EMBnet.news

Volume 13 Nr. 3
September 2007

- EMBnet AGM 2007
- HoxPred
- Linux for bioinformatics
- Keeping your data up to date

and more ...

Editorial

The EMBnet.news editorial team dedicates this issue to our colleague Martin Sarachu. Three EMBnet members leave us their testimony about the way he lived and worked, and his legacy to our community. In this issue you can also find the report on our AGM held in Torremolinos, Malaga, Spain. As you may remember, we have held a joint meeting with the "Red Iberoamericana de Bioinformatica" (RIB), greatly enhancing the chances of cross collaboration. Also in this issue, you may find technical papers on several subjects of interest for Bioinformaticians, ranging from protein classification to reviews of tailored Linux distributions for Bioinformatics and e-learning tools. The publication of the technical article about the BioMacKit by Erik Bongcam-Rudloff and Alvaro Martinez Barrio, announced in the previous issue, has been postponed to the issue 13.4. The EMBnet.news team wishes you a pleasant reading and, as usual, invites you to participate by sharing your experiences with the community by submitting articles.

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

proteinspotlight

> ONE MONTH, ONE PROTEIN <

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

`http://www.expasy.org/spotlight`

We provide the EMBnet community with a printed version of issue 86. Please let us know if you like this inclusion.

Cover picture: *Iphiclides podalirius*, Ragusa Ibla gardens, Sicily, Italy 2007. [© Daniele D'Elia]

Contents

Editorial Board:

Martin Sarachu

Martín Sarachu was born on August 5th, 1976. His parents have been highly appreciated graduate students in my lab and close friends. Therefore, we have had a deep and nice relationship from his childhood to now.

When Martín finished High School, he had many interests but not a defined vocation. At that time, he was already an expert in Informatics and therefore I proposed him to take care of our small MicroVax server. He learned by himself VMS and GCG and was very helpful for many biologists at that time.

In 1997, he installed a new Sun Work Station and learned again by himself Solaris and operated the new GCG programs. In that year our node was accepted in EMBnet upon a proposal of Peter Rice that had given a course in our lab. From there on Martín was the manager of the node and in 1999 he finally decided to undertake a University career in Informatics.

In 2001, he installed EMBOSS and looked for GUIs that would facilitate its use by biologists. He produced the first version of wEMBOSS based on a former program of Luke McCarthy.

In 2003, in collaboration with BEN (M. Colet, G. Bottu and R. Herzog), an improved version of wEMBOSS was developed and several releases were distributed.

Later on, he produced in collaboration with G. Bottu a series of wrappers for EMBOSS.

He produced a Live CD intended both for demonstrations of wEMBOSS and for people that have bad network connections and do not have access to a bio-server.

He organized a node PC cluster to install SRS and a file server to update data bases. He developed an automatic pipeline to process a large number of sequences obtained in EST studies.

He married Daniela at the end of 2005. In February 2006, he was diagnosed a Philadelphia leukaemia. He responded for a while to the treatment with the newest inhibitors but his disease required a bone marrow transplant in January 2007.

He was treated in an excellent hospital but his leukaemia was extremely aggressive and he passed away on September 9th. He has been an essential collaborator for all courses and activities organized by the AR.EMBnet node. Martin´s shy, modest and kind personality hid an extraordinary intelligence and fantastic capacity for learning. His honesty and willingness of hard work and doing things right received the appraisal from everyone that collaborated with him.

Oscar Grau

Martin and I met for the first time in April 2000 at University of La Plata, the host University of the Argentinean EMBnet node. I was invited by Oscar Grau to give an EMBnet course of bioinformatics and had therefore to install some pieces of software and databases.

Martin was terminating a graduate in computer science; he was at the same time the system manager and the software developer of the node. Besides the work, the strong friendship between Oscar and Martin created a so pleasant atmosphere.

We did a good job together and that's probably why at the first opportunity Oscar, Martin and I started a collaboration (Accord de Coopération Bilatérale Belgique-Argentine, SSTC 2003), the fruit of this collaboration is the well known wEMBOSS interface to the EMBOSS suite.

The contract offered to Martin the opportunity to visit us at two different occasions. I will never forget those moments of hard work but also the pleasant Sundays with Martin: the walk along the delightful banks of Molignée River, with Liliane, my wife, Martin and me; the "pedalo" trip with my children at the seaside…

Why the fate may be so cruel to some of us? Martin deserved a longer life, a lot of other good moments.

Goodbye Martin, you are always present in my mind!

Marc Colet

At the Belgian EMBnet Node we have a long tradition of integrating third party software like BLAST, CLUSTAL or SRS first under GCG and later under EMBOSS. During EMBnet meetings collaborators of other nodes showed interest but unfortunately, in their original form, the "wrapper" programs were not easy to install on someone else's computer. In March 2004 our colleague Valérie Ledent went to the University of La Plata to give a BEN style bioinformatics course. It was my task to login in telnet session and install the necessary software on their server. I was in regular Email contact with Martin and it was at this occasion that he proposed to dedicate himself to the task of finding a way to make the "wrappers" readily portable, without making the code much more complicated and without tinkering with the installation of EMBOSS itself. He came up with the simple but efficient solution to make code with inserted tags and an installation script that edits and recompiles the code so as to adapt it to the local situation.

By mid 2004 the first public release of wrappers4EMBOSS was ready and we continued to work together on it till end 2006, when Martin became too ill. Oscar let us know that Martin was eager to go back to the job as soon as he was allowed to leave the confinement of his germ-free environment. Alas, then came the shocking message that he had passed away...

What can I do? The best way to pay homage to Martin is to work hard to get into the maintenance of the part of wrappers4EMBOSS that was his province, so that this tool remains available for the researchers that have found it very useful.

Guy Bottu

I first learned about Martín Sarachu through his development of wEMBOSS, an amazing piece of software that changed the Bioinformatics community by providing a free, simple and powerful web interface for EMBOSS. I had my first opportunity to meet him in person during a 2002 CABBIO course in La Plata, thanks to Prof. Oscar Grau, where I was deeply surprised to discover he was such an awesome, most gifted youth with a handsome and charming personality. Over the time we had various opportunities for collaboration and to meet again, finding common interests in and out of Science.

Ever since I first met him, I have been constantly impressed by his amazing achievements and his kind disposition, which made him a firm supporter of everybody and a hard worker for the common good, which won him a large extent of friends and gratitude in the community. All this time I held the highest expectations for his future, and sincerely believed he was sure to become a major star in the world of Science.

If my respect for him was great from the beginning, it was even bigger after he was diagnosed of Philadelphia Leukemia and we all could witness how boldly he dealt with it, keeping high spirits, maintaining a high professional profile and still managing to help everybody while fighting the disease, to the point it was difficult at times to believe he might be sick.

I have met many of you in conferences, meetings and courses, and I know we all shared a similar appreciation and gratitude for Martín. Many of us have come to depend heavily on Martín and his work to provide tools for scientists and relied on his help to develop new works of excellence. In addition, literally tens, or hundreds of thousands of Life Science researchers all over the world have been relying every day on the works launched by Martín to perform their work and will continue doing so for years to come.

It is unquestionable that thanks to Martín and his efforts we have been able to see many developments in modern Science with a deep impact in Society. Thanks to the ease of use and power that he brought to Bioinformatics tools, scientists can now perform complex analysis to unravel the mechanisms of Life and apply them in Biotechnology and Medicine to improve our quality of life.

When Oscar told me Martín had passed away I was overwhelmed. It is always a pity when a young person dies, mostly so when he was brilliant and affable; but for all of us, it has been worse, as in addition we have also lost a true friend. I am certain we will all sorely weep him. I know I do.

