which have become available for many parasites, can lead to more efficient control of diseases through the design of better vaccines and drugs.

An important fish biodiversity project is under way in Malawi. The lakes of Southern Africa are heavily exploited for food production. Lake Malawi is known to house five hundred endemic fish species and follow up of decrease in stocks is necessary to assess the probable loss in biodiversity. DNA sequences for phylogenetic comparison are being used for better diversity management of the fish resources of the seven main rivers of Southern Africa which are the Zaire, Zambezi, Okavango, Limpopo, Orange, Ruvuma and Cunene. They run through eleven countries and make up 6.7 km² of catchment area.

The region also hosts a major project on the Conservation of Plant Genetic Resources. The SADC Plant Genetic Resources Centre (SPGRC) is located in Zambia and the project aims at strengthening capacities in other countries for: Assessment of genetic diversity, identification of

important genetic traits, conservation measures, utilisation and food security and sustainable livelihoods.

At the University of Mauritius, a bioinformatics group has been set up to evaluate the current needs and to initiate research. The group is presently working on microbial genomes comparison [5].

South Africa stands as an opportunity for neighbouring countries, as bioinformatics is well implemented among scientific institutions and universities. High quality research output is evidence for this [6-9]. A strong network has been set up and is functioning with many nodes which provide short courses.

The specific objectives of this project are:

- setting up of a network of scientists from the Southern African region,
- organise core facilities,
- Initiate the development of courses in bioinformatics,
- provide specific training to academics and students,
- assist research students in obtaining short-attachment in bioinformatics centres.

Scientists in the region have high hopes that training in the utilisation of bioinformatics resources and tools will enable them to enhance the quality of teaching and research being undertaken. Above all, such tools will accelerate development of diagnostics, vaccines and other medical and veterinary products that are directly relevant to national priorities.

An advisory working group has been set up and comprises of the following members:

- Prof. Luke Mumba, Director SANBio
- Dr. Y. Jaufeerally-Fakim, SANBio Bioinformatics Node Coordinator
- Mrs. Chimwemwe Chamdimba, M & E Manager, NEPAD Office Science and Technology
- Prof. Fourie Joubert, University of Pretoria
- Prof. Oleg Reva, University of Pretoria
- Dr. Etienne de Villiers, ILRI (International Livestock Research Institute) Kenya
- Prof. Erik Bongcam-Rudloff, The Linnaeus Centre for bioinformatics-SLU, Uppsala, Sweden
- Dr. Eija Korpelainen, CSC-IT Center for Science, Finland

A first meeting was held at the University of Pretoria, Computational Biology department in January 2009. Participants came from Malawi, Zambia, Namibia, Botswana, Mozambique, Tanzania, Mauritius and South Africa. A preliminary needs assessment was conducted to work out the types of training that will be required. The project planning meeting was held in Mauritius last July and the series of activities to follow was finalised. The areas for training were identified as follows:

1. *Genome Browsers*: an introduction to a browser (e.g., Human or Rice genome browser (TIGR), ENSEMBL, UCSC), general features, how to understand the display, and work with the features available.
2. *Biodiversity and Evolutionary Genetics*: tools for sequence alignment, motifs searching, finding recombinants (HyPhy/ PAML), phylogenetic programs; how to interpret output from such analysis.
3. *Pathogen genomics*: comparative genomics, how to compare genomes, find homologs or conserved sequences. Using for example the *Plasmodium falciparum* or *Mycobacterium tuberculosis* genomes.
4. *Assembly and Annotation*: tools for genome assembly; how annotation of different genomes is conducted.

It is expected that this initiative will enable more projects to follow and therefore securing extra funding for a flagship project will be an essential part.

Internet Access

The digital divide between Europe and Africa has been addressed by the e-Africa commission



Figure 1. Participants at a SANBio Regional Bioinformatics Training Course, University of Pretoria, South Africa, 2009.

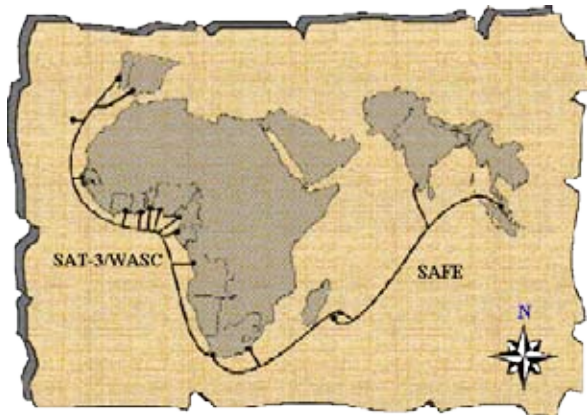


Figure 2. The South Africa Far East cable links Europe with countries of Western and Southern Africa and the Far East.

of NEPAD. The prices of internet access in Africa were among the highest in the world. This is about to change. High speed digital links have been made possible by the deep sea cable SAFE (South Africa Far East) which is 13,104 km and has points in Melkbosstrand and Mtunzini South Africa, St Paul in Reunion Island, Baie du Jacotet in Mauritius, Kochi in India and Penang Malaysia. It has capacity upgradable to 130 Gbits per second. This is important as it will allow easy and fast Internet traffic which is essential for sequence data requiring transfer of very large files.

In July this year the Seacom cable owned by African companies, went live thus bringing high speed connections to South Africa, Mozambique, Tanzania, Kenya, Uganda and Asia. It is 17,000 km and already many universities are benefitting from it (<http://www.seacom.mu/intro.html>).

The Regional Communications Infrastructure Program (RCIP 3) has brought high speed connection to Eastern and Southern Africa through the links from Malawi, Mozambique and Tanzania



Figure 3. Connection through SEACOM.

to the sea cable running up the Eastern coast of Africa.

References

1. Africa's Science & Technology Consolidated Plan of Action (2006). NEPAD Office of Science & Technology. D.S Print Media, Johannesburg. Pp72. ISBN : 978-0620-37633-4.
2. Gaoqian F, Aitken E, Yosaatmadja F, Kalilani L, Meshnick SR, Jaworowski A, Simpson JA, Rogerson SJ (2009) Antibodies to variant surface antigens of Plasmodium falciparum-infected erythrocytes are associated with protection from treatment failure and the development of anemia in pregnancy. *The Journal of infectious diseases* 200(2):299-306.
3. Kalilani L, Atashili J (2006) Measuring additive Interactions using odds ratio. *Epidemiologic Perspectives and Innovations* 3:5.
4. Bioinformatics in Africa [http://www.lirmm.fr/france_afrique/20060708_NEPAD.pdf]
5. Khoyratty Sher-ullah SS, Souza MT Jr., Jaufeerally-Fakim Y (2008) Structural analysis of catalase from two *Musa* accessions FHIA 18 and Williams and from *Ravenala Madagascariensis*. *In Silico Biology* 8: 413-425.
6. Collins NE, Liebenberg J, de Villiers EP, Brayton KA, Louw E, Pretorius A, Faber FE, van Heerden H, Josemans AI, van Kleef M, Steyn HC, van Strijp, Birkholtz LM, Wrenger C, Joubert F, Wells GA, Walter RD, Louw AI (2003) Parasitespecific inserts in the bifunctional Sadenosylmethionine decarboxylase/ornithine decarboxylase of *Plasmodium falciparum* modulate catalytic activities and domain interactions. *Biochem J.* 377:439-448.
7. Klockgether J, Wurdemann D, Reva O, Wiehlmann L, Tümmler B. (2007) Diversity of the abundant pKLC102/PAGI-2 family of genomic islands in *Pseudomonas aeruginosa*. *J. Bacteriol.* 189:2443-2459.
8. Reva, O, Weinel C, Weinel M, Böhm K, Stjepandic D, Hoheisel J, Tümmler B (2006). Functional genomics of stress response in *Pseudomonas putida* KT2440. *J. Bacteriol.* 188: 4079-4092.
9. Wells GA, Birkholtz LM, Joubert F, Walter RD, Louw AI (2006) Novel properties of malarial Sadenosylmethionine decarboxylase as revealed by structural modelling. *J. Mol. Graph. Model.* 24:309-318.



HERSHEY | NEW YORK

IGI Global's newest release
the
**Handbook of Research on
Computational Grid Technologies
for Life Sciences, Biomedicine, and
Healthcare**

edited by Mario Cannataro, University Magna Graecia of
Catanzaro, Italy

Life Sciences Implements Use of
Grid Technology

IGI Global's newest release, the *Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine, and Healthcare*, brings together state-of-the art methodologies and developments of grid technologies applied in different fields of life sciences. This Handbook of Research considers the use of grid technologies to support research and application of each information level where life science research takes place - a useful reference source for academicians, medical practitioners, and researchers involved in all areas of healthcare technologies. Italian EMBnet node members were authors of two chapters: chapter X: High-Throughput GRID Computing for Life Sciences and Chapter XXIX: The LIBI Grid Platform for Bioinformatics.

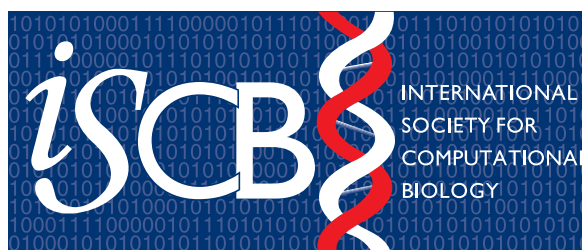
To learn more about this title, please see:

<http://www.igi-global.com/reference/details.asp?id=34292>

For additional information about this publication, to arrange an interview with the editor, or to request a copy for review, please contact Megan Childs at mchilds@igi-global.com.

Order inquiries may be directed to: 717-533-8845 x100, to cust@igi-global.com or to book wholesalers or journal subscription agents.

Contact: Megan Childs **FOR IMMEDIATE RELEASE**
Title: Marketing Communications Coordinator
Address: 701 E. Chocolate Ave., Hershey, PA 17033
Email: mchilds@igi-global.com
Tel.: 717-533-8845 x148
Fax: 717-533-7115
Web: www.igi-global.com



**The EMBnet is delighted to announce
its status of Affiliated Network to ISCB
starting from September 2009**



The Editorial Board is glad to announce that the
EMBnet.news is implementing

**a peer-reviewing process for
selected articles**

in almost all fields of Bioinformatics and
Computational Biology with a particular atten-
tion on "practical and applied bioinformatics".

EMBnet.news will accept submissions for peer-re-
viewed articles starting from November 20th.

Follow our news at the EMBnet web site and stay
informed

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

SYNAPSES:

Bridging the gap between Biologists and Bioinformaticians

2nd international Workshop on Pattern Discovery in Biology, Covenant University, Ota, Nigeria



José R. Valverde¹ and Ezekiel F. Adebiji²

¹ EMBnet/CNB, Centro Nacional de Biotecnología, CSIC. Madrid, Spain

² Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Nigeria

This has been the second Bioinformatics Workshop organized at Covenant University with the aim of bringing together Biologists and Bioinformaticians. The main goal of these workshops is to provide a way for specialists in both fields to get together and exchange knowledge and experience. The workshop is structured around a series of talks that introduce key concepts (of Biology for Bioinformaticians and of Bioinformatics for Biologists) and a series of hands-on practical sessions where relevant tools are demonstrated and tried first-hand by assistants. This basic scheme, provides a nurturing medium where specialists in both areas can talk, interact and get closer, shedding the seed of future collaborations.

The second Workshop on Pattern Discovery in Biology took place in the College of Science and Technology, Covenant University, Ota, Nigeria, from July the 6th to July 11th. Its location, very close to Lagos makes it an ideally well communicated place easy to reach. Covenant University, funded by the World Mission Agency (WMA) Intl Inc strives to become a major leading institution in the world; it is located in the broad, quiet campus of Canaanland and enjoys many modern facilities. This welcoming and relaxing

environment provides a warm environment for Workshop participants to mix and interact as well as the necessary technical support needed to be able to test and get acquainted with the tools demonstrated.

The workshop started with a warm Opening Ceremony where the Vice Chancellor, the Registrar, the Dean, College of Science and Technology and the Chairman, Workshop Steering Committee welcomed all participants and that helped set the key note for the rest of the sessions, stressing the main points of the meeting: easing mutual understanding, closing the interdisciplinary gap and promoting cooperation and lasting collaborations between participants. A keynote address helped introduce the main topics that were covered in the remaining of the sessions, and the leitmotif for the workshop, which this time centred on the need to partially automate using Bioinformatics, the design of new drugs to fight major diseases in the region.

The main body of the Workshop was structured around enlightening talks and practical sessions. The role of the talks was to introduce the key points of major topics both from the perspective of Biologists and the perspective of Bioinformaticians. Since the main goal is to bridge the gap and allow efficient interdisciplinary communications, the contents must go beyond basic introductions and get into enough advanced details to allow mutual comprehension. Reaching this level of detail was the major challenge for this workshop, and imposed relevant restrictions in the way talks needed to be structured (so as to cover both introductory and advanced topics) and the effort required by participants to follow the presentations. These difficulties were eased out by providing enough room for discussion after the talks, where participants could engage



The college of Science and Technology.



Workshop participants at the Closing Ceremony.

in conversations where not only obscure points were illustrated but where practical and personal points of interest and individual applications could be brought up and openly discussed and where free brainstorming was welcome.

The descriptive talks were complemented by a series of practical sessions, in a ratio close to 50%. The practical sessions helped address several complementary goals: first, they served to illustrate the points made during descriptive presentations, further clarifying key concepts and demonstrating how these are experienced in real practice; as a second benefit, they provided first hand personal experience in the use of the tools of the trade and the interpretation of the results they yield, ensuring every participant left with the expertise needed to identify and study biologically significant patterns generated by bioinformaticians and to use them to build predictive models to guide subsequent research; and third, they provided a joint environment for practical interaction among participants from different disciplines, in a space where they could share experiences and exchange comments in an actual, real set up similar to the one they find in their day-to-day work at the lab. To ensure achievement of these goals, a large computer room was used, where every participant could work with one computer (or even more if desired) and where they could freely interact with computers, speakers and their colleagues while working on a practical problem selected from real life.

As a part of our goal of making participants proficient in the field, and ensuring they can continue working once they go back home, we selected to demonstrate major points choosing

whenever possible from a large array of services publicly available on the World Wide Web, as these are the ones they will most likely want to use, and complemented them in parallel with equivalent demonstrations using locally installed tools and programs (e.g. Using Clustal on the web and ClustalX [1] and PhyloWin [2] on the desktop) so they could realize the limitations of web based approaches and the richness of the local applications approach.

In order to ensure participants had access to all the applications needed, we used a 4GB USB key/pen that is continuously being developed at EMBnet/CNB, in Spain[3]. The USB key/pen contains a minimalistic Linux system with a lightweight, yet complete, user interface (Puppy Linux) and can be either booted directly, run as a virtual machine under other system (windows, Mac, Linux) or as a slower emulated machine (if one lacks administrative privileges). The system has been populated with a large number of software packages, including publicly accessible systems like EMBOSS [4] or PHYLIP [5] as well as free -yet usually non distributable- systems like TRITON [6] or MODELLER [7] for which we had previously obtained permission from the authors to include and distribute them for educational purposes. This emulated machine proved to work acceptably well on the resources of the computers available (256MB RAM) even when dealing with complex problems like protein homology modeling, protein-ligand docking and even quantum mechanics modeling. As this key is still a work in progress, we shall not get into more details of its contents, configuration and availability here. The key and its contents were made available



Dr. Jose and Dr. Adebisi with some participants at the beach.



A few participants at the beach.

to participants, and a number of keys were also distributed at the end of the Workshop.

The workshop started with a general introduction to the field of Bioinformatics, putting a large stress in explaining major Biological and Computer Science concepts, the way they are mapped out on each field and the differences in terminology and conceptual approach to a common problem. After these basic concepts were illustrated, participants could move into the major tasks planned for the workshop: the generic problem of pattern discovery in Bioinformatics and its relationship with the close Biological concepts of similarity and homology, and then by extension its many applications in sequence comparison, evolution, genomics, genetic engineering and gene manipulation, proteomics, regulomics, functional prediction, etc..

As we have already mentioned, the leitmotif selected for this second international workshop was the problem of drug discovery and design, and understanding of drug resistance mechanisms. This is a topic of major interest in the region, where multi-drug resistant strains of serious diseases have spread due to a variety of reasons. The second part of the Workshop dealt therefore with these topics in more details, jumping from traditional sequence comparison into three-dimensional pattern discovery, homology modeling, understanding of the interactions between host and microorganism and between enzymes and substrates, and the methods available to analyse drug effects and drugs resistance.

According to the guiding line for the Workshop, the practical sessions were organized around a real-life example: resistance of *Mycobacterium*

tuberculosis to Isoniazid. We selected this example as there is a well understood body of knowledge in the field, it is one of the major diseases both in developed and developing countries, multiple-resistant strains have been characterized and identified as a major health challenge by WHO, and its molecular understanding is still a topic of great current interest. Using as a reference a very recent paper published by Wahab et al. [8] analyzing the mechanism of resistance on *InhA* (PDB 1ZID), we reproduced the work described in the paper on a similar system, the *KatG* ser315thr mutant, which is the most commonly observed [9] and were able to even proceed one step further using Triton [6], Modeller [7], Autodock [10] and Mopac [11].

The results of the Workshop were summarized in the Closing Ceremony, presided by the University Vice Chancellor, Prof. Aize Obayan, and attended by other members of the University's Management and all participants, who had an opportunity to express their feelings about the outcome of this hard working workshop week. The participants at the end of the ceremony received from the Vicechancellor their Certificates of Participation. According to the views expressed at this Closing Ceremony, the workshop was favourably evaluated by participants who asserted its value to bring together both (biologists and bioinformaticians) scientific communities and stir up collaboration, and expressed their interest in future continuation of this workshop series, an outcome that is encouraging and shows the effectiveness of this kind of initiatives.

We look forward to future re-occurrences of this workshop to build up momentum for research in Bioinformatics and the Life Sciences in the region, and hope to see the results of the cooperation ties in the form of work collaborations between the participants.

The workshop was wrapped up by a final day devoted solely to personal interactions, and during which participants were able to enjoy the leisure of a visit to Badagry and the beach at Ogungbe.

Acknowledgements:

This workshop was organized by Covenant University, Ota, Nigeria and supported by EMBnet.

Dr. David Oyedepo, Chancellor of Covenant University, Prof. Aize Obayan, Vice Chancellor,

Dr. Daniel Rotimi, Registrar, Prof. James Katende, Dean, College of Science and Technology, and in-fact the entire Board of Regents are to be thanked for their support in making this event possible. The Workshop Steering Committee (Dr. Ezekiel Adebisi, Ms. Ijeoma Dike, Mr. Conrad Omonimiyin, Dr. Wande Daramola, Mr. Oluwagbemi, Mrs. Ogunlana and Dr. Segun Fatumo) are to be thanked for the efforts invested in making this a successful and smoothly run event, the Workshop Resource Persons (Drs. Valverde, Rebai, Bewaji, Adebisi, Osamor and Fatumo) are thanked for their labor in steering the workshop, and all participants are to be acknowledged and congratulated for the big efforts and great undertakings achieved during it.

References:

1. Larkin MA, et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, 23, 2947-2948.
2. Galtier et al. (1996) SEAVIEW and PHYLO _ WIN: two graphic tools for sequence alignment and molecular phylogeny, *Bioinformatics*, 12-6, 543-548
3. Valverde, J. R. (2007) IBS-ES-07 course: The making of. *embnet.news*, vol 13, no. 4. 11-17
4. Rice et al. (2000) EMBOS: The European Molecular Biology Open Software Suite, *Trends in Genetics*, 16(6), 276-277.
5. Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle
6. Prokop et al. (2008) TRITON: a graphical tool for ligand-binding protein engineering. *Bioinformatics*, 24(17): 1955-1956.
7. Sali, A. and Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234(3):779-815.
8. Wahab et al. (2009) On Elucidating Isoniazid Resistance Using Molecular Modeling, *J. Chem. Inf. Model.* 49, 97-107
9. Johnson et al. (2006) Drug Resistance in *Mycobacterium tuberculosis*. *Curr. Issues Mol. Biol.* 7: 91-112
10. Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S. and Olson, A. J. (2009) *J. Comput. Chem.* in press. "Autodock4 and AutoDockTools4: automated docking with selective receptor flexibility."
11. MOPAC2009, James J. P. Stewart, Stewart Computational Chemistry, Colorado Springs, CO, USA, <http://OpenMOPAC.net> (2008)

Personal Account: Training in Grid Computing for Bioinformatics



Kanchana Senanayake

Institute of Biochemistry, Molecular Biology and Biotechnology (IBMBB), University of Colombo, Cumaratunga Munidasa Mawatha, Colombo, Sri Lanka

Since the Institute of Biochemistry, Molecular Biology and Biotechnology (IBMBB) was elected as the Sri Lankan EMBNet node at the EMBNet AGM in 2006 we have been in close contact with the Swedish EMBNet node. As a result of this association I was fortunate enough to get an invitation for a short visit to the Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden. This visit was funded by SIDA/SAREC and IBMBB. In this article I will be giving a brief insight in to my visit from an academic and a personal point of view.

Academic

The main objective of my visit was to get hands on training in grid computing and its applications in Bioinformatics. A 14 node cluster was given to the IBMBB under the NorduGrid project in 2006. **NorduGrid** is a Grid Research and Development collaboration aiming at development, maintenance and support of the free Grid middleware, known as the [Advance Resource Connector \(ARC\)](#)¹. The aim of this collaboration was to use grid computing for bioinformatics applications as well as porting existing bioinformatics tools on to the grid. These tools then can be used for high volume data processing application in various fields of Biology. Unfortunately the know-how was not there to make the best use of this resource at the IBMBB. The best way of describing this situation the Sri Lankan way is:

"Like giving an elephant, without the goad to control it"

So for the last three years this was not utilized at all for any type of research. In order to utilize this resource IBMBB needed someone with hands on

¹ <http://www.nordugrid.org/middleware/>

knowledge on Grid computing. As a result of this requirement I was requested to go to Sweden by Prof. Kamani Tennekoon, Director of IBMBB and Node Manager of the EMBNet Sri Lanka to gain the necessary knowledge and experience to better utilize this resource at IBMBB.

With help and support of the UPPMAX personnel I managed to assemble a 5 node cluster and install the OS and the middleware. The UPPMAX was kind enough to provide me 5 spare nodes to construct a fully functional cluster. These nodes had the same hardware configuration as the cluster that is at the IBMBB. The purpose of this exercise is to create a 5 node replica of the cluster in Sri Lanka but with the updated versions of the OS and the middleware, so I could do the same to our node and upgrade the cluster in Sri Lanka. After installing the cluster I managed to install and configure few bioinformatics tools and databases with the support of Prof. Erik Bongcam-Rudloff. This bioinformatics cluster can be accessed with the URL <http://biocluster.hgen.slu.se/>.

Although the main objective of this visit was Grid computing I was fortunate enough to get exposed to several other aspects of Bioinformatics as a result of the collective research culture at the Bioinformatics lab headed by Prof. Erik Bongcam-Rudloff. I was initially involved in a sequence assembly project with an MSc student at the lab which was his research project of the MSc. Through this exercise I learned different techniques and tools used for sequence assembly and their problems in execution for Illumina data. This made me realize the importance of a bioinformatics cluster to do these big data analysis work. I also got a chance to learn about Microarray data analysis using R and Bioconductor. This knowledge was gained through the participation in the one week course "DNA expression microarray data analysis using R and Bioconductor" conducted by the CSC - IT Center for Science in Finland. I also had a brief introduction to Weblab as the creators of this bioinformatics workflow had a visit to the SLU to do a local installation of the Weblab at the SLU.

Personal

From a personal point of view I thoroughly enjoyed the visit to this wonderful country and made some good and interesting friends both from Sweden and from other countries like Pakistan, China, Iran and Kenya.

The Swedish are very friendly people and a great host. I found that they are very serious about two things other than work, which are holidays and Coffee. The support that I got from these people for my work is maximum, specially from the people from SLU and UPPMAX. Without their support I could not have completed this training successfully.

Probably the highlight of this trip from a personal point of view is my visit to the Linnaeus Gardens. He is known as the father of modern taxonomy, and one of the fathers of modern ecology and happens to be my Favorite scientist. I think he was one of the greatest logical minds of the 18th century. This was one of those rare opportunities that you get to see your idol. I will also never forget the trip to Finland on a ship. This was the first time I was on a cruise ship which was a great experience for me. During this trip I saw so many wonderful sceneries and had expensive but excellent food.

I like to take this opportunity to thank the Department of Animal Breeding and Genetics of the Swedish University of Agricultural Sciences who was my host for providing me with all the facilities and the staff who provided me with all the support. I also like to thank the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX), University of Uppsala and the staff for providing me with the resources and the know how to make this training a success. I also like to thank SIDA/SAREC for funding this visit and the staff at International Science Program office who looked after me while I was in Sweden.

I look forward to continued and enhanced collaboration with Swedish EMBNet node and the other nodes in the EMBNet.



PairsDB protein alignment atlas - interface and database tables



Kimmo Mattila

CSC – IT Center for Science,
Espoo, Finland

On EMBnet news vol. 13 (4) [1] we presented the PairsDB protein sequence alignment database [2] that contains directly computed or hierarchically inferred pairwise alignments for all known protein sequences. Since 2007 several small modifications have been made to the database and the WWW interface (<http://pairsdb.csc.fi/>). The database itself has been updated twice. The latest release (1/2009) is based on the protein sequences collected from protein databases in February 2009.

In this article we provide an updated description about the interface and also discuss the general features of the most essential SQL database tables of PairsDB. The data collected to the PairsDB database is freely available and provides a unique resource for studying the currently known protein universe with the methods of bioinformatics and data mining.

Structure of the PairsDB database

PairsDB is based on a non-redundant set of protein sequences and their hierarchical clustering. The sequences of PairsDB are collected from UniProt, PDB and RefSeq databases. Identical sequences are merged into a single entry. In PairsDB sequences are considered identical only if they have the same length and 100% sequence identity. This first pruning of the source data produces a sequence set non-identical protein sequences called **NRDB100** (Non Redundant sequence DataBase).

In the second pruning step the NRDB100 sequences are clustered with CD-HIT program. 90% identity is used as the threshold level for the clustering. As a result the NRDB100 sequences are sorted into sequence families that contain a long representative sequence and group of shorter family members that are more than 90

% identical compared to the representative sequence. The representative sequences form the NRDB90 sequence set.

For the NRDB90 set a BLAST analysis is run in an all-against-all fashion. Using the BLAST results non-redundant sequence sets are created for 80%, 70%, 60%, 50%, 40%, and 30% sequence identity levels. As a final step an all-against-all PSI-BLAST analysis is run using the NRDB40 sequence set.

When data is retrieved from the PairsDB database, this hierarchical sequence classification and pre-calculated BLAST or PSI-BLAST alignments are used to construct a set of similar sequences and their alignments. For single query sequence the NRDB90 family and its representative sequence is first checked from the database. Also the alignment between the query and the representative sequences is retrieved. Using the pre-calculated BLAST results, other NRDB90 level sequences and their family members can then be promptly collected.

Below are some key figures from the 1/2009 version of PairsDB. This data gives an overview of both the size of PairsDB and also of the currently known protein universe.

- The total number of protein sequences collected from source databases (UniProt, PDB and RefSeq) **13,4 million**.
- Number of unique sequences (NRDB100) was **7,3 million**. Of these, 36% were found only once in the source databases. About 24% of the unique sequences were found to exist in more than one organism.
- Number of sequence families that are less than 90% identical to other sequences (NRDB90) is **4,4 million**. 80% of these families contain only one sequence.
- Number of sequence families that are less than 40% identical to other sequences (NRDB40) is **2,3 million**. 69% of these families contain only one sequence.
- Number of BLAST matches within the NRDB90 sequence set: **9428 million**
- Number of PSI-BLAST matches within the NRDB40 sequence set: **5003 million**

Finding name for your sequence

PairsDB interface is operated using the UniProt, RefSeq or PDB sequence names like CYC_HUMAN, NP_061820.1 or 1J3S-A (this refers to the A-chain

of PDB entry 1J3S). If you do not know the name of your sequence you can use the "Sequence Space Filter" to check it. Sequence space filter is found in the top bar of the PairsDB interface. With this search tool you can try to find the sequence name by searching the sequence descriptions finding sequences that match 100% to your query sequence or a fragment of it. Often already a fragment of 10-20 amino acids is enough to identify your sequence. If the sequence is not found, the reason may be that it was not yet in the public databases when the last PairsDB data set was collected. Sequence Space Filter can also be used to collect sequence data sets using combination of several search criteria. For example you could easily collect all sequences that

are from a certain taxonomic group and contain a given InterPro domain.

BLAST and PSI-BLAST based searches

PairsDB provides two ways to look for similar sequences for your query sequence. BLAST in NRDB90 level and PSI-BLAST in NRDB40 level. Both of them use the same logic to construct the sequence relationships from the database. Here we discuss only about the BLAST search interface but the same features exist also in the PSI-BLAST interface. The BLAST search interface can be opened from the BLAST link in the top bar of the interface. To start the search, define the "Query sequence" and press "Search" button. Remember that you should feed the name of

Figure 1. The BLAST query interface of PairsDB.

CYC_HUMAN is a cross-reference for XP_001140708.1, which is represented by A8MV93 in NRDB90.

Set	Shortcuts	Acc.No.	Identifier	Database	Description
NRDB	I B P	XP_001140708.1	114687932	RefSeq	PREDICTED: similar to cytochrome c [Pan troglodytes]
NRDB90	I B P	A8MV93	A8MV93_HUMAN	UniProt	Putative uncharacterized protein ENSP00000381989 - Homo sapiens PE=3 SV=1

BLAST Results for XP_001140708.1 expanded to NRDB100

Filtering conditions for the hit sequences

- E-value is less than 0.01

Matches: 942 Displaying first 50

Match Overview

I	ID	Score	E-value	Shortcuts	Acc.No.	Organism	Description
				I B P	XP_001140708.1	Gorilla gorilla gorilla, Pan troglodytes, Pongo abelii, Homo sapiens	PREDICTED: similar to cytochrome c [Pan troglodytes]
				I B P	A8MV93	Homo sapiens	Putative uncharacterized protein ENSP00000381989 - Homo sapiens PE=3 SV=1
		559	3.8E-56	I B P	XP_001095458.1	Macaca mulatta, Macaca sylvanus	PREDICTED: similar to cytochrome c (isoform 2 [Macaca mulatta])
		474	2.7E-46	I B P	B5MCJ8	Homo sapiens	Putative uncharacterized protein ENSP00000384933 - Homo sapiens PE=4 SV=1
		474	2.7E-46	I B P	Q7YR71	Trachypithecus cristatus	Cytochrome c - Trachypithecus cristatus
		474	2.7E-46	I B P	XP_519702.1	Pan troglodytes	PREDICTED: similar to cytochrome c [Pan troglodytes]

Figure 2. BLAST result page of PairsDB.

the sequence to the "Query sequence" field, not the actual sequence data.

As a first search step the NRDB90 level representative sequence for the given query sequence is retrieved. Then BLAST hits for the representative sequence are collected at the NRDB90 level. After this the hit list is expanded to NRDB100 level so that only those sequence neighborhood members that have overlapping match region with the query sequence are selected to the result set.