José R. Valverde

The EMBnet Annual General Meeting 2007

and EMBnet-RIBIO collaborative workshop

The Annual General Meeting 2007 of the EMBnet Stichting was organised in Torremolinos, Malaga (Spain) by Oswaldo Trelles, Jose Ramon Valverde and Oscar Grau.

We've got a participation of more than 80 people from 28 countries: Argentina, Australia, Austria, Belgium, Brazil, Chile, China, Colombia, Costa Rica, Cuba, Egypt, Finland, France, Greece, Kenya, Mexico, Netherlands, Norway, Pakistan, Portugal, Russia, Slovakia, Spain, Sri Lanka, Sweden, Switzerland, UK and Uruguay.

EMBnet-RIBIO collaborative workshop

Starting on June, 11th, the meeting was very intense over 3 days with 37 presentations, 2 poster sessions and a round table discussion.

Among the invited speakers we were honoured by the presence of famous bioinformaticians like Prof. Amos Bairoch, and Prof. Alfonso Valencia.

Topics:

- Computational genomics and Evolution
- Structural Bioinformatics
- Systems Biology and Databases
- Education and Training
- Cooperative Projects

Round table summary

After an intensive discussion five topics emerged as potential collaborative projects:

- Standards in Bioinformatics (JR Valverde)
- GRID distributed data, services and workflows (JR Valverde)
- Education (e-learning) (L. Falquet, JR Valverde)
- Light-weight Genome Browser P2P (L. Falquet, T. Attwood)
- Metagenomics (D. D'Elia, D. Holmes)

A RIBIO-EMBnet coordination group was created, with for EMBnet:

- JR Valverde (ES), L. Falquet (CH), S. Kossida (GR), T. Attwood (UK), D. D'Elia (IT)

and for RIBIO:

- J. Collado-Vides (MX), D. Holmes (CL), O. Trelles (ES), AT de Vasconcelos (BR)

This coordination group will stay in contact via video conferencing and will survey possible EU FP7 calls (or other sources of funding) that would fit with one or more of the potential projects.

This meeting was a great success showing the enormous potential of new collaborations that might appear in the future. This kind of meeting should be conducted with other networks like Asian-Pacific Bionet or African Bioinformatics network with the idea of creating a World-Wide network.

Ethnic party

The traditional "Ethnic party" took place on Wednesday evening on the beach. The delegates usually bring specialities of their countries. Noticed among others were special fruit wines from Belgium, typical seafood cans from Chile, and delicious bocadillos from Colombia. The local organisers served a wonderful and tasty paella accompanied by a refreshing sangria.

Annual General Meeting of the EMBnet Stichting

Present were the delegates from Argentina, Australia, Beligum, Brazil, Chile, China, Colombia, Costa Rica, Cuba, Finland, Hungary, Mexico, Pakistan, Portugal, Russia, Slovakia, Spain, Sweden, Switzerland, ILRI-BECA (Kenia), and UMBER (UK), while India, Italy, and Norway were represented by appointed proxies. Absent with apologies EBI (UK), South Africa, ICGEB (Italy). Other absent Canada, Israel, Poland, ETI (Netherlands), IHCOP-BGMO (Italy), MIPS-GSF (Germany).

Several observers from Belgium, Spain, Sweden, Switzerland, Greece (applicant), Sri Lanka (applicant).

The Minutes of the AGM 2006 in Sweden-Finland were approved with an amendment to the financial report correcting a sum.

Financial report

Our treasurer Oscar Grau provided a detailed report of the financial situation, annual fees and expected expenses. Counting a few pending fees from previous years, the assets are in slight augmentation compared to last year.

Re-election of nodes

As the EMBnet Stichting impose, a renewal of the membership is conducted every 3 years. This time the following nodes were up for re-election: Canada, Finland, Israel, Norway, Sweden, ICGEB and ETI.

Reports of their current status were provided on time for Finland, Norway, Sweden, and ICGEB. No reports were obtained from Canada, Israel and ETI. All members except Israel were re-elected. The israelian node is suspended and up for discharge at next AGM.

Confirmation of activity

Brazil was up for discharge since last AGM. The presence of Ana Teresa Ribeiro de Vasconcelos confirmed that they are still interested to be member of EMBnet. She is ready to serve as national delegate for the Brazilian Bioinformatics organisation, in reality a network of interconnected laboratories.

Brazil is confirmed on board of EMBnet.

Candidate new nodes

Greece, represented by Dr. Sophia Kossida of the Biomedical Research Found-ation (BRF) of the Academy of Athens. It is a non-profit institute dedicated to understanding, treating, and preventing human ailments through biomedical research. BRF seeks to serve science and medicine, and to participate fully in global innovation through its commitment to the true integration of biology, medicine, and informatics. The Bioinformatics & Medical Informatics team within the BRF has got several research interests within the field of Bioinformatics (`www.bioacademy.gr/bioinformatics`).

Sri Lanka, represented by Prof. Kamani Tennekoon of the Institute of Biochemistry, Molecular Biology and Biotechnology (IBMBB) of the University of Colombo. IBMBB provides human resource training at Masters and PhD level in Molecular Life Sciences and allied fields and has well equipped laboratories. Infrastructure for bioinformatics includes high end Intel Xeon workstations and a Grid computing facility connected to a similar facility at the University of Colombo, School of Computing and to University of Uppsala (sweden). It is proposed to house mirrors of biological databases which can be accessed by the researchers in the Sri Lankan Universities and Research Institutes which are networked through Lanka Education and Research Network.

Both national nodes of Greece and Sri Lanka were elected as new members of EMBnet.

Report from the Executive Board

The secretary Robert Herzog reports about the successful monthly Virtual General Meetings (VGM). The attendance was satisfactory with between 15 to 25 attendees. No technical difficulties were encountered and very fruitful discussions and decisions were taken during those meetings, namely:

- The selection and organisational model of this AGM
- Discuss the ongoing activities of the EMBnet program committees
- Start discussing an application for one or more FP7 grants
- Effectively revive the quarterly newsletter EMBnet.news
- Take decisions about the acquisition of the EMBnet hardware
- Discuss the options about the EMBnet CMS (Drupal was chosen and implemented by Cesar Bonavides)

All members are encouraged to take part of these VGMs.

The EB took several actions during the reporting period. The account M. Huub Clemens was officially dismissed and replaced by our treasurer Oscar Grau. The account of the Fortis bank was converted to an internet accessible account, allowing easier management from anywhere in the world. The official address of the Stichting was moved from The Netherlands to Belgium:

The EMBnet Stichting
Université libre de Bruxelles
Bioinformatics Laboratory
c/o Robert Herzog
Rue des Prof. Jeener & Brachet 12
B6041 Gosselies – Belgium

It is noted that for any official document to be produced by the Stichting (e.g., an invoice), we should also mention the registration number of the EMBnet Stichting in The Netherlands, namely:

41058777- Stichting EMBnet

Reports from the committees

The Education & Training Committee (E&T PC) activities were reported by Vassilios Ioannidis.

Regular e-meetings were organised using Marratech or Skype. Lisa Mullan resigned at the end of 2006 and was replaced by Jingchu Luo. Unfortunately no news was obtained from Isabelle Marques.

Activities of the E&T PC were mainly the creation of a repository for courses of the EMBnet community via a CMS, the investigation on the licensing of this material, and the creation of new material like QuickGuides, animated tutorials and interactive courses via Marratech. The Moodle CMS was choosen and is now active from our main web site:

`http://education.embnet.org`

Other administrative issues related to mailing lists and certification are under investigation. The E&T PC asks again all members to contribute to the material. Especially welcome are courses in various local languages which is one of the strength in EMBnet.