The hit sequence list can also be filtered using following features:

- e-value (can vary between 1 – 0);
- exclude fragments, hypothetical or transmembrane proteins;
- select only hits that are from certain source database (UniProt, PDB or RefSeq) or that are included on certain NRDB hierarchy level;

- domains from InterPro, SCOP, CATH or ADDA domain databases. For InterPro and ADDA standard database identifiers are used. For SCOP and CATH domains PairsDB uses coding system, that can be checked from help pages of PairsDB;

- Taxonomy ID number;
- subregion of the query sequence.

Retrieving and filtering the data takes 10s to few minutes depending on the size of the result set and the selected filtering methods and output formats. The number of hits to be reported is by default limited to 50 but can be expanded up to 10,000.

BLAST Results

The BLAST results page starts with information about the query sequence and the corresponding representative sequence in NRDB90 level. After that, filtering conditions, used in the query,

are listed. By default the actual results are shown as match overview table, stacked multiple alignment and pairwise alignments.

Match Overview

The Match Overview table lists the found BLAST hits. The first column displays the location of the matching region between the hit and representative sequence. The original query sequence is represented as a red bar and its NRDB90 representative sequence as a green bar. The matching sequences that originate from NRDB90 are shown as dark yellow bars while the corresponding NRDB100 level family members are presented as light yellow bars.

Using the shortcuts (I,B,P) you can directly go to the sequence info, BLAST or PSI-BLAST page of any of these sequences. Note also that one hit in NRDB100 level can represent several entries in the source databases. Thus if the result list seems to lack a UniProt entry or PDB structure that should be there, it may be presented by some other sequence name. For example UniProt entries CYC_GORGO, CYC_HUMAN and the A chain of PDB entry 1J3S have identical sequences so they are presented by only one hit, in this case named as 1J3S-A.

Stacked Multiple Alignment

The stacked multiple alignment shows those regions of the hit sequences that align with query sequence. The density of the colour refers to how well conserved a specific amino acid is in the alignment. In the stacked alignment the hit sequence regions that do not align with the query sequence, are not shown. Thus this query-anchored stacked alignment is NOT a multiple sequence alignment. Stacked multiple alignments are not shown for query sequences that are longer than 1000 amino acids.

Pairwise Alignments

This section displays the pairwise alignments between the query and hit sequences. The score and E-values refer to the values of the NRDB90 level BLAST hits thus they are not exactly correct values.

Other output options

You can modify the BLAST results display in the BLAST query page: You can print the hit sequences or stacked multiple alignment in fasta format or switch of some part of the output. Often the

most time consuming part of the PairsDB result processing is constructing and downloading the HTML presentation of the stacked and pairwise alignments. Using only Match overview presentation can make PairsDB to act much faster.

PairsDB SQL Tables

The PairsDB www-interface allows an easy way to use the PairsDB database as a handy substitute for BLAST. However the real power of PairsDB can be obtained by using the database directly through SQL queries. CSC does not provide tools that would allow any user to submit free MySQL queries to the database, but the database content is freely available at the FTP site of CSC:

<ftp://ftp.funet.fi/pub/sci/molbio/pairsdb/>

There are two limiting issues in using the data, however. Firstly, the size of the current PairsDB version is about 1,5 TB. Another drawback is that the database is not well documented.

All together the PairsDB consists of 50 different tables. We present here the most important tables of the system to help potential users to get started. Installing instructions for the PairsDB tables can be found from the README document at the FTP site.

nrd

The NRDB table is the most central table of the database. It contains all the unique sequences that form the non-redundant data set. For each unique sequence a unique id number: **nid**, is assigned. This nid number is used in all PairsDB tables to identify the sequence. In addition to the nid number the nrd table contains columns for the actual sequence string, description, sequence length, date and a filter column that describes the position of the sequence in the PairsDB hierarchy. Zero value in the filter column means that the sequence is obsolete and no more in use in the other tables.

Each nid has accession number and identifier values too, however you should note that the accession number, identifier and description, presented in the nrd table are not necessary the only ones that in the source databases refer to this sequence. The possible other values can be found from the cross _ references table.

cross_references

This table contains information about the names and accession numbers that have been used for

a certain sequence (i.e. nid) in the source databases. So if you would like to know the nid of a sequence you are working with (say RIMM_ECOLI). You could check it with SQL query:

```
SELECT nid FROM cross_references WHERE
identifier="RIMM_ECOLI";
```

cross_references table also has column to identify the source database where the accession number was used (1 = UniProt, 3 = PDB and 12 = RefSeq), and the sequence description present in the source database.

pairsdb_90x90 and psiblast_40x40

The all-against-all BLAST results for the NRDB90 sequence set and the all-against-all PSI-BLAST results for the NRDB40 sequence set are stored into tables pairsdb_90x90 and psiblast_40x40. These very large tables have identical structures. The two first columns hold the nid numbers of query (query_nid) and hit sequences (sbjct_nid). The e-value is stored to the third column as the logarithm of the actual value (log(e)). The following six columns contain information about the alignment between the two sequences. In addition to the starting, and end residues of the actual structure of alignment is stored too. The alignment structure is stored to query_ali and sbjct_ali columns in a format where +X means X aligning residues and -X X gaps in the alignment. So for example BLAST alignment:

```
Query:      ALES-SAS
           | | |::
Hit:       A--SESVA
```

Would be stored in following format:

```
query_ali:  +4-1+3
sbjct_ali:  +1-2+5
```

The last two columns of this table contain the score and identity percent of the alignment. As pairsdb_90x90 and psiblast_40x40 tables contain billions of rows, indexing of the columns that will be used in the queries is essential.

pairsdb_100x90 and pairsdb_100x40

If the query sequence does not belong to NRDB90 or NRDB40 sequence sets, one has to be able to check what is the nrdb90/40 level representative sequence for the query sequence and how the sequence aligns with the repre-

sentative sequence. This information is stored to the pairsdb_100x90 and pairsdb_100x40 tables. The alignment between the reference sequence (rep_nid) and the member sequence (mem_nid) of the sequence family is coded in the same way as in the pairsdb_90x90 and psiblast_40x40 tables.

Acknowledgment

PairsDB was developed by Prof. Liisa Holm and Dr. Andreas Heger, and it is maintained jointly with CSC.

References

1. Mattila K (2007) PairsDB protein alignment database. EMBnet news 13 (4):22-24.
2. Heger A, Korpelainen E, Hupponen T, Mattila K, Ollikainen V, Holm L (2008) PairsDB atlas of protein sequence space. Nucleic Acids Res. 36(Database issue):D276-D280.

A More elaborative way to check codon quality: an open source program



Rakesh Kumar Shardiwal¹ and Dr. Sohrab Sartaj Sayed²

¹ Genseq Sdn Bhd, Cyberjaya, Malaysia

² JK Agri Genetics, Hyderabad, India

Introduction

Protein-coding genes are translated into amino acid polypeptides following the genetic code. The sequence of a gene directly determines the sequence of amino acids in the protein it produces [1]. In a reading frame of protein, each group of three consecutive nucleotides in the DNA (or RNA) sequence corresponds to an amino acid residue that will be incorporated into the protein sequence. These nucleotide triplets are called "codons". The correspondence between the codons and their coded amino acids constitutes the genetic code [2]. Genetic code elements have large number of redundancy. A direct result of the redundancy is the observation of codons that codes for the same amino acid (synonymous codons). These codons are very rarely used with equal frequency.

According to the study of Sharp and Li [3] in *Escherichia coli* and yeast *Saccharomyces cerevisiae*, there is a clear positive correlation between degree of codon bias and level of gene expression and it is desirable to quantify the degree of bias in each gene in such a way that comparisons can be made both within and between species. Codon bias is correlated with a corresponding bias of tRNA, which is a wide arrangement for optimizing the gene expression. On the other side, it is suggested that heterologous gene expression is not as sensitive to codon bias as previously thought, but that it is quite sensitive to other characteristics of the heterologous gene [4-5,9].

An optimal codon will get you more expression with good translation rate. On the other side non-optimal codons has been postulated to reduce translation rate, probably due to a relative scarcity of cognate tRNA species. Non-optimal codons have bit advantage in to maintain a low cellular concentration of the proteins that they encode [6,7-9].

Relative Synonymous Codon Usage (RSCU) measures the relative frequency that each codon suits to encode a particular amino acid.

Methodology

We have implemented an algorithm to optimize the codon, which is based on a simple effective measure of synonymous codon usage bias. The Relative Synonymous Codons Uses (RSCU) value for a codon is simply the observed frequency of that codon divided by the frequency expected under the assumption of equal usage of synonymous codons for an amino acid [5].

Thus,

$$RSCU_{ij} = \frac{X_{ij}}{\frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}} \quad [1]$$

Where X_{ij} is the number of occurrences of the j th codon for the i th amino acid, and n_i is the number (from one to six) of alternative codons for the i th amino acid. In the absence of any codon usage bias, the RSCU value would be 1.00. A codon that is used less frequently than expected will have a value of less than 1.00 and vice versa for a codon that is used more frequently than expected. The RAC (Relative Adaptiveness of a Codon) is calculated based on RSCU value, the frequency of use of that codon compared to the frequency of the optimal codon for that amino acid:

$$w_{ij} = RSCU_{ij} / RSCU_{imax} = X_{ij} / X_{imax} \quad [2]$$

Where $RSCU_{imax}$ and X_{imax} are the RSCU and X values for the most frequently used codon for the i th amino acid. Codon usage data have been compiled for *trpR* gene lowly expressed regulatory gene and *dnaK* gene of *Escherichia coli* to obtain reference RSCU and RAC value.

Process Flow

The codon quality of coding sequences can be depicted in two different ways. The simplest way of depiction is to plot the codon usage frequency that can be found in common codon usage tables [5,10]. A more elaborate way to depict the codon quality is to convert the codon usage frequency into relative adaptiveness values. In contrast to the codon usage frequency the relative adaptiveness takes into account the number of codons which code for the respective amino acid. Selection of appropriate codon plays a major role in the determination of codon usage in all organisms; this program is implemented as Object Oriented way to get more efficient and accurate result to select most preferable codons. Our translation for each coding sequence (CDS) is based on genetic codes [2] and RSCU values. The basic principle for deriving relative adaptiveness values out of codon usage frequency values is the following. The codon usage table (Table 1) for trpR gene of *Escherichia coli*, [6,8-9] lists the following values for Glycine and Glutamate codons:

Table 1. The codon usage table for trpR gene of *Escherichia coli*.

AmAcid	Codon	Number	RSCU	RAC
Glu	GAG	1	0.18	0.10
Glu	GAA	10	1.81	1.00
Gly	GGA	7	0.55	0.30
Gly	GGT	23	1.82	1.00

The frequency of each codon is listed in the 3rd column named "Number". Comparing the RAC values for the best glutamic codon GAA (1.81) with the best glycine codon GGT (1.82) shows

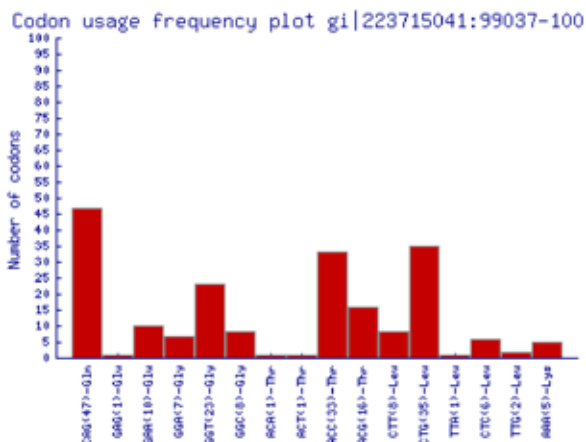


Figure 1. Plotting codon usage frequencies.

clearly that codon frequency values for one amino acid cannot be compared to those of other amino acids even within the same codon usage table. The codon quality of the following sequence stretch is analysed by plotting codon usage frequencies (Figure 1) and relative adaptiveness values (Figure 2).

Inputs

Our Program support three kinds of Input

1. If you have your customize sequence then you can use like this

```
my $seqobj = Bio::Tools::CodonOptTable
->new(
  seq =>
  'ATGGGGTGGGCACCATGCTGCTGCTGCTGAATTTGG
  GCACGATGGTGTACGTGCTCGTAGCTAGGGTGGGT
  GGTTTG',
  Id => 'GeneFragment12',
  accession_number => 'Myseq1',
  alphabet => 'dna',
  is_circular => 1,
  genetic_code => 1,
);
```

2. If you want to read from file

```
my $seqobj = Bio::Tools::CodonOptTable
-> new(
  file => "contig.fasta",
  format => 'Fasta',
  genetic_code => 1,
);
```

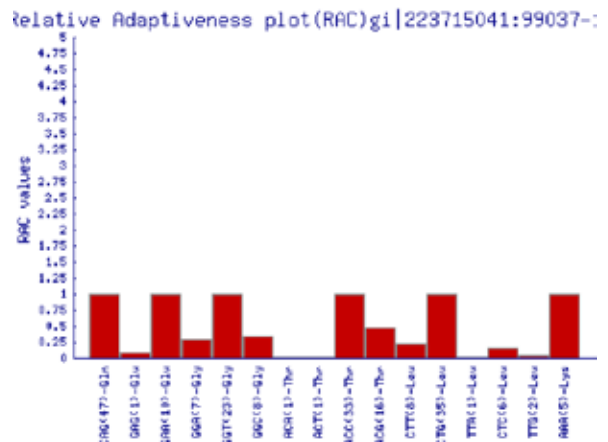


Figure 2. Relative adaptiveness usage frequencies.

3. If you have accession number only, so the program will download that sequence from NCBI and get you the optimization frequency.

```
my $seqobj = Bio::Tools::CodonOptTable
-> new(
ncbi_id => "J00522",
genetic_code => 1
);
```

3.1.2 Input Parameters

Seq	=> sequence string
display_id	=> display id of the sequence (locus name)
accession_number	=> accession number
primary_id	=> primary id (Genbank id)
desc	=> description text
alphabet	=> molecule type (dna,rna,protein)
id	=> alias for display id
file	=> file location
format	=> file format
ncbi_id	=> NCBI accession number
genetic_code	=> 1 (Default)

Output

The program will produce three kinds of Output.

1.1 RSCU and RAC Table along with amino acid name of the codons

```
my $myCodons = $seqobj -> rscu_rac_
table();
if ($myCodons)
{
for my $each_aa (@$myCodons)
{
print "Codon : ", $each_aa->{'codon'}, "\t";
print "Frequency : ", $each_aa->{'frequency'}, "\t";
print "AminoAcid : ", $each_aa->{'aa_name'}, "\t";
print "RSCU Value : ", $each_aa->{'rscu'}, "\t";
print "RAC Value : ", $each_aa->{'rac'}, "\t";
print "\n";
}
}
```

2.1 Graph between RSCU and RAC for more statistical analysis

```
$seqobj -> generate_graph($myCodons, "my
output.gif");
```

3.1 Most preferred codon for the sequence

```
my $preferred_codons = $seqobj ->
preferred_codon ($myCodons);
while ( my ($amino_acid, $codon) =
each(%$preferred_codons ) )
{
print "AminoAcid : $amino_acid \t Codon
: $codon\n";
}
```

Web Interface

The current version of Bio::Tools::CodonOptTable is a 0.07 is a open source pure perl and bioperl program and users can use it with common gateway interface (CGI) perl and make good tool for codons optimizations.

Here is an example tool created with [CodonOptimizer](#)¹ [11]

Availability

<http://search.cpan.org/~shardiwal/Bio-Tools-CodonOptTable-0.07/lib/Bio/Tools/CodonOptTable.pm>

Results and discussion

In this study we have explored the potential of the RSCU and RAC bias table in gene expression. These RSCU and RAC is being used to optimize the codons to get higher expression of desired protein. Our program is based on Sharp and Li [3] study in *Escherichia coli* and yeast *Saccharomyces cerevisiae*.