The Technical Management Committee (TM PC) activities were reported by Nils-Einar Ericsson and Cesar Bonavides.

The ownership of the embnet.org domain has not yet been transferred to the EMBnet Stichting. We plan to achieve this final and formal step of moving the domain administration shortly. A solution for the backup of the main server is still under evaluation. Uppsala University charges 12 euros per 10Gb of data. Marratech was bought by Google Inc. and the free server option is no longer available. Fortunately the current license we bought is not limited in time. The Drupal portal is fully functional and already many members use it to update their information and provide valuable input. *Every member is encouraged to maintain and update its own data.*

JR Valverde mentions that a project management system (dotProject) is also installed on the main server and available to the members. Moodle has now an RSS feed functionality.

The Publication & Public Relation Committee (P&PR PC) activities were reported by Pedro Fernandes.

The P&PR PC has allocated most of the efforts in preparing and publishing EMBnet.news.

In 2006 Vol 12 had two issues and in 2007, Vol 13 will certainly regain speed and have 4 issues. The current issue of EMBnet.news (13.2) has been made available in paper at the Collaborative Bioinformatics 2007 meeting, in order to better enhance EMBnet's image, and take our messages further throughout the world.

A poster on EMBnet, prepared by Kimmo Mattila was presented in Brazil and will also be on display in Torremolinos, at Bioinformatics 2007, the joint meeting of EMBnet and RIB (Red Iberoamericana de Bioinformatica). Another poster on EMBnet was prepared and presented by Domenica D'Elia, at the Annual Meeting of the Bioinformatics Italian Society (BITS).

Lubos Klucar has prepared a discussion forum for our activities on the new webserver of EMBnet. This forum enables us to upload submitted articles and make the editing of the issues less labour intensive.

The P&PR PC believes that correctly funded publicity activities enhance EMBnet's outreach. The way we operate now is artificially sustained by the fact that we publish electronically at virtually no cost. In truth we should be using a proper budget to get known in scientifically and politically influent circles, and this is not being done. The P&PR PC believes that joining efforts with other networks can enhance EMBnet's position in fundraising situations, namely in what concerns RIB, and is ready to provide support to such activities.

After a discussion it appeared that a periodic activity report gathering information from all nodes could be the best way to fulfil this task.

Report from the Task Force FP7

The task force activities were reported by Laurent Falquet. The Task Force FP7 was created after an initial discussion started by JR on the emb-adm mailing list.

We met 5 times from March to May via Marratech with a variable list of participants:

Erik (SE), JR (ES), Domenica (IT), Andreas (IT), Valérie (BE), Richard (BE), Etienne (BECA), David (BE), Sofia (GR), Lubos (SK), Cesar (MX), Emiliano (CO), Shahid (PK), Laurent (CH).

We collected information on the support we could get from our local institutions or countries to prepare a proposal. The data is available on the dotProject web site (installed by JR). Unfortunately because we started to discuss in March only, none of the possible projects was ready for beginning of May, we missed the first round of Calls.

Lessons from our first experience:

- Be active in watching Calls
- Be reactive when a Call could fit
- Just a few weeks/months to respond with a full proposal!
- Warning: submission can be in one or two steps

During RIBIO-EMBnet round table discussions, a coordination group was created (See first page).

Past and Future activities

The RIBIO collaborative meeting held at the same place on June 11-13. As all members present attended this meeting, JR Valverde briefly mentions the great success of this meeting where about 40 people from the Latin-American region joined the EMBnet members (see first page).

The Pakistan scholarship program

Chohan Shahid comments about this project by the Pakistan government to fund the stay abroad of about 50 students order to obtain a doctorate in bioinformatics. Money will be made available for travel, residence and day-by-day expenditures for the students. The program needs to identify countries, institutes and supervisors to serve as hosts for this project. EMBnet members are encouraged to react positively. Seven countries have been mentioned in an initial phase (Austria, France, Germany, The Netherlands, Norway, Sweden and China). Other might join provided our Pakistan managers obtain agreement from their authorities.

APBionet

Jingchu Luo comments about the present status of APBionet. This is not (yet) a structured entity like EMBnet. There are no AGM, no fees, no funding, etc. It is more like a loose network of partners with the same domain of interest, in the same area of the World. Winston Hide from South Africa is among the most active partners.

SIMDAT

As Valérie Ledent and Richard Kamuzinzi (BEL) gave presentations about SIMDAT and its possible impact on EMBnet during the RIBIO meeting, no further comment is given here.

Federated Blast

David Coornaert (BEL) gives a presentation on a possible way to exploit relatively modest hardware resources in order to optimize the speed of BLAST searches against large databanks. He shows that the speed is directly dependent on the proper configuration of the processor-storage media connection. Given a properly accomplished distribution of the data on several modest computers, a configuration where all the data are stored in RAM allows for very fast BLAST runs. David suggests that a distribution of large datasets among disperse computers deserves experimentation (a prototype running on three nodes over a Belgium-Slovakia connection proved perfectly functional).

MRS

Robert Herzog briefly comments about the availability as open source of the MRS databank management software, produced by Maarten Hekkelman at the University of Nijmegen. This is an easy to manage (only Perl and Make scripts are used for overall management) and efficient alternative to SRS, which is no longer available for free, even for academic partners. One of the unique functionalities of MRS reside in its rewritten BLASTP algorithm that allows to search a subspace of the protein sequence databanks expressed as an MRS query (e.g. search my sequence of interest against all sequences from the protein databanks where species is "mus" and description line contains the word "tumor" or the word "cancer"). MRS has been described in recent issues of embnet.news.

EMBRACE

As Erik Bongcam-Rudloff gave a presentation about EMBRACE during the RIBIO meeting, this is not repeated here.

HealthGRID

Erik mentioned he could not attend the HealthGRID meeting this year, but contacts with EMBnet are continued. EMBnet effectively sponsored the initial HealthGRID meeting a couple of years ago.

Elections

Executive Board

Robert Herzog announced he wants to step down after serving three years in the EB. Robert is now retired from his position at the university and is progressively loosing contacts with his formers co-workers inside the Belgian EMBnet node. During the official dinner the chairman of EMBnet thanked Robert for his 3 years of duty and offered him a special gift.

We all would like to thanks Robert and wish him a nice retreat spending time on his favourites hobbies like aeromodelling!

Two members stepped forward as candidates for a position in the EB, namely Teresa Attwood (UMBER) and JR Valverde (National node Spain). Both members are eligible, as JR announces he intends to step down from his position in the TMPC.

A secret vote elected JR Valverde for 3 years. Welcome JR and be ready for the good work!

For next year, the EB of EMBnet consists of Erik Bongcam-Rudloff, chairman, Oscar Grau, treasurer, Laurent Falquet, secretary and Jose-Ramon Valverde, member.

Education & Training

Except for Jingchu Luo, all positions in the ET PC are open for re-election. Candidates are Sophia Kossida (GR), Vassilios Ioannidis (CH), Georgina Moulton (UMBER) and Valérie Ledent (BE). All were elected, Sophia Kossida (see previous page) and Georgina Moulton are new members.

For next year, the ET-PC consists of: Valérie Ledent (Belgium) - chairperson, Vassilios Ioannidis (Switzerland) – secretary, Jingchu Luo (China) - treasurer, Georgina Moulton (UMBER, UK) and Sophia Kossida (Greece) - members.

Technical Management

No change with the step down from JR, the TM PC decided to keep 4 members and consists of: George Magklaras (Norway), Nils-Einar Eriksson (Sweden), Cesar Bonavides (Mexico) and David Coornaert (Belgium).

Publications & Public Relations

Pedro Fernandes was up for re-election and was brilliantely re-elected.

For next year, the P&PR PC consist of : Kimmo Mattila (Finland), Lubos Klucar (Slovakia), Domenica D'Elia (Italy) and Pedro Fernandez (Portugal).

Date & place of next meeting

Due to lack of time, the date and place for the next EMBnet AGM is delayed until the next VGM. Candidates are invited to present their project during next VGMs according to the following schedule (every second Tuesday of the month at 4pm CET):

- Sept. 11, 2007 First presentation
- Oct. 9, 2007 Detailed budget
- Nov. 13, 2007 Vote for next AGM

Concluding remarks

Our chairman Erik Bongcam-Rudloff thanks everybody for the good work and closes the meeting at 18:40. Cesar Bonavides gets the opportunity to comment a little bit more about the use of the Drupal software behind the EMBnet portal and mentions he opened access to "external registered users" in order for RIBIO participants to be able to contribute to the think tank about possible source of financing, notably the 7FP of the EU. Participants disperse for the dinner at 21:00 in the Casa Juan restaurant in Torremolinos.

Laurent Falquet & Robert Herzog

HoxPred: an automated procedure to classify Hox proteins

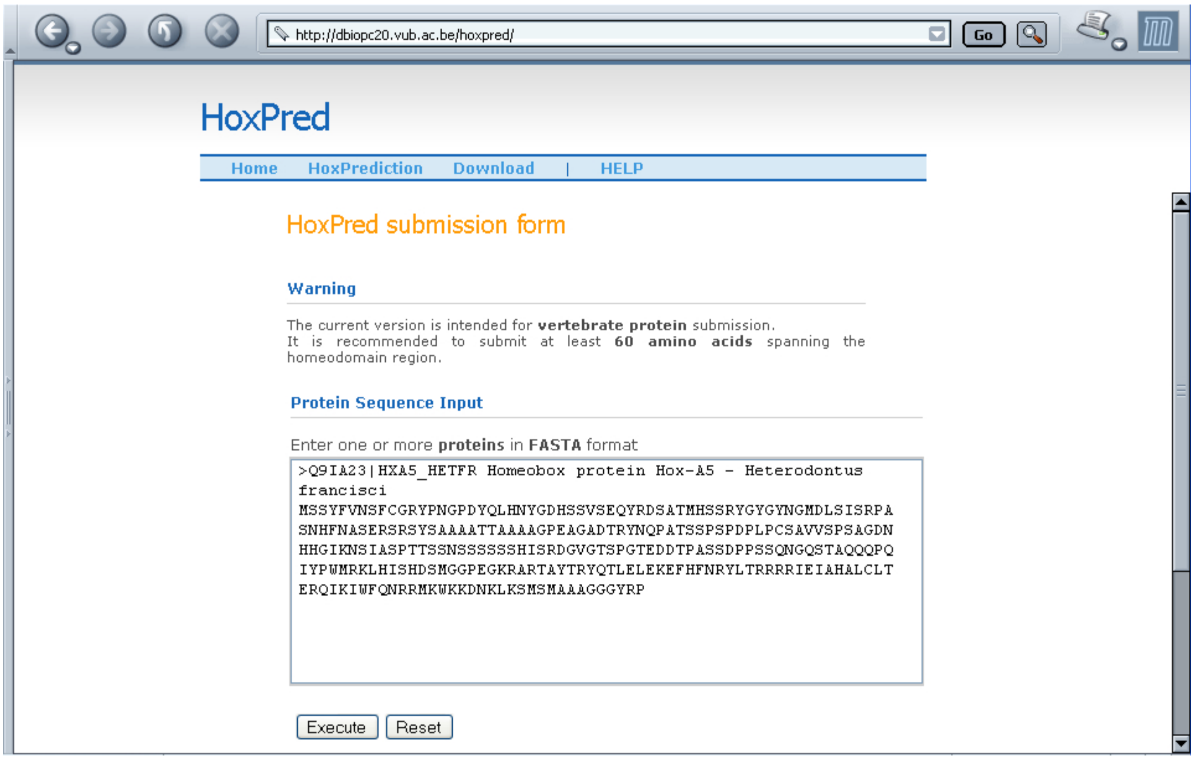
Morgane Thomas-Chollier and Valérie Ledent
BEN, ULB Campus Plane CP, Bruxelles, Belgium

Introduction

Correct identification of individual Hox proteins is an essential basis for their study in diverse research fields of molecular and evolutionary biology. This protein family is best known for its crucial role in patterning the anterior-posterior axis of animal embryos and in tetrapod limb development.

Hox genes actually belong to the family of homeobox transcription factors characterised by a 60 amino acids region called homeodomain. Besides, the genomic organisation of Hox genes in clusters is common to most animals. An ancestral Hox gene cluster, supposed to have arisen from tandem duplications in early eukaryotes, has been retained in bilaterians. This ancestral cluster has been duplicated early in the vertebrate lineage. Mammals Hox genes are organised in four clusters whereas teleost Hox genes are generally arranged on 7 clusters, due to an additional duplication specific to teleost fishes. Lineage-specific gene loss has subsequently occurred, leading to diverse presence/absence combinations of Hox genes.

Common methods to classify Hox proteins in their group of homology rely on sequence similarity and phylogenetic analysis. These methods commonly focus on the homeodomain. Classification of Hox proteins is thus hampered by the high conservation of this short domain. Since phylogenetic tree reconstruction is time-consuming, it is not suitable to classify the growing number of Hox sequences.



HoxPred submission form

Warning

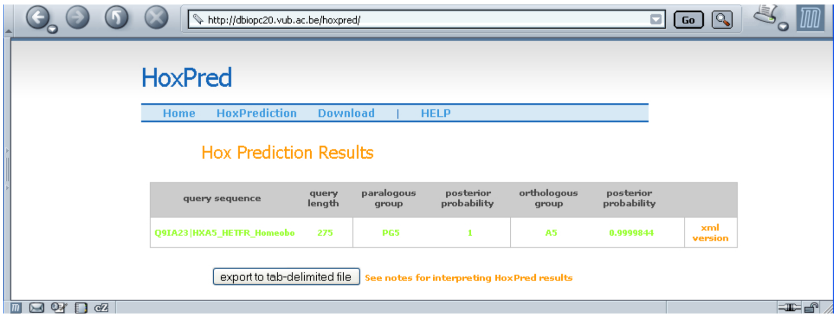
The current version is intended for **vertebrate protein** submission.
It is recommended to submit at least **60 amino acids** spanning the homeodomain region.

Protein Sequence Input

Enter one or more **proteins** in **FASTA** format

```
>Q9IA23|HXA5_HETFR Homeobox protein Hox-A5 - Heterodontus
francisci
MSSYFVNSFCGRYPNGPDYQLHNYGDHSSVSEQYRDSATMHSSRYGYGYNGMDLSISRPA
SNHFNASERSRSYSAAAATTAAAAGPEAGADTRYNQPATSSPSPDPLPCSAVVSPSAGDN
HHGIKNSIASPTTSSNSSSSSSHISRDGVGTSPGTEDDTPASSDPPSSQNGQSTAQQQPQ
IYPWMRKLHISHDSMGGPEGKRARTAYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLT
ERQIKIWFQNRRMKWKKDNKLKSMSMAAAGGGYRP
```

Figure 1. HoxPred Web interface: Submission Form.



http://dbiopc20.vub.ac.be/hoxpred/
Go
HoxPred
Home
HoxPrediction
Download
|
HELP

Hox Prediction Results

query sequence	query length	paralogous group	posterior probability	orthologous group	posterior probability	
Q9IA23|HXA5_HETFR_Homeobo	275	PG5	1	A5	0.9999844	xml version

export to tab-delimited file
See notes for interpreting HoxPred results

Figure 2. HoxPred Web interface: Hox Prediction Results.