In order to improve this situation, we have developed a Perl module that relies on the BioPerl bundle and implements the algorithm to optimize the codons for better gene expression. Furthermore, this module let the user to perform simple experiments with codons without having to develop a program or Perl script. We have used Object Oriented approach to solve this problem and provided a simple API (Application Programming Interface).

Our program has the ability to handle complete genome and draw graph of codons based on frequencies and RSCU. In future work, we will develop more comprehensive interface methods to annotate sequence to give more informative results.

Conclusion

This Perl Module is available in CPAN (Comprehensive Perl Archive Network), and can

¹ <http://bioinformatics.chhotikhatu.com/main.html>

also be downloaded. A web-based application is also available (see availability).

References

1. Lewin R (1996). Patterns in Evolution - The New Molecular View. Scientific American Library, New York.
2. The Genetic Codes [<http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi>]
3. Sharp PM, Li WH (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15(3):1281-95.
4. Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol.* 2(1):13-34. Holm L (1986) Codon usage and gene expression. *Nucleic Acids Res.* 14(7):3075-3087.
5. Ikemura T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol.* 2(1):13-34.
6. Grantham R., Gautier C., Gouy M., Jacobzone M., Mercier R. (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.* 9(1):r43-r74. Gouy M., Gautier C. (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* 10(22):7055-7074.
7. Holm L. (1986) Codon usage and gene expression. *Nucleic Acids Res.* 14(7):3075-3087.
8. Sharp P.M. (1985) Does the 'noncoding' strand code? *Nucleic Acids Res.* 13(4):1389-1397.
9. CodonOptimizer [<http://bioinformatics.chhotikhату.com/main.html>]

Designing Primer Pairs and Oligos with OligoFaktorySE



Laurent Gatto¹ and Colas Schretter²

¹ DNAVision, Microarray Unit, Gosselies, Belgium

² Université Libre de Bruxelles, Belgium

Abstract

OligoFaktorySE (Standalone Edition) is a free software for Apple Mac OS X 10.4 or more recent versions. It designs long oligonucleotides for DNA microarrays, primer pairs for PCR amplifications, siRNAs, and more... The innovative user interface emphasizes usability and is aimed at assisting researchers for a painless, rapid, automated, and reliable design. OligoFaktorySE is currently distributed as freeware on the Math & Science section of Apple Downloads (http://www.apple.com/downloads/macosx/math_science/oligofaktorystandaloneedition.html) and on its dedicated website (<http://homepages.ulb.ac.be/~cschrett/oligofaktory>).

Introduction

Primer and oligonucleotide design are important applications of bioinformatics in molecular biology. This is reflected by the abundance of oligonucleotide design softwares that are available. Regular primer design tools (Primer3) are targeted towards computer-savvy users and require scripting skills to automate the design of thousands of primer pairs or oligonucleotides. Alternative tools are often controlled by a graphical user interface (GUI) but only allow the design of a small number of oligonucleotides.

OligoFaktorySE is the software presented in this paper. Its main particularity is that the interface runs exclusively on Mac OS X 1.4 or newest versions of the operating system from Apple. The de-

sign of the interactive interface follows the latest's Apple Human Interface Guidelines (HIG) to easily design large batches of oligonucleotides while tuning finely the design constraints. Furthermore, the interface features advanced visualization of results.

The OligoFaktory has been first introduced to the EMBnet community in 2005 with a short article describing its main features [1]. An equivalent web service was also announced in a Bioinformatics application note [2]. Today, only the standalone version (SE) for Mac OS X is maintained.

Main Features of OligoFaktorySE

OligoFaktorySE allows the researcher to import DNA sequences in FASTA format. It also uses its own XML format to store the sequences, the oligonucleotides, and the design parameters. Several complementary actions are at hand:

- the **Design Oligos** action allows for the automated design of specific long oligonucleotides for the development of microarrays;
- the **Design Primers** action allows for the automated design of specific PCR primer pairs on an arbitrary number of regions;
- the **Design siRNAs** action allows designing optimal 19bp siRNA with the modern method described in [3];
- the **Detect Repetitions** action identifies all repetitions and microsatellites under user-specified constraints;

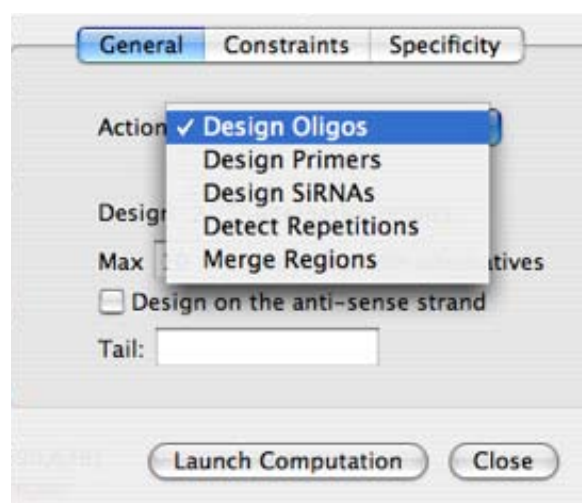
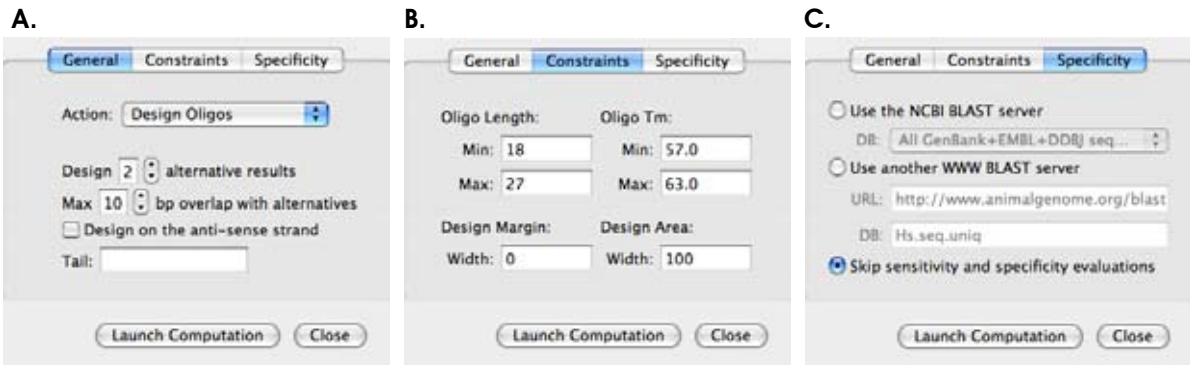


Figure 1. **Actions.** One of the five complementary design actions can be selected at each iteration of the interactive design session.

Figure 2. **Parameters.**

A. For each design action, a dedicated panel showing general design constraints appears to the user.

B. Hard constraints such as the allowed ranges for the oligonucleotide length and melting temperatures can also be set.

C. Sensitivity and specificity of oligonucleotides can be optionally checked against a BLAST database.

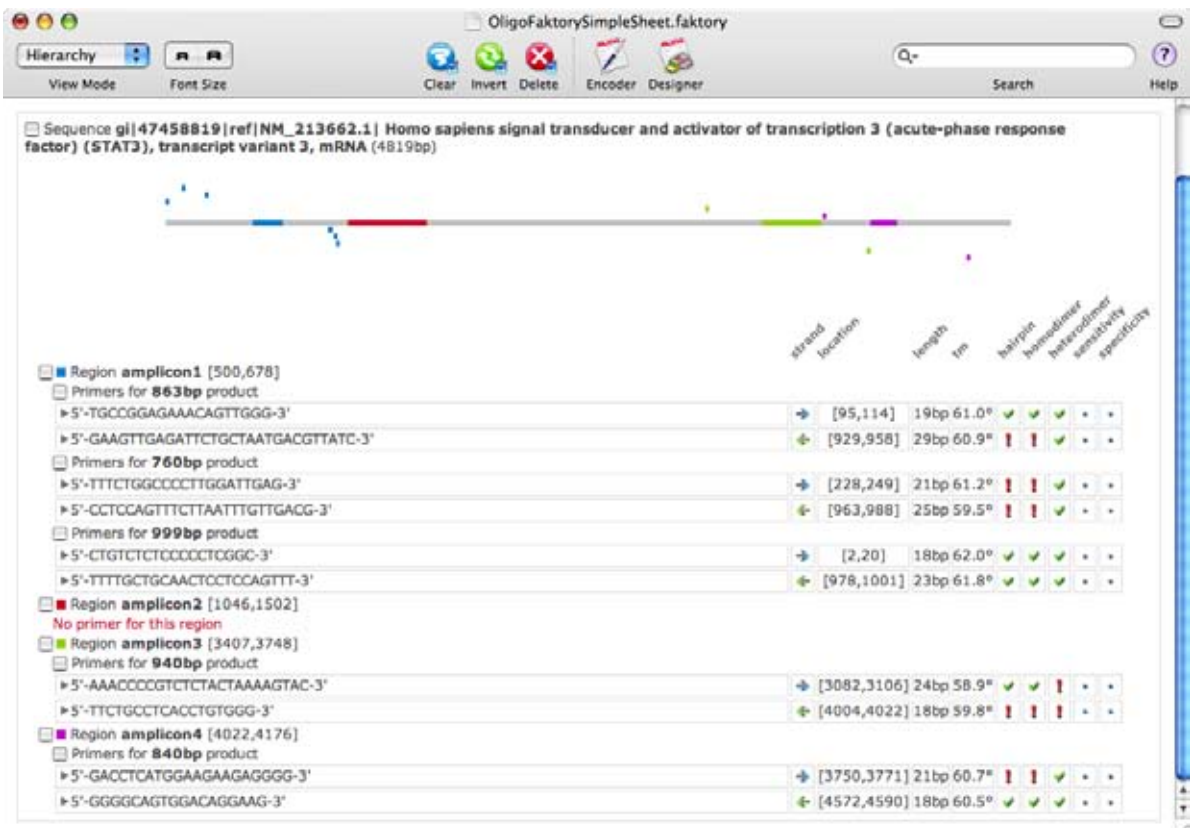


Figure 3. **Visualisation.** A colourful bar graph representation illustrates the relative position and length of primers around amplified regions.

- the **Merge Regions** action merges regions which are too close to allow designing specific primer pairs.

Each of these actions has a set of parameters that can be set by the user: design constraints like oligonucleotide length, preferred product length

for PCR, melting temperatures, location of the oligonucleotides or 5'-end tail.

The underlying design algorithm is based on approaches from statistical mechanics [4] and implements successive selection of best candidates on various criteria like minimization of secondary structures (homo- and hetero-dimers).



Figure 4. **Statistics.** Statistics are computed to visualize the distributions of relevant features for large batches of results.

Furthermore an optional evaluation of specificity and sensitivity can be performed using online or custom offline BLAST databases.

At the end of the design, the results can be directly visualized. The output includes the list of oligo sequences together with their corresponding locations on the query sequences, their lengths, and their melting temperatures. Easy-to-spot warning flags are shown in case of problems with secondary structures and/or with specificity evaluation. These results can be presented in full details using a hierarchical view, a short listing that emphasizes leaf results, while hiding locations information or a statistics view summarizes the results with charts showing the distributions of main features.

Finally, the results can be exported in several formats including OligoFactorySE's own XML based format (including the oligonucleotides, the design warnings and initial input data), FASTA format (oligonucleotides only) and CSV format for easy importation in spreadsheet softwares such as Excel.

Interactive User Interface

The interface presents consistent application windows, from the data import and design parameter specifications to the visualization of results. However, the novelty of OligoFactorySE goes beyond the multiple actions presented above and the graphical interface.

The main originality is certainly the interactive user interface. Initial design constrains that are

→	[95,114]	19bp	61.0°	✓	✓	✓	•	•
←	[929,958]	29bp	60.9°	!	!	✓	•	•
→	[228,249]	21bp	61.2°	!	!	✓	•	•
←	[963,988]	25bp	59.5°	!	!	✓	•	•
→	[2,20]	18bp	62.0°	✓	✓	✓	•	•
←	[978,1001]	23bp	61.8°	✓	✓	✓	•	•

Figure 5. **Warnings.** Easy to spot flags indicates worst-case thermodynamic properties of oligos, suggesting further inspections.

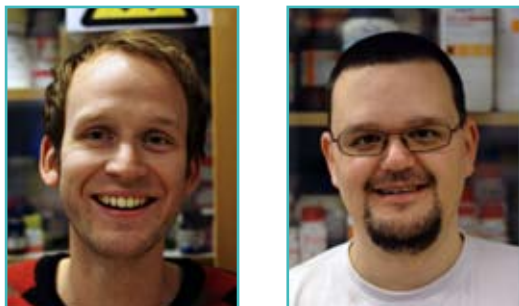
set by the user are often stringent, at least in a first design round, to maximize the chances to yield functional results. However, there are always a set of design that do not succeed and that need to be repeated with reduced parameter stringency. The interactive design paradigm allows extracting several design subsets that can then be manipulated independently: the successful oligonucleotides can be stored and the other ones can be recycled for additional designs. At the end, the several design results can be merged together again.

As OligoFactorySE native format is XML-based, these can be generated and parsed by third party tools. For instance, if hundreds of exons need to be amplified independently, one can easily generate a PCR primer input file based on the gene sequences and exon locations. This input file then just needs to be imported by OligoFactorySE, the design parameters set and the design can start. Similarly, final results files can be parsed to be incorporated in oligonucleotide databases.

References

- [1] Schretter, C. (2005) Discover OligoFactory Standalone Edition. EMBnet.news, 11(4):24-25.
- [2] Schretter, C. and Milinkovitch, M.C. (2006) OligoFactory: a visual tool for interactive oligonucleotide design. Bioinformatics, 22(1):115-116.
- [3] Reynolds, A. et al. (2004) Rational siRNA design for RNA interference. Nature Biotechnology, 22(3):326-30.
- [4] Schretter, C. and Milinkovitch, M.C. (2005) Oligonucleotide design by multilevel optimization. Technical Report, ULB.

Bioclipse 2: towards integrated biocheminformatics



Ola Spjuth* and Jonathan Alvarsson

Department of Pharmaceutical Biosciences,
Uppsala University, Uppsala, Sweden

* Corresponding author
email: ola.spjuth@farmbio.uu.se

Introduction and history

Bioclipse [1] is a free and open source workbench for the life sciences with advanced functionality in bioinformatics and cheminformatics. It allows users to work with resources and entities in the life sciences, such as chemical structures,

sequences, spectra, and alignments. Bioclipse 2, which was released in July 2009, constitutes a complete rewrite of the Bioclipse version published on EMBnet.news in 2007 [2] and provides more features and new graphical components that simplifies integrated life science research and development. The Bioclipse project has as of July 2009 accumulated over 28.000 downloads since its original release in 2007, and also been awarded 3 international prizes for its innovative architecture and intuitive interface.

Architecture

Bioclipse is built on Eclipse (<http://www.eclipse.org>), which is an open source framework that evolved from being an Integrated Development Environment (IDE) into a universal platform for constructing software applications. This provides Bioclipse with advanced plugin architecture, where all functionality is contributed via plugins. Bioclipse defines common interfaces for biological and chemical entities, such as IMolecule for chemical structures, and ISequence for biological sequences. Other plugins can operate on these entities without being aware of each other's existence, for example a tool that visualizes sequences graphically.