The project to develop a bioinformatics tool able to automatically perform the classification of Hox proteins was born from a close collaboration between the Belgian EMBnet Node (BEN) at the Bioinformatics laboratory, Université libre de Bruxelles, and the Developmental and stem cell biology laboratory of Prof Luc Leyns and the Amphibian Evolution laboratory of Dr. Franky Bossuyt at the Vrije Universiteit Brussel.

HoxPred Procedure and access

Morgane Thomas- Chollier, a PhD student, has developed HoxPred, an automated procedure to classify Hox proteins in their groups of homology. The method relies on a discriminant analysis that classifies Hox proteins according to their scores for a combination of proteic generalised profiles. 54 generalised profiles dedicated to each Hox homology group were produced de novo from a curated dataset of vertebrate Hox proteins. Several classification methods were investigated to select the most accurate discriminant functions. These functions were then incorporated into the HoxPred program accessible via SOAP and via an easy-to-use Web interface at `http://cege.vub.ac.be/hoxpred/` (Figures 1 and 2).

HoxPred in action

Applied on a curated dataset of vertebrate Hox proteins, HoxPred shows a mean accuracy of 97% in cross-validation tests. We demonstrated that HoxPred is appropriate to decipher Hox proteins from whole genomes by applying it on two telost fishes, Oryzia latipes (medaka) and the recently sequenced Gasterosteus aculeatus (stickleback). HoxPred predictions are largely correct even though teleost Hox sequences are known to be divergent consequently to additional duplication of their Hox clusters.

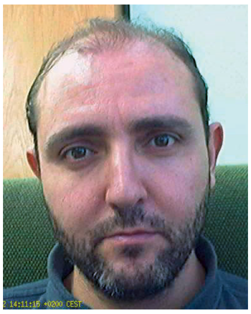
We also tested HoxPred on a wide range of proteins by applying it to the UniProt databank. The Hox content of many organisms is often analysed by PCR surveys that produce very short sequence fragments. The Uniprot databank thus comprises many short Hox sequences (<60 residues). As profile scoring-system is length-dependent, input protein fragments for HoxPred should be at least as long as the profile (60 residues) and span the homeodomain. Despite this limitation, we observed that predictions are quite robust to short fragments and we showed that HoxPred could help identifying homologous groups in PCR surveys.

These results were recently published in the journal BMC Bioinformatics ("HoxPred: automated classification of Hox proteins using combinations of generalised profiles " by Thomas-Chollier M., Leyns L. and Ledent V, BMC Bioinformatics 2007, Jul 12; 8(1): 247).

Remarks

HoxPred can efficiently contribute to a better annotation of Hox in vertebrate. It correctly discriminates Hox sequences from non-Hox homeoboxes, including the closely related paraHox proteins. HoxPred classifies sequences into paralogous and orthologous homology groups. The program is particularly appropriate for automatic classification of Hox proteins into their paralogous groups. As orthologous group predictions show a higher risk of missclassification, it is better to corroborate them with additional supporting evidence.

Taking education beyond the classroom



Enrique de Andrés Sáez, David J. García Aristegui, Germán Carrera, Alfredo Solano and José R. Valverde

EMBnet/CNB, CNB/CSIC,
C/Darwin, 3, Madrid 28049

An Agony in Eight Fits [1]

Fit the First: The Landing

"...I have said it thrice:
What I tell you three times is true."

Earlier this year, the Biocomputing Unit of CNB had to organize an international course with a strong practical, hands-on component. The road to success was laden with stumps, traps and emotions. This is a tale of the problems we faced, the solutions we probed and the final result of our efforts. We believe many will see their own experiences reflected in our own odyssey and hope that you find this tale as exciting to read as it was for us to live.

Fit the Second. The bellman's speech

"Friends, Romans, and countrymen, lend me your ears!"

Attention, spoiler ahead. If you don't want to know the end of this story in advance, please do jump over this short introduction over to the next section.

On February 2007 we gave a hands-on course on **Workflows, Web Services and Grid Computing** at CNB/CSIC with a strong hands-on practical component. According to the polls at the end of the course, it was a success among students (although they missed even more practical sessions). During this course students had access to remote computers using each a local Linux machine with full functionality which gave them a powerful environment to perform all the tasks of the course. Further to it, they were given these machines and could take them home (*for free!!*), fully configured so they could continue working seamlessly at home with them.

Fit the third. The baker's tale

"But oh, beamish nephew, beware of the day,
If your Snark be a Boojum! For then
You will softly and suddenly vanish away,
And never be met with again!"

The course was an ambitious collaborative initiative organized by the EMBRACE [2] and 3D-EM [3] networks of excellence, the EGEE [4] project and EMBnet [5]. Actually it was two courses: one aimed to bioinformaticians and a second one aimed to 3D-EM electron microscopists, both being held simultaneously and sharing the last practical day so wet-lab scientists and application developers could work together to know each other ways.

Having two simultaneous courses imposed a major restriction: we needed to have two rooms for training, but our facilities at CNB had been already widely over-allocated for more than a year. Luck came to favour us in the way of a building expansion that were to be completed just one month in advance to the course, and we were allowed to make use of two labs for the course as long as their owners had not moved in yet. This gave us some hope although just in case we also made reservations of two UAM [6] training computer rooms.

Building works delayed a bit and it was only two weeks earlier that the new labs were available. Obviously their owners would not be able to come in before the course, so we decided to use them. For the experimental course on 3D-EM it might be an excellent solution, but when we came to see the labs it soon became obvious that they were not good for us as visibility was seriously hampered.

We might do with a seminar room, but how would we do the practices? We had bought 10 laptops recently for training, but both courses needed computers, and we had already decided that the wet lab students were to use them. Oh, well, who cares? We are a biocomputing shop and have lots of used computers, we could resort to **use retired old machines as X terminals and do all the work on servers**. But when we tried, these machines resulted too big to fit comfortably on our seminar room, and they had no wireless support which meant deploying an '*ad hoc*' temporary wired network making it even less suitable as a training solution.

Thus we resorted to use one of the **University training computer rooms**. By now we were one week ahead of the course, and being a cautious bunch of guys we decided to make a full test of all the tools we were to use and ensure everything would work smoothly.

On Monday, one week before the course we went to the room we had reserved. Our first surprise came when we tried to **dual-boot into Linux** and machines systematically failed. A short talk with the manager confirmed that a network problem had isolated the room a few days ago from the Linux boot servers somewhere else in the University (a carterpillar had accidentally severed some network lines), and they could not assure when the service would be restored.

Nothing serious, we thought: we can **build a knoppix** [7] **like tailored distribution and boot from it**. As we were already short on time we hurried to our office and took several live CD/DVD distributions (hey *we are cautious*) and came back to the computer room only to discover that -wisely enough- the system administrators had disabled booting from anything but the network (well, to discover that and to feel like fools for not having thought of it). At this point we were four working days ahead, we could only boot on Windows, not from CD/DVD and still hadn't started testing the tools we needed on the computer room. For a paranoid team like ours it was starting to look worrying.

Now, we bet that you have some time or other been in a similar situation or fear you might find yourself in one, otherwise you would have not been reading this far. So by now, you may be pondering what we did to escape this tar pit. Read on and you'll find out.

Fit the Fourth: The hunting

"I said it in Hebrew—I said it in Dutch—
I said it in German and Greek:
But I wholly forgot (and it vexes me much)
That English is what you speak!"