In Bioclipse, all functional source code contributed by plugins is collected in Bioclipse Managers; e.g. BioJava [3] contributes functionality via a

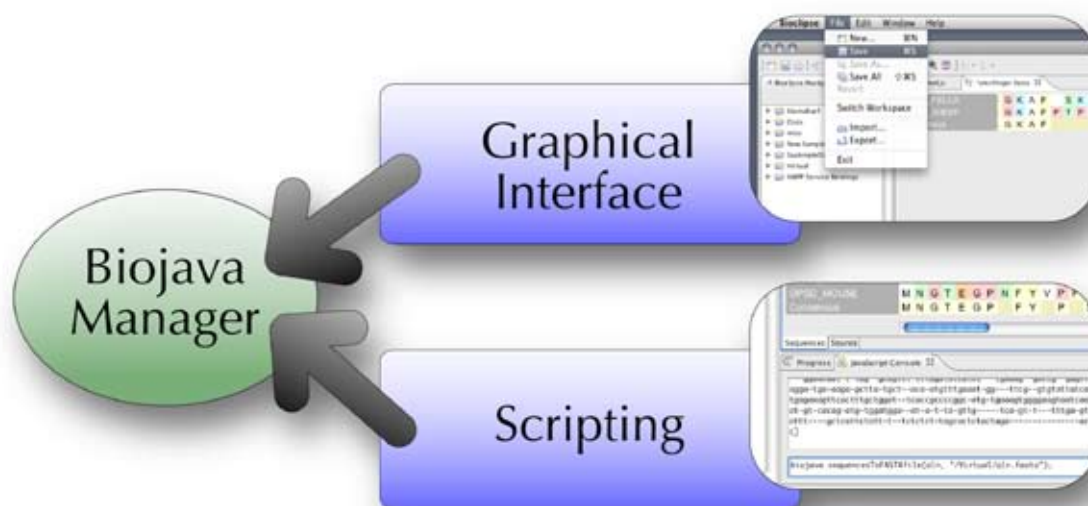


Figure 1. Overview of the Bioclipse architecture describing the use of Managers to collect functional code. The same manager is reachable both from the graphical interface and the JavaScript console, making all functionality available from the GUI and the Bioclipse Scripting Language.

```

seq1 = biojava.DNAfromPlainSequence("CTCGCTTAGAGATA", "myseq");
rnal = biojava.DNAtoRNA( seq1, "myRNA" );
prot1 = biojava.DNAtoProtein( seq1, "myProtein" );
// "myseqs/zf.fasta" to a file with 2 proteins
seqs2 = biojava.proteinsFromFile("myseqs/zf.fasta");

```

Figure 2. Examples on creating and reading sequences in Bioclipse Scripting Language.

BioJavaManager. The Manager objects are built with the help of Spring (<http://www.springsource.org>) and published into the scripting environment. Hence, the same objects that are called from the GUI are also reachable from scripts (see Figure 1), which is named Bioclipse Scripting Language (BSL). The reference BSL is based on JavaScript, and users can invoke all functionality in Bioclipse by typing commands in the JavaScript Console. A JavaScript Editor is also included, which allows for scripting entire analyses. It is already an appreciated feature to use the graphical editors of Bioclipse together with the scripting language to solve biological problems.

Bioinformatics

The core framework for Bioinformatics in Bioclipse 2 is primarily based on Biojava [3], which is available from the BioJavaManager. Figure 2 shows some examples on how to create and read sequences on the JavaScript Console.

Bioclipse also has bioinformatics plugins that take advantage of remote functionality, such as Web services. Examples include WSDbfetch for retrieving data from public repositories, and Kalign for sequence alignments [4].

```

biows.queryEMBL("X56734")
biows.queryRefseq("NM_000410")
biows.queryUniProtKB("INSR_HUMAN")

```

(a)

```

seqs = biows.queryEMBL("X56734,X56735");
aln = kalignws.alignDNA(seqs);
biojava.sequencesToFASTAfile(aln, "save here");

```

(b)

Bioclipse 2 also features new GUI components, such as a new Sequence Editor (see Figure 4) which allows for editing and visualization of sequences, including DNA, RNA, protein sequences, as well as pairwise and multiple alignments.

Cheminformatics

The core framework for Cheminformatics in Bioclipse 2 is primarily based on The Chemistry Development Kit (CDK) [5], which is available via the CDKManager. Figure 5a shows an example of how CDK can be used to create molecules via the JavaScript Console, and open them in the chemical editor JChemPaint. Bioclipse also includes features to query public repositories for chemical substances, for example PubChem. Figure 5b shows a script for querying PubChem for substances annotated with H1N1, downloading them, and visualizing them in the MoleculesTable. As in bioinformatics, the querying functionality is also available from the GUI using a wizard (see Figure 3c).

Bioclipse contains many graphical editors to empower scientists in cheminformatics. One example is the interactive 3D visualization tool Jmol (see Figure 6).

Conclusions

The Bioclipse project aims at providing a workbench with the commonly needed features in chem- and bioinformatics, and also to enable scientific research and development spanning multiple fields. There are ongoing projects to further develop the platform and existing features, but also many new initiatives that widen



(c)

Figure 3. a) Three different commands to query public repositories for sequences. b) Script to query EMBL for two DNA sequences, align them using Kalign, and write the alignment to a FASTA file. c) The first page of a graphical wizard for executing the same queries as in a).



Figure 4. Screenshot from Bioclipse showing the Sequence Editor open with a file containing multiple sequences. The outline (middle frame) shows an overview of the sequences and allows for simple navigation. The JavaScript console (bottom frame) enables scripting of Bioclipse.

the scope of Bioclipse into more fields. The list includes toxicity assessment, site-of-metabolism predictions, integrated local and networked databases, and QSAR analysis. Social features such as integration with MyExperiment [6] are already available, as are features for working with semantic technologies like RDF/OWL. The Bioclipse Wiki (<http://www.wiki.bioclipse.net>) and the Bioclipse Blog (<http://bioclipse.blogspot.com/>) holds the most recent information regarding the Bioclipse development.

License and Availability

Bioclipse 2 is released under the Eclipse Public License (EPL), a flexible open source license that

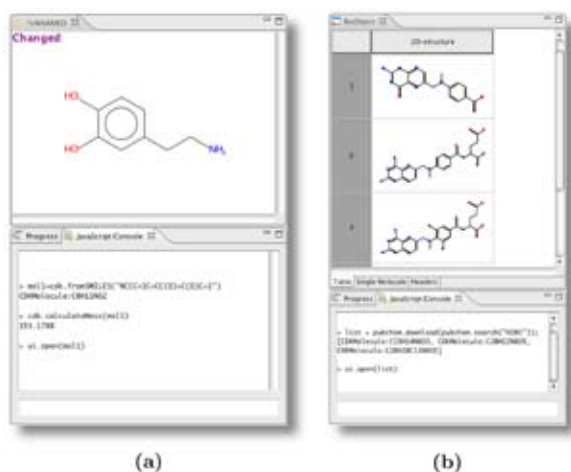


Figure 5. a) Screenshot from Bioclipse showing creation of the compound Dopamine from SMILES in the JavaScript Console, calculation of its mass, and opening of the molecule in JChemPaint. b) Screenshot showing a PubChem query with visualization of the resulting molecules in the MoleculesEditor.

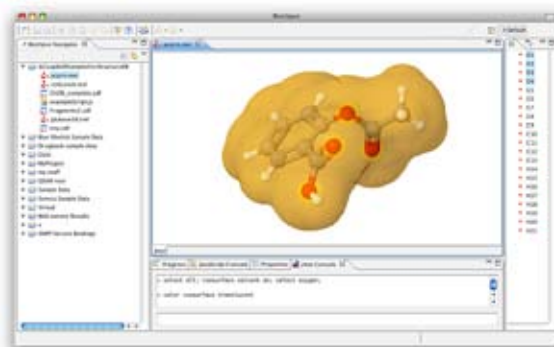


Figure 6. Bioclipse integrates advanced visualization components. Here the interactive 3D visualization tool Jmol is used for displaying an isosurface of aspirin. The Jmol Console (bottom) is used to enter Jmol commands to affect the visualization. Many of these commands are also available from the Jmol menu.

allows additional plugins to be of any license. Bioclipse 2 is implemented in Java and supported on all major platforms. Source code and binaries are freely available at <http://www.bioclipse.net> and development versions are available from <http://pele.farmbio.uu.se/bioclipse-devel/>.

References

- [1] Spjuth O, Helmus T, Willighagen EL, Kuhn S, Eklund M, Wagener J, Murray-Rust P, Steinbeck C, Wikberg JES: Bioclipse: an open source workbench for chemo- and bioinformatics. *BMC Bioinformatics* 2007, 8:59.
- [2] Spjuth O: Using Bioclipse to integrate bioinformatics functionality. *EMBNET news* 2007, 13 (1), 5-11
- [3] Holland RCG, Down TA, Pocock M, Prlc A, Huen D, James K, Foisy S, Drager A, Yates A, Heuer M, Schreiber MJ: BioJava: an open-source framework for bioinformatics. *Bioinformatics* 2008, 24(18):2096-2097.
- [4] Labarga A, Valentin F, Anderson M, Lopez R: Web services at the European Bioinformatics Institute. *Nucleic Acids Res* 2007, 35(Web Server issue):W6-11.
- [5] Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL: Recent developments of the Chemistry Development Kit (CDK) - an open-source Java library for chemo- and bioinformatics. *Curr Pharm Des* 2006, 12(17):2111-2120.
- [6] De Roure, D., Goble, C. and Stevens, R. The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows. *Future Generation Computer Systems* 2009, 25

Ensembl: A New View of Genome Browsing



Giulietta M. Spudich and Xosé M. Fernández-Suárez

European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambs, UK

www.ensembl.org

Abstract

An increasing number of methods are being developed to sequence and compare whole genomes, and to detect functionally important coding and non-coding regions. Genome browsers face the challenge of displaying an integrated picture of these data for the biological community. Ensembl allows life scientists to browse through genomic data using an extensive website, and programmers to access the same data directly through the Perl API. This paper explores



Figure 1. Ensembl data is divided into four tabs, reflecting objects in the database.

the browser and underlying data, providing a walk-through of information for one gene with a focus on variation. We hope to demonstrate the power of the browser in this popular, world-wide genomic tool.

About Ensembl

Ensembl[1], a joint project between the EMBL's EBI and the Wellcome Trust Sanger Institute, was started in 2000. The focus has been on chordates, and in the last few years, Ensembl has offered a genome browser and access to underlying databases for a rapidly increasing number of vertebrate species (currently 50 species... and counting.) To extend the Ensembl platform to invertebrates, a sister project, Ensemblgenomes, at www.ensemblgenomes.org, has been recently launched.

Evolving bioinformatics methods have allowed Ensembl to increase analysis and annotation of the genome. Gene sets are determined not only for the fully-sequenced genomes through the Ensembl gene-building pipeline[2], but for low-coverage (2X) genomes, via the "low-coverage pipeline" developed at Ensembl. Comparative studies have increased, and every species in

Home > Human
Location: 1:114,356,437-114,414,375 Gene: PTPN22

Gene-based displays

- Gene summary
 - Splice variants (4)
 - Supporting evidence
 - Sequence
 - External references (2)
 - Regulation
- Comparative Genomics
 - Genomic alignments (0)
 - Gene Tree (image)
 - Gene Tree (text)
 - Gene Tree (alignment)
 - Orthologues (35)
 - Paralogues (5)
 - Protein families (1)
- Genetic Variation
 - Variation Table
 - Variation Image
- External Data
 - Personal annotation
- ID History
 - Gene history

Gene: PTPN22 (ENSG00000134242)

Tyrosine-protein phosphatase non-receptor type 22 (EC 3.1.3.48) (Hematopoietic cell protein-tyrosine phosphatase 70Z-PEF)

Location: [Chromosome 1: 114,356,437-114,414,375 reverse strand.](#)

Transcripts: There are 4 transcripts in this gene. [hide transcripts](#)

Name	Transcript ID	Protein ID	Description
PTPN22-001	ENST00000359785	ENSP00000352833	protein_coding
PTPN22-004	ENST00000420377	ENSP00000398228	protein_coding
PTPN22-201	ENST00000207489	ENSP00000304749	protein_coding
PTPN22-202	ENST00000354605	ENSP00000346621	protein_coding

[Gene summary help](#)

Name: [PTPN22 \(HGNC \(curated\)\)](#)

Synonyms: [Lyp, Lyp1, Lyp2, PTPN8](#) [to view all Ensembl genes linked to the name [click here](#)]

CCDS: This gene is a member of the Human CCDS set: [CCDS883](#), [CCDS884](#)

Gene type: Known protein coding

Prediction Method: Gene containing both Ensembl genebuild transcripts and [Havana](#) manual curation, see [article](#).

Figure 2. <http://tinyurl.com/nw6vcv> The gene summary page for PTPN22. Four transcripts are shown in the table, each has a unique ENST identifier. These identifiers are stable across Ensembl releases. One is circled, it provides a link to the transcript tab for PTPN22-202.

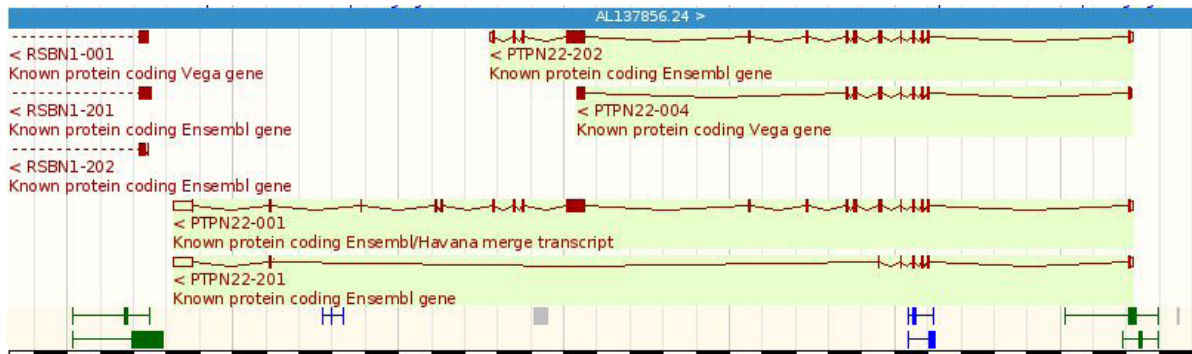


Figure 3. The transcripts are diagrammed below the genome (blue bar) in the gene summary page for PTPN22 (the url is shown in figure 2). Exons are drawn as boxes, and connecting lines show intronic sequence. Filled boxes are coding sequence. The four transcripts are on the reverse strand of the genome.

Ensembl builds phylogenetic trees to determine orthology, paralogy, and ancestral alleles [3]. Whole genome alignments between species pairs, or even multiple genomes (31 mammals) are available. Variations across populations, breeds, and strains are mapped, and a first set of disease-relationships linked to human polymorphisms is available. Functional genomics is also taking a lead, with the ENCODE project revealing potential promoter and enhancer elements in 1% of the human genome, and currently extending to a full genome analysis[4].

Ensembl Data

To cope with an expanding amount of information, Ensembl keeps evolving to enhance the user experience. Based on feedback collected at browser workshops, questions to the helpdesk, and world-wide user surveys run by Ensembl, the website has been designed to display a vast amount of information in an organised manner. The data in the website is now separated into tabs: location, gene, transcript and variation (Figure 1). Users may browse a region of the genome, or focus on homology, variations, or sequence for just one gene or even one splice variant (Ensembl transcript). Data external to the project such as expression profiles from ArrayExpress[5] are also

accessible. The tabs allow easier addition of the anticipated flood of phenotypic data, variations, and regulatory regions from projects such as HapMap[6], 1,000 genomes[7] and ENCODE.

Case Study – Gene, Transcript, Variation and Location Tabs

For this investigation, we use version 55 of the Ensembl browser. To view the same pages upon future releases of the website, view the archive site for version 55:

<http://Jul2009.archive.ensembl.org/index.html>

Let's browse Ensembl by entering *human PTPN22* gene into the search box on the main page at www.ensembl.org

Clicking on the Ensembl gene ID ENSG00000134242 takes us to the gene summary, where we find four transcripts. These isoforms result from alternative splicing of the gene.

At the left of the gene tab are links to the sequence, whole genome alignments, gene trees, variation, and regulatory information.

The transcripts (splice variants) are drawn below the genome (the blue bar). Features below the blue bar are on the reverse strand. Ensembl transcript diagrams show exons as boxes, and

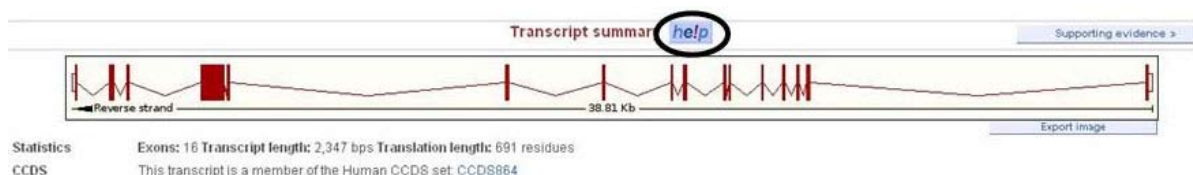


Figure 4. <http://tinyurl.com/kmzmqhg> A larger view of the transcript structure is shown in the transcript summary. The number of exons, length of the spliced transcript in nucleotides, and number of amino acids in the corresponding protein product are displayed. The help button is circled.



Figure 5. <http://tinyurl.com/n39o6z> The general identifiers section for the ENST00000345605 transcript shows IDs in other databases that contain a matching sequence to the Ensembl transcript. The extent of the match is shown as percent identity (%id). The align link, circled, shows the sequence comparison between the Ensembl transcript or protein and the external match. Links to the external databases (such as NCBI RefSeq, and UniProtKB) are encoded in the ID.

intronic sequence as connecting lines. Filled boxes are coding sequence, unfilled boxes are UnTranslated Regions (UTR). For example, PTPN22-202, the first transcript in this diagram, has sixteen exons. To learn more about the transcript, either click on the diagram and follow the link to the transcript ID, ENST00000345605, or click on the ID in the table (circled in Figure 2).

Clicking on the transcript ID opens the transcript tab. The transcript diagram is larger in this view, and a summary is available, showing the length of the mRNA (Figure 4).

Each Ensembl view has a page-specific help article, accessible by the "help" button circled above. This transcript is a member of the consensus coding sequence set (the CCDS[8]); this information is written below the diagram. CCDS sequences are agreed upon by Ensembl, Sanger's Vega/Havana[9, 10] team, UCSC[11] and NCBI[12].

View other IDs for this transcript by clicking on general identifiers at the left of the page (Figure 5). For example, the Ensembl protein sequence matches to the NCBI RefSeq[13] protein NP _

Variations in Watson:

ID	Type	Chr: bp	Ref. allele	Individual genotype	Ambiguity	Transcript codon
rs1599971	INTRONIC	1:114377093	A	A/G	R	-
rs1970559	SARA (Same As Ref. Assembly)	1:114377148	T	T/T	T	-
rs2476601	NON_SYNONYMOUS_CODING	1:114377568	A	A/G	R	CGG

Figure 6. <http://tinyurl.com/lclgnm> The table above shows an excerpt from the population comparison page. The three variations (rs1599971, rs1970559, and rs2476601) are found in James Watson's genome. The first is intronic, the second is the same as the reference sequence GRCh37, and the third shows a non-synonymous allele, indicating there is more than one possible amino acid at that position. The variations IDs provide links to the variation tab, such as rs2476601 (circled).



Figure 7. <http://tinyurl.com/megb64> The variation summary for rs2476601, an NCBI dbSNP identifier.

012411.3 with a Blast Reciprocal Hit score of 99%. Click on the NP (known protein) identifier to jump to the PTPN22 protein in NCBI. Or, click the align link to view the sequence alignment between ENSP00000346621 and NP_012411.3.

Let's explore variations such as polymorphisms mapped to this transcript. The population comparison page displays all the variations, such as SNPs and insertion-deletion mutations (indels), across populations (Figure 6). The ID and position of the variation (such as intronic, non-synonymous coding, etc.) are noted in the first two columns, and the remainder of the table includes the allele in the individual (or strain or breed for non-human species), and the source of the polymorphic data.

An image with this same information drawn graphically is available in the next link (*comparison image*). Both the table and image can be customised using the *configure this page* link at the left. This menu allows selection of individuals, variation types and/or sources to be displayed.

Click on any variation from the table or image to open the variation tab, a focused set of pages for one variation. In this example, let's click on rs2476601, a non-synonymous coding SNP found in James Watson's genome[14].

The variation summary is the first link in the tab, show in Figure 7. Here we find the SNP source (NCBI dbSNP[12]). Any other IDs that this SNP is known by are listed under "Synonyms".

The links at the left of the variation tab are specifically for rs2376601. Click on Phenotype Data (circled in Figure 7) to see that the NHGRI GWAS catalogue[15] relates this SNP to Crohn's Disease, Rheumatoid Arthritis, and Type I Diabetes (Figure 8).

You can easily jump back to the gene or transcript displays by clicking on the appropriate tab. Let's explore the fourth tab in this set, the location tab.



Figure 8. <http://tinyurl.com/mcnbye> The variation tab houses the “phenotype data” page. This shows any associations between a variation and disease phenotype in the NHGRI GWAS catalogue. The study in PubMed that shows the association of the variation to the phenotype is also listed.



Figure 9. <http://tinyurl.com/lfnnga> The top panel of the region in detail page, centred on the PTPN22 gene (highlighted). Protein coding genes from Ensembl are shown in red, genes from the Vega/Havana project are blue. Grey blocks and identifiers show pseudogenes.

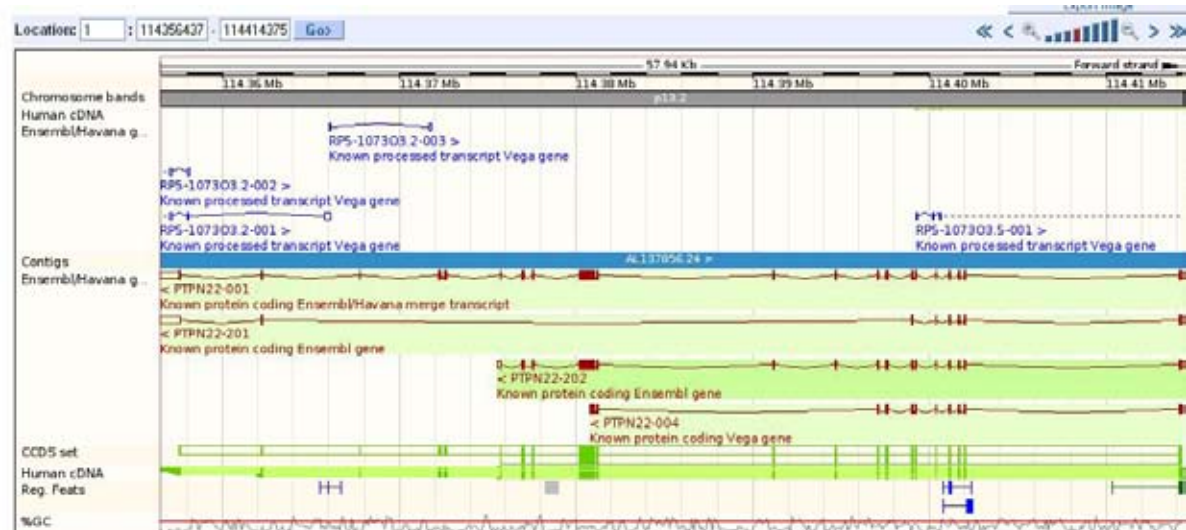


Figure 10. The main panel of the region in detail page is centred on PTPN22. The four transcripts are drawn as in the gene summary page (Figure 3). Coding sequence in the CCDS set along with human cDNA sequences in EMBL-Bank are aligned to the genome. Regions of alignment are displayed in green, filled boxes. Gaps in the alignment are shown by empty boxes. Aligned sequenced support the (red) exons in PTPN22 transcripts.

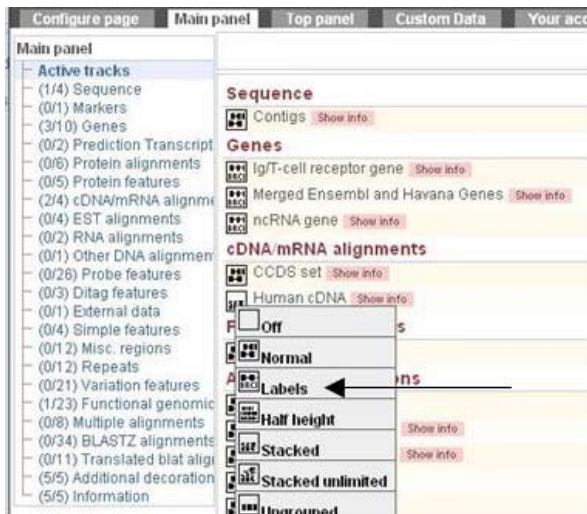


Figure 11. The active tracks (i.e. selected tracks) in the region in detail page. The menu is revealed using the configure this page link. Tracks may be selected in the main or top panel (see tabs) or user data may be uploaded, along with DAS sources (custom data tab). Menus of displayable tracks are shown at the left. The human cDNA track is collapsed; expand the track by clicking on 'normal' or 'labels' (shown by the arrow in the figure.)

Clicking on the location tab (circled in Figure 8 opens the "region in detail" page. For long-time users of Ensembl, this was the ContigView page (Ensembl versions 50 and previous).

The top panel shows a large (1Mb) region of the genome centered on the PTPN22 gene (highlighted in figure 9). Neighboring genes on either side of PTPN22 include RSBN1 and AP4B1. Any gene can be clicked on to view the ID, and/or jump to its gene or transcript tab.

The red box outlining the PTPN22 gene is expanded in the panel below (the main panel, figure 10).

All transcripts (four) of the PTPN22 gene are displayed. These transcripts can be from the Ensembl genebuild, the Vega/Havana manual curators, or they may be a merged transcript, agreed upon by both projects. This is the case for PTPN22-001.

The CCDS set is also drawn in Figure 10, along with the human cDNA alignments. The cDNA track is collapsed by default. To expand the cDNA alignments, click *configure this page* (Figure 11). The active tracks menu will appear (these are the tracks displayed in the "region in detail" page.)

Now each cDNA can be clearly seen (Figure 12). The dark boxes within each entry (such as BC0716701) show the alignment of the cDNA to the genome. Unfilled boxes are gaps in the alignment. (Gaps are expected for cDNA and protein alignments to the genome, as intronic sequence will not be present in cDNA and protein sequences). Click on any green cDNA diagram for an information box, showing the source, as in the example above. These cDNA alignments are updated with every new release of Ensembl, for human.

Case Study Summary

We started this walk-through by searching for a gene symbol (PTPN22). However, Ensembl also allows search by genomic region, accession number, variation ID, clone ID, or disease or phenotype.



Figure 12. The region in detail page with the human cDNA track expanded. Click on any cDNA alignment to view a pop-up box of information. For example, BC071670.1 is a human tyrosine phosphatase record in EMBL-Bank. The CCDS track has been turned off in this view.

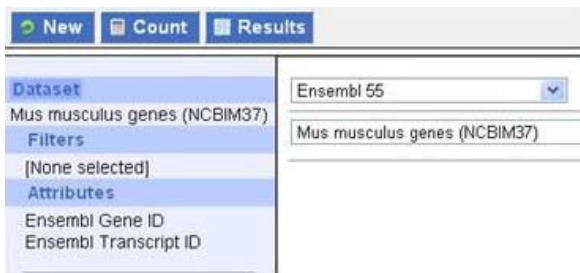


Figure 13. <http://www.ensembl.org/biomart/martview> The BioMart tool allows fast access of data in Ensembl. Sequences and annotation may be exported in FASTA or tabular format. In this example, all mouse genes in Ensembl version 55 have been selected. Output information appears as columns in the table, and are selected in "attributes" (in this case, the Ensembl gene and transcript IDs.).

We quickly learned there are four isoforms (splice variants) that come from the Ensembl genebuild, and/or the Vega/Havana project. The location tab, "region in detail" page displayed human cDNA alignments for the gene locus. This allows users to view support for each exon by entries in databases such as EMBL-Bank[16]. Protein alignments can also be viewed in the "region in detail" page.

One transcript was also explored (PTPN22-202, also named ENST00000354605 in Ensembl). The protein product for this transcript, ENSP00000346621, matched well to the RefSeq peptide NP036543.3, which was seen in the *general identifiers* view.

Finally, variations in different individuals were compared in the population comparison page and the comparison image. One specific variation, rs2476601, was explored in depth. Ensembl views showed the source of this variation was dbSNP, though it is also present in other variation sets, and that three diseases are associated with variation at this nucleotide position.



Figure 14. A heat map showing page impressions to the Ensembl browser in May, 2009. Europe, the USA, Australia and Japan show heaviest use.

Other Access

Ensembl data need not be accessed through the browser. The tabs reflect the organisation of the data in our publicly accessible databases, which are queried by a large number of bioinformaticians. A Perl API is supported, which is heavily accessed by our user community, and is kept current with Ensembl releases (every two months).

<http://www.ensembl.org/info/docs/api/>

In addition, BioMart allows fast mining of Ensembl data for programmers and non-programmers alike (Figure 13) [17] [18].

Ensembl Scope

Who's using us? A recent heat map of page impressions on our website shows a worldwide community of users (Figure 14). Countries shaded darkly show highest usage, and lighter countries access the browser less. Interestingly, this reflects locations of our worldwide workshops.

Ensembl offers workshops in the browser for free. Get to grips with our data by hosting a workshop. Details are here:

<http://www.ensembl.org/info/about/outreach/>

In addition to workshops, we have tutorials on our website, and a YouTube channel of videos providing task-based walk-throughs of the browser.

<http://www.youtube.com/user/EnsemblHelpdesk>

Blog statistics also show a worldwide readership (Figure 15). Posts range from the direction of genomics to details about upcoming species or variation sets in Ensembl.

Find out about upcoming workshops, species, and more on our blog.



Figure 15. Blog readership is marked by red "pins" on the map. "Recent visitor map" and statistics are from StatCounter (<http://www.statcounter.com/>).

<http://ensembl.blogspot.com/>

Despite the relatively high number of queries to the Ensembl helpdesk, answers are returned within one to two days. Questions or comments may be submitted through a form on the Ensembl browser, or directly emailed to helpdesk@ensembl.org.

Conclusions

The Ensembl project aims to provide high-quality genome annotation for vertebrate genomes. Users can freely access all data from various sources, using the extensive pages of the browser at www.ensembl.org, or through BioMart or the Perl API. Ensembl provides displays under four separate tabs to allow flexible addition of the new genomics data to come. Our worldwide users access both the website and the blog.

References

- Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, et al. (2009) Ensembl 2009. *Nucleic Acids Res* 37(Database issue): D690-7.
- Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, et al. (2004) The ensembl analysis pipeline. *Genome Res* 14(5): 934-941.
- Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, et al. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* 19(2): 327-335.
- ENCODE Project Consortium. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447(7146): 799-816.
- Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, et al. (2009) ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res* 37(Database issue): D868-72.
- International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164): 851-861.
- [Anonymous]. 1,000 genomes. .
- Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, et al. (2009) The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* 19(7): 1316-1323.
- Ashurst JL, Chen CK, Gilbert JG, Jekosch K, Keenan S, et al. (2005) The vertebrate genome annotation (vega) database. *Nucleic Acids Res* 33(Database issue): D459-65.
- Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, et al. (2008) The vertebrate genome annotation (vega) database. *Nucleic Acids Res* 36(Database issue): D753-60.
- Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, et al. (2009) The UCSC genome browser database: Update 2009. *Nucleic Acids Res* 37(Database issue): D755-61.
- Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2009) Database resources of the national center for biotechnology information. *Nucleic Acids Res* 37(Database issue): D5-15.
- Pruitt KD, Tatusova T, Maglott DR. (2007) NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35(Database issue): D61-5.
- Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, et al. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452(7189): 872-876.
- Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 106(23): 9362-9367.
- Stoesser G, Baker W, van den Broek A, Camon E, Garcia-Pastor M, et al. (2002) The EMBL nucleotide sequence database. *Nucleic Acids Res* 30(1): 21-26.
- Haider S, Ballester B, Smedley D, Zhang J, Rice P, et al. (2009) BioMart central portal--unified access to biological data. *Nucleic Acids Res* 37(Web Server issue): W23-7.
- Mullan L. (2006) Mining ensembl. *EMBnet.News* 12(1): 12-13.