For many people that would have been the end of the story but we were *tough*. We tried to **go for Windows using an X-server on it**, deferring all the work to our main servers: we allocated nodes from our grid cluster, one for each student and pre-configured them with the tools, but then the Windows Xserver was not well fitted to our needs and besides, we saw this as a horrible hack as we would be taking the machines off the very Grid facility the students were supposed to learn and use and some tools would not be fully usable.

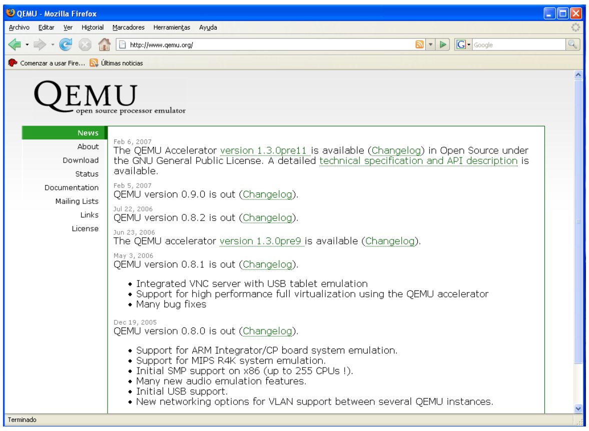
The obvious approach in this case was to **use a virtualizer**: we could not run Knoppix on the computer room machines, but we could run a live distribution *on virtual machines*. We checked we could install and run applications and hurried to our rooms to burn a new set of DVDs, this time containing a live image and the virtualizing software, **VMware Player** [8]. Cautious as you should have discovered by now we are, we gave it a first try on our machines and it run wonderfully. While some of us started installing the software, others run to the computer room to re-verify it again and once more feel like fools for it didn't work. Obviously enough, we could not install the virtualizing software.

We had expected to run VMware Player on the machines, just as we could run any other application, but had forgotten an important issue: a virtualizer gives the hosted system direct access to many of the host computer resources to allow it to run efficiently. This requires administrator privileges which we all had on our machines but lacked on the University computers and hence we could not use it. Asking the University system managers to do it for us was not an option as we were now too short on time.

Two days to go and still in high spirits (hey, we are *tough* guys), we decided to go for the next option (yes, we always have a hidden card under the sleeve) which was to **use an emulator**: this also uses a virtual machine, but this time it does

Copyright © 1998-2007 VMware, Inc. All rights reserved. Protected by one or more of U.S. Patent Nos. 6,397,242, 6,496,847, 6,704,925, 6,711,672, 6,725,289, 6,735,601, 6,785,886, 6,789,156, 6,795,966, 6,880,022, 6,944,699, 6,961,806, 6,961,941, 7,069,413, 7,082,598, 7,089,377, 7,111,086, 7,111,145, 7,117,481, 7,149,843 and 7,155,558; patents pending.

Figure 1. Starting VMWare Player installation (after free downloading from `www.vmware.com`).


QEMU - Mozilla Firefox
Archivo Editar Ver Historial Marcadores Herramientas Ayuda
http://www.qemu.org/
Google
Comenzar a usar Fire... Últimas noticias
QEMU
open source processor emulator
News
About
Download
Status
Documentation
Mailing Lists
Links
License
Terminado

Feb 6, 2007
The QEMU Accelerator version 1.3.0pre11 is available (Changelog) in Open Source under the GNU General Public License. A detailed technical specification and API description is available.

Feb 5, 2007
QEMU version 0.9.0 is out (Changelog).

Jul 22, 2006
QEMU version 0.8.2 is out (Changelog).

Jun 23, 2006
The QEMU accelerator version 1.3.0pre9 is available (Changelog).

May 3, 2006
QEMU version 0.8.1 is out (Changelog).

- Integrated VNC server with USB tablet emulation
- Support for high performance full virtualization using the QEMU accelerator
- Many bug fixes

Dec 19, 2005
QEMU version 0.8.0 is out (Changelog).

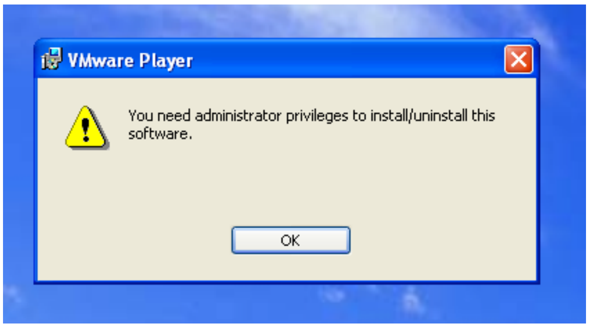
- Support for ARM Integrator/CP board system emulation.
- Support for MIPS R4K system emulation.
- Initial SMP support on x86 (up to 255 CPUs !).
- Many new audio emulation features.
- Initial USB support.
- New networking options for VLAN support between several QEMU instances.

Figure 3. The home page of QEMU (`www.qemu.org`).

not give the hosted system access to the hosting hardware; instead, it takes instructions one by one and translates them simulating a complete machine in software, with no aid from the local hardware. As a result, an emulator needs not administrator privileges and can be run in any machine. We substituted VMware by **QEMU** [9], a popular open software emulator for a large number of architectures and continued installing software on the virtual machines while doing the tests.

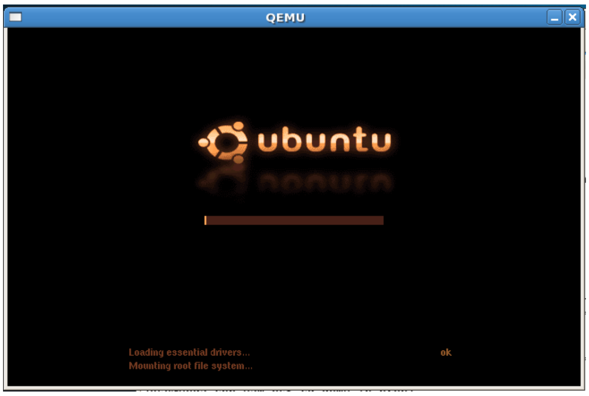
This time our test run flawlessly on the computer room machines: we could install QEMU, run it and launch the virtual machines with Linux without any need for administrator privileges. Great! time to hurry up as we had only three days to start (and two of them fell on week-end). Working together we could soon install all the software needed and check it on our machines in one day. To speed work, we took advantage of the *QEMU accelerator*: a system module that converts QEMU in a virtualizer rather than an emulator. This allows QEMU to run very efficiently, at almost native hardware speeds, altough obviously, it requires administrator privileges to be installed (just like VMware). Indeed being open source, QEMU does an excellent work both as an emulator and a virtualizer, competitive with commercial products like VMware (it is reported to be only 4% less efficient) with the added advantage that it can provide both (emulated and virtualized) kinds of virtual machines and create disk images that are compatible with VMware. For us this was an excellent solution: we could install all the software needed at almost native speeds using the QEMU virtualizer and students would be able to run the systems (about 4-10 times slower) using an unprivileged account.

The problem came when at the end of the day we tried again to run the systems: our earlier tests had behaved nicely, but now it crawled to an unbearable snail speed. We realized that our tests had been done with hard disk images, not



You need administrator privileges to install/uninstall this software.

Figure 2. Installation of VMWare Player fails due to lack of administrator privileges.



```
Loading essential drivers...                              ok
Mounting root file system...
```

Figure 4. Screenshot of the Ubuntu DVD provided to students booting under QEMU.

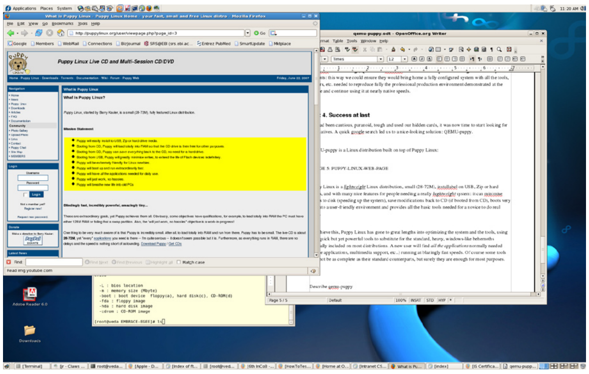
directly from DVD, and the extra tools had imposed a very heavy toll: we had installed Ubuntu Linux10, with a web server, tomcat, all the development and support environments, sshfs to access remote data, we wanted to use eclipse which required java interpretation, and so on... Still, as students would be able once home to install QUEMU and the virtualizer module to run the machine from hard disk, we decided to go ahead: this way we could ensure they would bring home a fully configured system with all the tools, servers, etc. needed to reproduce fully the professional production environment demonstrated at the course and continue using it at nearly native speeds.

Fit the Fifth. The beaver's lesson

They returned hand-in-hand, and the Bellman, unmanned
(For a moment) with noble emotion,
Said "This amply repays all the wearisome days
We have spent on the billowy ocean!"

We had been cautious, we had been paranoid, we had been tough and and we had used our hidden cards, it was now time to start looking for alternatives. A quick Google search led us to a nice-looking solution: QEMU-puppy, a Linux distribution built on top of **Puppy Linux** [11]:

Puppy Linux is a **lightweight** Linux distribution, small (28-72M), installable on USB, Zip or hard drives, and with many nice features for people needing a really lightweight system: it can minimize writes to disk (speeding up the system), save modifications back to CD (if booted from CD), boots very fast into a user-friendly environment and provides all the basic tools needed for a novice to do real work.



What is Puppy Linux - Puppy Linux Home - your fast, small and free Linux distro - Mozilla Firefox

http://puppylinux.org/user/viewpage.php?page_id=3

Puppy Linux Live CD and Multi-Session CD/DVD

What is Puppy Linux?

Puppy Linux, started by Barry Kauler, is a small (28-72M), fully featured Linux distribution.

Mission Statement

- Puppy will easily install to USB, Zip or hard drive media.
- Booting from CD, Puppy will load totally into RAM so that the CD drive is then free for other purposes.
- Booting from CD, Puppy can save everything back to the CD, no need for a harddrive.
- Booting from USB, Puppy will greatly minimize writes, to extend the life of Flash devices indefinitely.
- Puppy will be extremely friendly for Linux newbies.
- Puppy will boot up and run extraordinarily fast.
- Puppy will have all the applications needed for daily use.
- Puppy will just work, no hassles.
- Puppy will breathe new life into old PCs

Blindingly fast, incredibly powerful, amazingly tiny...

qemu-puppy.odt - OpenOffice.org Writer

4. Success at last

U-puppy is a Linux distribution built on top of Puppy Linux:

E 5: PUPPY-LINUX-WEB-PAGE

Describe qemu-puppy

```
-L : bios location
-m : memory size (Mbyte)
-boot : boot device  floppy(a), hard disk(c), CD-ROM(d)
-fda : floppy image
-hda : hard disk image
-cdrom : CD-ROM image

[root@veda EMBRACE-EGEE]# ls
```

Figure 5. Home page of Puppy Linux.

To achieve this, Puppy Linux has gone to great lengths into optimizing the system and the tools, using tiny, quick, but yet powerful tools to substitute for the standard, heavy, windows-like behemoths normally included on most distributions. A new user will find *all the applications* normally needed (office and network applications, multimedia support, etc..) running at blazingly fast speeds. Of course some tools may not be as complete as their bigger counterparts, but surely they are enough for most purposes.

Fit the Sixth: The barrister's dream

"They sought it with thimbles, they sought it with care;
They pursued it with forks and hope;
They threatened its life with a railway-share;
They charmed it with smiles and soap."

QEMU-Puppy [12] is a Linux distribution that fits into a USB key, and can be booted natively or on top of QEMU. If booted off the USB key, the PC will run Linux at native speeds. If that is not possible (as was our case) then it can be started by a double click on the 'puppy' icon on any system supported by QEMU (Linux, most UNIX systems and Windows) to be run under emulation. If in addition, the host operating system supports the QEMU virtualizer (Linux, Windows..), then it may be installed to run the system under virtualization at nearly native speeds. This makes QEMU-Puppy ideal for situations when one needs a system to carry around (e. g. work and home, Internet-cafés, travel, conferences...). All that is needed is a PC or workstation and a free USB port to plug the key in.

The ability to boot on QEMU adds various advantages: first one needs not exclude the pre-existing OS on the machine (as would be the case on an USB boot), second a direct boot may not recognize all devices, under QEMU it can be run concurrently with the host OS (hence giving access to both worlds simultaneously), a virtual machine provides an "idealized" fully featured hardware (hence you may use the underlying devices -sound, video, network- from the QEMU system transparently, and it is easier to secure (the virtual machine accesses the network through QEMU's built-in NAT server). QEMU-Puppy may be run using the full screen or as a window, and you can switch among host and guest OS pressing simultaneously the [CTRL] and [ALT] keys.

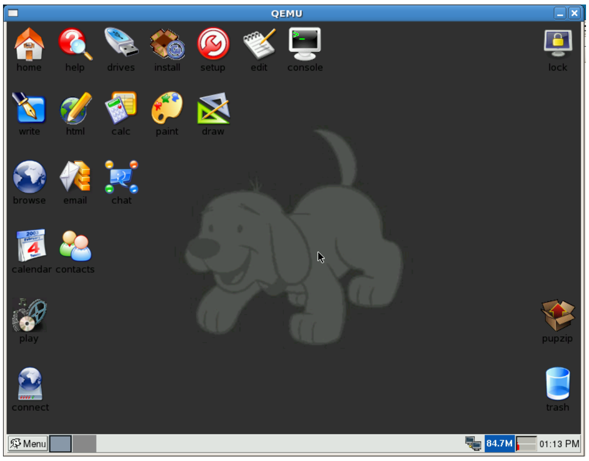


Figure 6. QEMU-Puppy running.

Having a USB key that can be both booted natively and as a virtual QEMU machine is one of the nice tricks of this distribution, but not the only one: as the virtual machine (if run under QEMU) accesses the Internet through a NAT private address, some protocols requiring a returning external connection (most notably FTP) may fail. QEMU-Puppy also includes a "virtual" FTP "server" to fix this problem so that running FTP from within the virtual machine will work seamlessly. It also supports SMB (so you can transfer files between QEMU-Puppy and a host windows machine easily, and of course NFS to share files with a UNIX host.

As QEMU-Puppy is based on Puppy, it runs very fast -even under emulation- and provides a complete environment. The file system is divided into several pieces, some of which can be exchanged to provide extra functionality (like e.g. adding development tools) and it comes with a simple yet effective package manager to automate installation of software.

In short: QEMU-Puppy provides a *very lightweight and fast* environment that can be carried around on a USB key, directly bootable or runnable under emulation (virtualization if you have the privileges), complete, easy to use, configure and maintain. Using QEMU-Puppy is as simple as booting off the key or double-clicking on its icon and the system may be extended easily.

Fit the Seventh: The banker's fate

"Leave him here to his fate—it is getting so late!"
The Bellman exclaimed in a fright.
"We have lost half the day. Any further delay,
And we sha'nt catch a Snark before night!"