Linux distributions for bioinformatics: an update



Antonia Rana¹ and Fabrizio Foscarini

Joint Research Centre, European Commission

Introduction

The article provides an updated view on the world of Linux distributions tailored for bioinformatics analysis. The main driver for producing these distributions is to provide an easy-to-use, user friendly environment for non IT specialised users without strong requirements on the knowledge of the technology. Most commonly, intended users of these distributions are students of bioinformatics-related courses. Around 2007, quite a number of Linux distributions, which wrapped almost all available open source tools for bioinformatic analysis, appeared on the Internet. Most of them were assembled by universities and their main purpose was to use them as a tool for teaching and learning. Live CDs which did not require installation were particularly useful for this purpose. The Linux distribution around which almost all of them were built was Knoppix [3], a Linux flavour whose main characteristic was, in fact, that it was an easy customizable live distribution. In the latest two years, new technology trends have emerged in the world of Linux distributions addressed at novice users: the ability to boot from a USB flash drive (practically replacing the CD-ROM, not requiring a CD drive, providing data persistency and reusable) and the availability of distributions as Linux environments to be

run as a virtual machine, in parallel with the host operating system, a feature which has the advantage of giving occasional users or students the possibility to use their usual environments while becoming familiar with a new operating system. This is reflected in the Linux distributions for bioinformatics that we have reviewed in this article. A trend that has been noticed in respect with the review we made in 2007 is the tendency to use Ubuntu as base distribution which is in fact replacing Knoppix and to provide the bioinformatics bench environment also as a virtual machine which can be run inside the popular VMWare environment in parallel with the host operating system. While reviewing the distributions in this article we have paid particular attention to their user friendliness and ease of use.

Bio-Linux

Bio-Linux [4], developed and distributed by the NERC Environmental Bioinformatics Centre, has evolved since our review in 2007, its home page has also changed. Its developers describe it as "...a fully featured, powerful, configurable and easy to maintain bioinformatics workstation" and in fact it is rich with applications and documentation. In its current versions, 5.0, the most notable new features are the possibility to boot it from a USB stick, as well as a LiveDVD and to install it on the hard disk. The Linux distribution on which it is based has also changed from Debian to Ubuntu. This change benefits from all the features and advantages of Ubuntu over Debian, without losing the Debian characteristics since Ubuntu is also based on Debian.

Bio-Linux provides about 500 bioinformatics programs. The complete list is available on its website. The structure and organisation of the bioinformatics programs has not changed since our last review: the bioinformatics applications are accessible via a submenu (**Bioinformatics**) of the Applications menu. Bioinformatics software is installed under `/usr/local/bioinf`. The general layout includes a directory with the base name of the package, under which a directory for each update of the software is installed. This makes it easy for users to locate packages which must be run using the command line.

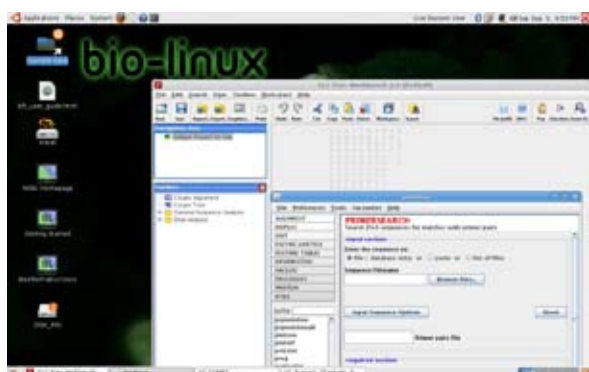
We have tested the LiveDVD version. Hardware recognition went fast without any problem or configuration required from the user. However, during the whole start-up process we did not see

¹ The views expressed are purely those of the writer and may not in any circumstances be regarded as stating an official position of the European Commission.

any information about what the system is doing, we are only shown a progress bar. This can make you feel somehow uncomfortable until the whole process ends: if you use it on a PC where you have data, you want to see what is happening, make sure that the start-up process is progressing, etc.

Once started the system is fast, compatible with the need to access the DVD media when running a new program. Support for the most popular LAN, wireless and bluetooth drivers are provided. The start screen displays the icons: **Getting started** which illustrates the system, **Bioinformatics docs** for easy access to the Bio-Linux bioinformatics documentation system, **NEBC Homepage**, **Install** and **Sample data**. They are very useful to become familiar with the system and how to use it. The Install facility is also conveniently located on the desktop for easy installation on the hard drive. This is a characteristic shared also by all the other distributions which are based on Ubuntu. The security of the system is guaranteed by the installation of a personal firewall (iptables) and ssh for secure remote login.

Bio-Linux has a very rich suite of bioinformatics programs, but what is also very important is that it provides very extensive documentation on the bioinformatics software as well as on the system itself, how to install new packages, how to update it, how to install a bootable USB stick, etc. Its website is also very informative with a lot of useful information.



Download: http://nebc.nox.ac.uk/tools/bio-linux/bl_download

BioBrew and NPACI Rocks with BioRoll

The BioBrew (<http://biobrew.bioinformatics.org/>) distribution has not been upgraded since our last review (version 4.1.3). In fact, already in 2007 the NPACI Rocks cluster distribution with its optional BioRoll package which contains bioinformatics software looked like a candidate to its replacement. NPACI Rocks is a Linux distribution tailored for clusters and was the operating system underlying BioBrew. The main feature that distinguished BioBrew from all the other distributions was that it provided "off-the-shelf" cluster functionality. Currently, this capability can be implemented using NPACI Rocks (current version 5.2) and installing on top of it, its optional package called BioRoll which comprises a large set of bioinformatics tools.

Download: www.rocksclusters.org

DNALinux

DNALinux [5] is the distribution among those reviewed in 2007 that has, more than the others, radically changed. The most notable is that it is no longer distributed as a LiveCD/DVD but only as a virtual machine that can be run inside the VMware player on a Windows OS. The virtual machine bundles together the operating system and the bioinformatics application. Similarly to a live distribution, a virtual machine does not require installation on a dedicated computer and in addition, it can be run in parallel with the host operating system, so you can continue using your PC while running your bioinformatics application in DNALinux.

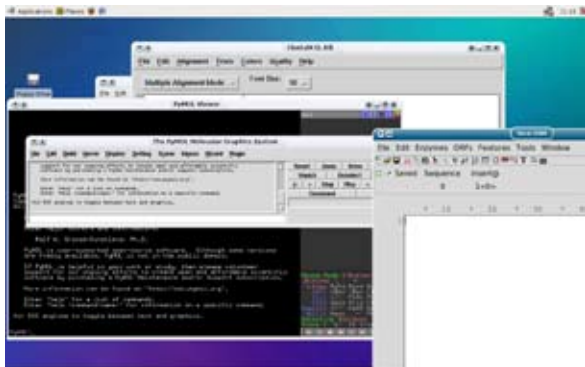
This approach shares with the live CD approach the advantage that users don't need to modify existing installations, it also shares its main disadvantage: the relative low speed of loading applications, in addition running a virtual machine implies higher memory usage.

Another difference from the DNALinux version reviewed in 2007, which was a live CD, is the Linux distribution it is based on. Slax has been replaced with Xubuntu, the light version of the popular Ubuntu linux distribution. Its authors motivate the choice of Xubuntu over Ubuntu as the first is faster thanks to the lighter desktop environment it uses.

The latest version of DNALinux is also included in the book *Python for Bioinformatics*², for this reason it is also called DNALinux Virtual Desktop Py4Bio.

DNALinux provides a large number of pre-installed bioinformatics software, the complete list is available at <http://www.dnalinux.com/installed-software.html>. However we have not found a menu or easy access indications for them. Some of the packages are located in the home directory of the user that is logged into the virtual machine. Some of the tools with a graphical user interface are available under the **Science** or **Education** menus.

DNALinux can be downloaded only using the bit torrent protocol. This can result in longer download times and may be not an optimal solution for environments in which bit torrent traffic is blocked.



Download: <http://www.dnalinux.com/>

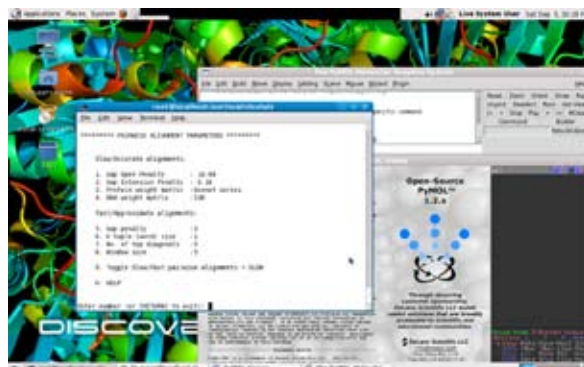
Open Discovery

OpenDiscovery is a new Linux distributions for bioinformatics which was not available at the time of our first review. According to its authors [6], besides providing the usual bioinformatics software (e.g. sequence analysis), OpenDiscovery has been developed with the capability to perform complex tasks like molecular modelling, docking and molecular dynamics. Like Bio-Linux, OpenDiscovery is capable of booting from USB flash drives, live DVD and can be installed on hard disk. Unlike most of the "updated" bioinformatics distributions which seem to prefer Ubuntu over knoppix, the choice which was popular in 2007, OpenDiscovery has chosen Fedora as its base distribution. Open Discovery integrates a

comprehensive range of bioinformatics software. The complete list is available on its homepage.

We tested the live version of OpenDiscovery and, again with a look at user friendliness, we have noted that the startup process is very straightforward although, as for Bio-Linux, seeing only a progress bar instead of the usual list of Linux start-up operations makes us a little nervous when running a live CD. The startup is quick, though and at the end we have a Linux desktop which, without any problem, has started up our wireless network interface and plugged into the network.

The desktop is not rich: you will see the home icon, the classical **Computer** icon and the option to install the system on the hard drive. There is no *Getting started* information or shortcut to the bioinformatics application which would make it easier for novice users to become familiar with the environment. The security of the system is increased with the presence of common security tools such as: a personal firewall (iptables), and secure remote login (ssh). The firewall is configured to exclude any incoming connection.



Download: <http://opendiscovery.org.in/>

BioPuppy

BioPuppy [7] is also a newly found Linux distribution for bioinformatics which is released still in beta version. It is based on a Linux distribution known as LinuxPuppy the main advantage of which is its compactness. It contains all the tools available in the basic LinuxPuppy plus bioinformatics tools.

As for the other distributions which have been updated or new in this review, according to [7] it can be run as a live CD, from a USB stick or installed on the hard disk. We downloaded the live CD version, however testing it was not possible: the start-up process is not as straightforward

² Sebastian Bassi, *Python for Bioinformatics*, Chapman & Hall

as for the other distributions, hardware recognition requires the intervention of the user who is asked to provide information about the monitor for instance. The overall process was not successful. This test was done on a notebook, as different hardware might behave differently we also tested it on a netbook (Acer AspireOne with an USB CD drive) and on a desktop PC. The netbook gave the same result, while BioPuppy started on the desktop PC. The problem seems to be related to the fact that being a very small distribution the choice of drivers available is quite limited. However, BioPuppy comes with a personal firewall installed and has its own package manager to add/delete software (.pet packages). Although it is a very compact distribution, the most popular bioinformatics tools are included (e.g. EMBOSS, HMMER, Clustal-W, Clustal-X, blast, Garlix, Phylip).



Download: <http://biopuppy.org>

BioSLAX

BioSLAX was just emerging when we did our first review. It is now available as a Live CD/Live DVD and bootable from a USB flash drive as well as, of course, for installation on a hard disk. It is released by the National University of Singapore and it is, in fact, an evolution of the APBioKnoppix and APBioKnoppix2 that we reviewed in 2007. BioSLAX is rather different from the other two, though, being based on the SLAX linux distribution (a compressed Slackware flavour of the Linux Operating System).

According to its authors [8], SLAX was chosen over knoppix because knoppix was found not very easily expandable: in order to update an application or add a new one, a new remastering of the distribution was required. This made the distribution highly inflexible. On the other hand, SLAX works by overlaying "application modules" on top of the base Linux OS, thus making the

dedicated bioinformatics distribution built on top of it, modular. Applications can be made into modules which can be inserted either dynamically or via a special folder in the BioSLAX USB/DVD distribution.

The current version of BioSLAX is 7.5 and it is based on SLAX 6. It is available for download in four formats:

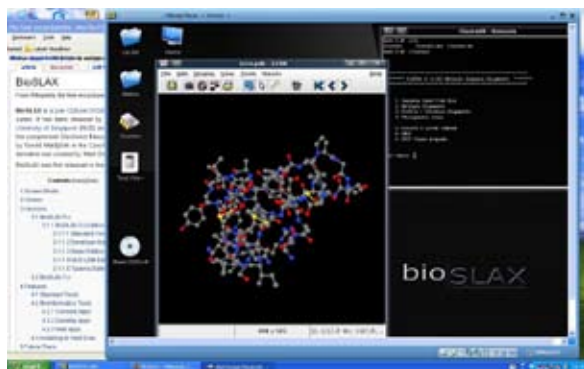
- **Power Developer DVD:** this is the full version with the complete suite of bioinformatics applications which includes also development tools (compilers and Linux kernel headers required for compilation of new applications). This comes as a Live DVD but it can be installed using the BioSLAX installer provided;
- **BioSLAX for NUS LSM courses:** This version is the full power developer version customized for students and teachers at the National University of Singapore;
- **BioSLAX for VMWare (LSM version):** Again the same 7.5 version but created as a virtual machine for use within the virtualisation software VMWare;
- **BioSLAX with Taverna:** In this case the standard 7.5 distribution includes Taverna for workflow management.

We had some problems in testing the LiveDVD version (the over 800MB ISO file is an ISO CD format and not a DVD and could not be burned). We were interested, on the other hand, to test the virtual machine version.

The startup of the virtual machine is quite fast and it is very handy if one has to use the tools sporadically to have your standard PC environment underlying the virtual machine environment although one needs to get used to the keyboard keys combination to get in and out of the virtual machine.

The desktop presents three icons with links to **cgi-bin**, **htdocs**, **home** and **system**. Bioinformatics tools are easily found in the dedicated **BioSLAX** menu which is conveniently further organised into five submenus: **Documentation**, **Console Apps**, **Desktop Apps**, **WebApps** and **BioSLAX Installer**. **Console Apps** provides access to all the tools that are run on the command line (e.g. blast, clustalW, EMBOSS, phylip, primer3, etc), by clicking on one of the menu items in this section, a console window is opened with the PWD set to the directory of the launched application where the executable is located. **Desktop Apps** comprises all the applications which have

a graphical user interface (e.g. ClustalX, jEMBOSS, NJPlot, Pymol, etc.) and finally **WebApps** starts web based applications (e.g. wEMBOSS). In the latter case, Firefox is started opening the launched application. It is interesting to notice that Firefox is equipped with a bookmark toolbar which provides easy and convenient access to bioinformatics-related websites such as Entrez, Bioinformatics.org, NCBI, etc.



Download: <http://www.bioslax.com>

BioconductorBuntu

BioconductorBuntu [9], and the distribution in the next section, are somehow different from the others we have described so far in that they have been tailored to a specific type of bioinformatics analysis: DNA microarray analysis using web-based tools. BioconductorBuntu is also a custom distribution of Ubuntu Linux. It has been created to simplify the process of setting up a microarray processing environment collecting together all the necessary analysis tools for this task in an easily installable and distributable format. The distribution is available as a live CD but, as for the other distribution based on Ubuntu, it is easily installable on the hard disk.

BioconductorBuntu provides a user friendly web-based graphical user interface to many of the tools developed by the Bioconductor Project (hence the name of the distribution). Because many of the tools it provides are accessible via a web interface, the best use of this distribution is to install it on a server and allow network access for microarray analysis. The python scripting environment underlying the Bioconductor modules facilitates the server side integration of additional modules.

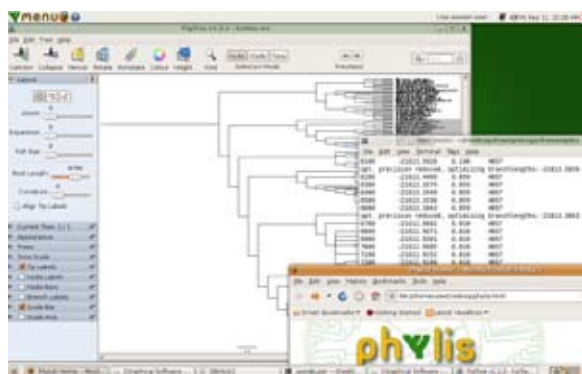
Download: <http://www3.it.nuigalway.ie/agolden/bioconductor/version1/biocBuntu.iso>

phylIs

Similarly to the previous one, phylIS (Phylogenetic Linux for Informatics and Systematics) [10] is also dedicated to a particular type of bioinformatics analysis, namely phylogenetics and phyloinformatics and contains the open source tools that are useful for this type of analysis. The distribution, first released in 2008 and based on Ubuntu, has been configured to include most commonly used phylogenetic software, it is a light distribution streamlined to focus on phyloinformatic research so that computational power is used most effectively for this purpose. Several CPU intensive programs are also available in their parallel version (MPI), so that with the proper hardware processing speed can be improved. Although it does not contain all the range of bioinformatics tools, PhylIS contains popular scripting languages including Perl (with BioPerl), Python (with BioPython), and R.

PhylIS is distributed as a live CD and, like all the other distributions reviewed in this article, it can be also installed on the hard disk. It is based on the Ubuntu Linux distribution.

The system is well documented both on the website and on the desktop where a folder called **Examples** contains sample files for the tools installed and the documentation file phylIS.html, is available which provides an introduction and the list of the software installed also with the indication of its location and command name you would use to start the tool from the command line. A folder called **Graphical software** also located on the desktop, provides easy access to tools with a graphical user interface.



Download: <http://www.eve.ucdavis.edu/rthomson/phylis/>

Package repositories

We did not find new package repositories, on the contrary some that were available in 2007 are no longer available on the internet. Those which are still available were updated, others have restricted their access to registered users. The status of the tools version is collectively shown for all distributions in table 1.

Debian Med

Debian Med has been updated to be included in the latest release of the Debian operating system (Lenny). Although this project is mainly dedicated to medical informatics and medical imaging, the set of bioinformatics tools included in the set of packages is increasing. In the last version, EMBOSS has been included together with one of its web interfaces, EMBOSS Explorer, which allows you to use EMBOSS either locally or on the network. All the major free programs for multiple sequence alignment and structural bioinformatics have also been included. All the programs for sequence analysis and bioinformatics are collected in the med-bio package.

Download: <http://www.debian.org/devel/debian-med/index.html>

Distributions that were not updated

Of the distributions we reviewed in 2007, some were not updated but are still available in the same version on their websites (this is the case for BioKnoppix, BioBrew, Vlinux, VigyaanCD, Quantian, and Goebix), others are not available any longer and in some cases the original websites are also not available (this is the case for AR.EMBNET, DebianBioinformatics, BioLand, APBioKnoppix2 and BioLinux-BR).

Conclusions

After two years we have had again a close look at what the open source world makes available to scientists and students who need a bioinformatics workbench for their analysis. We have noted that, like we anticipated two years ago, updating the base distribution and the bioinformatics tools can be an issue. An aspect which is important if selecting a distribution of choice is the documentation and the availability of a "getting started" introduction. Also important is a

Table 1.

	Blast	Bioperl	ClustalX/ CLustalW	EMBOSS	Glimmer	HMMER	Phylip	Primer3	T-Coffee	Gromacs
Bio-Linux	2.2.19-1	1.4	1.83-3	6.0.1-6	2.13-4	2.3.2-5	3.68-3	1.1.4-0	6.30-1	
DNAlinux	2.2.20	1.5.2	1.83	5.0.0	2.13	2.3.2-3	3.67-1	1.1.1-1	5.31-1	3.3.3-2
Vlinux			1.83	2.9.0	2.0	2.1.1	3.6b	0.9	1.37	3.2.1
BioKnoppix		1.2.1	1.82	2.8.0			3.5/3c			
APBioKnoppix2		1.4	1.83	3.0.0		2.3.2			3.27	
Vigyaan		1.4	1.83	2.10.0	2.13					3.2.1
Quantian	2.2.12	1.4	1.83 (+ClustalW-MPI)			2.1.4	3.61		2.50	3.3-2
GöBIX			1.83	4.0.0					4.9.3	
Biorpms	(ncbi-6.1)	1.4	1.83	3.0.0		2.3.2	3.6a3-5	1.0.0	2.03	
Biolinux ³	2.2.8	1.4	1.83	2.9.0			3.61	0.9		
Rocks + Bio roll	(ncbi 6.1.4)	1.5.1	2.0.11	6.0.1	3.02	2.3.2	3.66	1.0.0	7-81	4.0.4
BioSLAX	2.2.17		1.83	3.0.0		2.3.2	N/A	N/A	3.9.3	
AR.EMBNET	2.2.9		1.83	2.10.0						
BioPuppy		1.4	1.83-1		2.13-1	2.3.2-5	3.67	1.0.0	2.0.3	
PhylIS	2.2.17	1.5.2	1.83	5.0.0-2	2.13-1	2.3.2-3	3.67	1.1.1	5.31-1	
DebianMed	2.2.21		2.0.10	6.1.0	3.0.2	2.3.2	3.68	1.1.4	5.7.2	4.0.5
Package Current Version*	2.2.21 July 2009	1.6.0 Jan 2009	2.0.11 Apr 2009	6.1.0 July 2009	3.0.2 May 2006	2.3.2	3.68 Aug 2008	1.1.4 Apr 2008	8.06 July 2009	4.0.5 May 2009

Black: current and updated

Blue: not updated

Red: no longer available for download

³ http://www.biolinux.org/wiki/index.php/Main_Page available only to registered users

clear indication of where the tools are located and how to launch them. Having a dedicated menu certainly helps the novice users who are likely to be the target of live or USB based distributions.

From a technological point of view, we have noted a trend towards using Ubuntu or its variations as the base distributions with a few exceptions (Fedora and SLAX) and to provide the distribution as a virtual machine as the only choice or in addition to the historical live CD. In Table 1 we have summarised the current versions for the distributions we have discussed here keeping also those which we discussed in 2007. While some tools tend to be quite static, others do evolve. If you are looking for a distribution that you will install in your laboratory, choosing one which can be easily and frequently updated can make a difference.

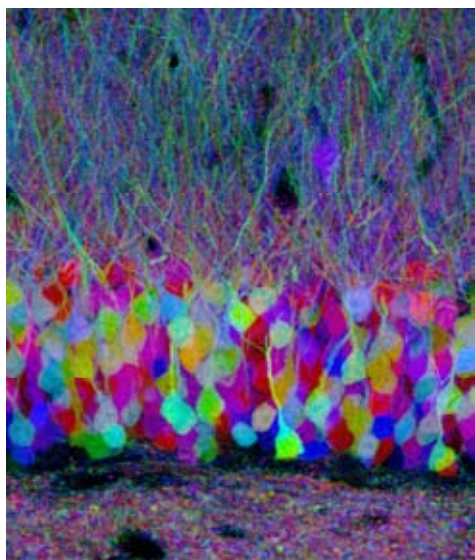
References

- [1] Rana, A., 2007, Linux for bioinformatics: dedicated distributions for processing of biological data – Part 1: Live distributions, EMBnet.News Vol.13 No.2
- [2] Rana, A., Foscari, F., 2007, Linux for bioinformatics: dedicated distributions for processing of biological data – Part 2: Repositories and Complete Systems, EMBnet.News Vol.13 No.3
- [3] knoppix, <http://www.knoppix.net>
- [4] Bio-Linux 5.0, <http://nebc.nox.ac.uk/tools/bio-linux/bio-linux-5.0>
- [5] Bassi, S. and Gonzalez, V., 2007, DNALinux Virtual Desktop Edition. Available from Nature Precedings <<http://dx.doi.org/10.1038/npre.2007.670.1>> (2007)
- [6] Vetrivel, Umashankar; Pilla, Kalabharath, 2008, Open discovery: an integrated live Linux platform of Bioinformatics tools.(Software), Bioinformation, January 1, 2008
- [7] BioPuppy Linux, <http://biopuppy.org/>
- [8] BioSLAX, <http://bioslax.com>
- [9] Geeleher, P., Morris, D., Hinde, J.P., and Golden A., BioconductorBuntu: a Linux distribution that implements a web-based DNA microarray analysis server, Bioinformatics 2009 25(11):1438-1439
- [10] Thomson, R. C., phyLIs: A simple GnU/Linux Distribution for phylogenetics and phyloinformatics, Evolutionary Bioinformatics 2009:5 91–95

paint my thoughts

Cele Abad-Zapatero¹ and Vivienne Baillie Gerritsen

Drawing is probably not a talent the layman would normally associate with Science. Yet it has been an essential ingredient in the life of many scientists for the advancement of their field of research, among them, the Spanish neurobiologist Santiago Ramón y Cajal (1852-1934). Cajal contributed greatly to our understanding of the brain, not only in his writings but also by way of the fine drawings of his observations, which have always been heralded as a key element in conveying the evidence necessary to establish the neuron theory of the anatomy and physiology of the brain. Almost a century later, the world of brain research has gone one step further. Thanks to genetic recombination, scientists are getting proteins to draw for them. What is more, in colour and 3D... The artist's name is GFP – green fluorescent protein – a protein whose fluorescent properties have inspired many a researcher since its chance discovery in the 1960s.



A brainbow. Neurons are glowing all the colours of the rainbow by way of recombinant GFP.

Courtesy of J. Livet and J.W. Lichtman
Center for Brain Science, Harvard University

Not so long ago, the capacity to observe and translate an observation into a coherent drawing was a crucial component of a scientist's life – something on which he or she could base an emerging theory, or strengthen an existing one. In fact, the art of illustration is probably one of the qualities which built the very foundations of many fields of research today, until

photography and computers took the relay. Natural philosophers have been drawing plants since the beginning of the first millennium, although the most popular illustrations date back to the 18th century when various classification systems were thought up in an attempt to identify specimens. Astronomers painstakingly recorded the movements of planets. Anatomists and embryologists documented the different stages of animal and plant development, and palaeontologists, like archaeologists, spent hours recording fossil remains and ancient sites with a pencil and a notebook. Today, however, scientists can count on progressive tools such as cameras, powerful microscopes, telescopes, and, surprisingly, biotechnology to help them record what they observe.

GFP has been one of the top ten proteins in laboratories for years now because of its capacity to glow. It was discovered in the 1960s in *Aequoria victoria*, the Pacific Northwest jellyfish. It took a further forty years before its 3D structure was solved, and researchers were able to admire its rounded barrel shape – known as the β -can – with a chromophore hidden in its centre, where it is protected from assaults such as photochemical damage. Scientists pounced on the opportunity: here was a protein which could be used as a biological beacon.

¹ Cele Abad-Zapatero is an established scientist whose writing is at the crossroads of Science and Art. He is the author of 'Crystals and Life: A Personal Journey', International University Line 2002, and his play 'Bernal's Picasso' was staged at Argonne National Laboratory in 2008.

Tagged to numerous molecules in many different creatures – from the slime mould to humans – its glow can signal a specific molecule's location and movement in an organism. And what if GFP shone different colours? Then scientists could follow the concomitant migrations of different molecules in a given tissue... It didn't take long before scientists learned how to modify GFP fluorescence the way you would tailor a suit, and variants can now beam all the colours of the rainbow. The exciting part is that, by the very nature of any tissue, GFP can brush a three-dimensional image.

There have been several successive developments in chemistry and molecular biology that made these achievements possible. Recently, Lichtman and his team managed to combine different colour variants of GFP with a sophisticated system of genetic recombination – known as the Cre/Lox – which made it possible to paint, literally, inside the brain. The different recombination systems – suitably termed *Brainbows* – permit an exquisite recombination of the genetic elements and, by virtue of GFP, produce an amazing rainbow with a subtle

palette of hues, colours and textures. And when such technology is applied to certain areas of the brain, it can create images of an amazing beauty and, what is more, scientific insight.

The variants of GFP created by molecular recombination, such as has been described above, illuminate with magnificent colours the images that Cajal could only painstakingly illustrate in black and white. But even more exciting is the fact that GFP will not only give a far clearer – and more colourful – idea of the progression of given brain molecules in space, but also in time...thereby flirting with the meanders of our mind, our memories, our feelings and our conscience. The GFP paintbrush will sketch, in a wide array of colours and patterns, the neuronal events that make us what we are and what we feel, what we crave and what we despise. There is no doubt that Cajal would be astonished by – but also very proud of – the foundations that he laid down in his makeshift laboratory in a corner of his kitchen, with only his ink drawings and his precious Zeiss microscope as tools for introspection.

Cross-references to Swiss-Prot

Green fluorescent protein, *Aequoria victoria* (Jellyfish) : P42212

References

1. Livet J., Weissman T.A., Kang H., Draft R.W., Lu J., Bennis R.A., Sanes J.R., Lichtman J.W.
Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system
Nature 450:56-63(2007)
PMID: 17972876
2. Rapport R.
Nerve endings: The discovery of the synapse
Book. Publisher: W.W.Norton & co., 224 pp, 2005.
ISBN 0 393 06019 5
3. Lichterman B.L.
Book review of 'Nerve endings: The discovery of the synapse'
on www.bmj.com (2006)
<http://www.bmj.com/cgi/content/extract/332/7536/308>
4. Baillie Gerritsen V.
The greenest of us all
Protein Spotlight, issue 11, June 2001

National Nodes

Argentina

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

Australia

RMC Gunn Building B19,
University of Sydney, Sydney

Belgium

BEN ULB Campus Plaine CP
257, Brussels

Brazil

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

Chile

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

China

Centre of Bioinformatics,
Peking University, Beijing

Colombia

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

Costa Rica

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

Cuba

Centro de Ingeniería
Genética y Biotecnología, La
Habana

Finland

CSC, Espoo

France

ReNaBi, French
bioinformatics platforms
network

Greece

Biomedical Research
Foundation of the Academy
of Athens, Athens

Hungary

Agricultural Biotechnology
Center, Godollo

India

Centre for DNA Fingerprinting
and Diagnostics (CDFD),
Hyderabad

Italy

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

Mexico

Nodo Nacional 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 Nodes

CASPUR

Rome, Italy

EBI

EBI Embl Outstation, Hinxton,
Cambridge, UK

ETI

Amsterdam, The Netherlands

ICGEB

International Centre for
Genetic Engineering and
Biotechnology, Trieste, Italy

IHCP

Institute of Health and
Consumer Protection, Ispra,
Italy

ILRI/BECA

International Livestock
Research Institute, Nairobi,
Kenya

MIPS

Muenchen, Germany

UMBER

School of Biological
Sciences, The University of
Manchester,, UK

for more information visit our Web site

www.embnet.org



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://www.embnet.org/download/embnetnews>

Publisher:

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

Submission deadline for the next issue:

November 20, 2009