By now we were already off-time. We had spent Friday developing a DVD with all the tools as a virtual QEMU machine, and the weekend was on top of us. Time for some homework. After making a QEMU-Puppy master USB key we left for home. During the weekend it was a trivial snap to substitute the default environment by a developer file system, and using DotPup/PupGet (the graphical package managers) adding most extra tools was painless (we could add very powerful programmers' text editors, compilers, etc... with little effort). Once the basic tools in place we easily compiled add-on tools like SSHFS [13] (a file system based on SSH and FUSE [14] which would allow secure sharing of data with the remote home directory in our Grid User Interface host). All along, some of us would run in parallel the final tests of the DVD with QEMU, Ubuntu and the tools over the weekend to make sure it did really work as expected.

Monday came and we finally had some usable solutions. The course would start late in the morning with a welcome address: we invested the few hours left burning the needed copies of the master DVD, cloning the USB key (did we mention it is very easy to clone QEMU-Puppy?) and printing some minimal introduction instructions. All in time for the students to collect these last-minute materials just after the welcome talk: a short introduction to the concept of virtual machines and the systems we had built before lunch and we were ready to dive in depth in the complexities of Workflows, Asynchronous Web Services and Grid computing.

The students were able to run all the tutorials (some of them at slow speeds) successfully, working with our servers from real Linux machines, fully configured with all the professional tools needed, just as the ones we use ourselves in our everyday work (note that we use a wider choice of Linux distributions besides Ubuntu or Puppy, but the tools were all there), and what is even best of all, the students were able to take these very Linux machines back home with them, with all their work saved and all the tools and services needed so they could continue working home.

Fit the Eighth: The vanishing

"In the midst of the word he was trying to say,
In the midst of his laughter and glee,
He had softly and suddenly vanished away—
For the Snark was a Boojum, you see."

Sure, some sessions had to run a bit slower and hence we had less time for practices, but we had saved the day (the week actually), students were happy and they brought back home a fully functional system with them to continue working. From our point of view, specially after all the troubles, this had been a great success.

In our case it was a sad combination of circumstances that led us to endure this much harshness, but we now have a new weapon in our arsenal, and a powerful one: as many others we often have to teach courses out of our institutions, in places where students will only have access to a Windows machine, often with no choice for booting, and indeed we have to move around (travelling or just home/work). Formerly we would resort to remote servers to demonstrate the practical tutorials, but from now on we have a new choice: we can create non-expensive virtual machines, configure them easily to add all the tools the users will need and provide them each with their own preconfigured server, one that they can bring back home and continue working, and better yet, one that they don't need to boot into, one they can use simultaneously with their operating system of choice while working normally.

From our point of view we have jumped a quantum leap: we are reaching a point were students can really become autonomous and continue using their newly acquired professional skills independently after the courses, and where we can concentrate on just solving doubts instead of fixing problems. What is even more interesting, using QEMU-Puppy-Linux we are now looking to a new horizon on Bioinformatics: as USB keys grow in size and squeeze in cost and MacOS X moves to Intel, we plan to add major user components (like EMBOSS) to the base system so that future users will be freed from the tyranny of dual booting and have a complete professional environment that integrates seamlessly into their existing system adding a plethora of new tools and opening for them the gates of freedom.

If you, like us, have ever found yourself in a similar situation, please do consider having a look at these technologies (VMWare, QEMU, Live distributions, Puppy and QEMU-puppy) and joining our effort [15] to bring them into Bioinformatics, as you will probably find them worth the effort and may also jump to the next evolutive step in the ladder to freedom, training and education.

References

1. The Hunting of the Snark. An Agony in Eight Fits, by Lewis Carroll
2. EMBRACE (http://www.embracegrid.info)
3. 3D-EM NoE (http://www.3dem-noe.org/)
4. EGEE (http://www.eu-egee.org)
5. EMBnet (http://www.embnet.org)
6. UAM (http://www.uam.es)
7. Knoppix (http://www.knoppix.org)
8. VMWare (http://www.vmware.com)
9. QEMU (http://www.qemu.org)
10. Ubuntu (http://www.ubuntu.com)
11. Puppy Linux (http://puppylinux.org)
12. QEMU-Puppy (http://www.erikveen.dds.nl/qemupuppy/)
13. SSH Filesystem (http://fuse.sourceforge.net/sshfs.html)
14. FUSE (http://fuse.sourceforge.net)
15. SBG, Structural Biology on the Grid team (http://bioportal.cnb.uam.es/sbg/)

Linux for bioinformatics:

dedicated distributions for processing of biological data – Part 2: Repositories and Complete Systems

Antonia Rana and Fabrizio Foscarini[1]

European Commission, Joint Research Centre, Institute for Health and Consumer Protection (IHCP), Via E. Fermi 1 - 21020 Ispra (VA) - Italy

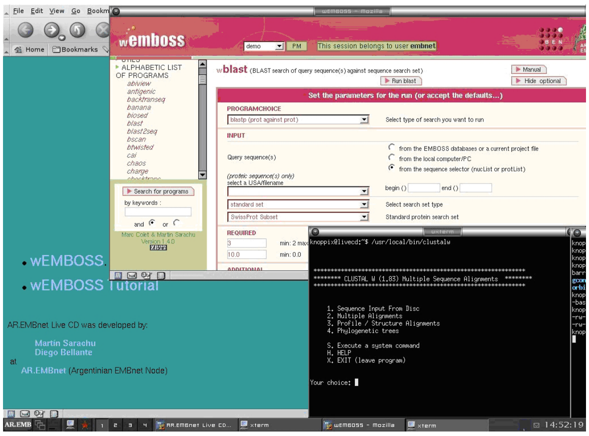
antonia.rana@ec.europa.eu

Introduction

In this second part of our overview on Linux distributions for bioinformatics, we will complete the review of live distributions, add some information on package repositories, i.e. collections of packages ready to be installed on the appropriate Linux distribution without the need to go through the "configure-make-install" process, and finally conclude with some considerations on distributions for installation with particular attention to those that can be used to set up a computer cluster for bioinformatics.

AR.EMBNET live CD 0.1

The first live distribution that we cover in this part is AR.EMBNET Live CD.



wblast (BLAST search of query sequence(s) against sequence search set)

Set the parameters for the run (or accept the defaults...)

PROGRAMCHOICE: blastp (prot against prot) — Select type of search you want to run

INPUT

Query sequence(s): ○ from the EMBOSS databases or a current project file ○ from the local computer/PC ● from the sequence selector (nucList or protList)

(proteic sequence(s) only) select a USA/filename — begin () end ()

standard set — Select search set type

SwissProt Subset — Standard protein search set

REQUIRED

3 — min: 2 max

10.0 — min: 0.0

AR.EMBnet Live CD was developed by:

Martín Sarachu
Diego Bellante
at
AR.EMBnet (Argentinian EMBnet Node)

```
knoppix@livecd:~$ /usr/local/bin/clustalw


 **************************************************************
 ******** CLUSTAL W (1.83) Multiple Sequence Alignments ********
 **************************************************************


     1. Sequence Input From Disc
     2. Multiple Alignments
     3. Profile / Structure Alignments
     4. Phylogenetic trees

     S. Execute a system command
     H. HELP
     X. EXIT (leave program)


Your choice:
```

Figure 1. AR.EMBNET live CD.

Version 0.1 of AR.EMBNET Live CD is distributed on the Argentina EMBnet website. It has been released on 11/10/2005. Like most of the live distributions reviewed in part 1 of this article, it is based on KNOPPIX, version 3.6. Although this is quite an old version for KNOPPIX (currently at version 5.1), the hardware auto-recognition and configuration procedure performed well, it is able to recognise FireWire and USB but not wireless network card. The system starts a graphical desktop environment with higher resolution using IceWM. This is the only window manager available differently from the others live distributions which use KDE and the rich set of system utilities provided with this desktop environment.

When the system starts an empty desktop is presented and the Mozilla browser is opened on the file /KNOPPIX/resources/index.html which displays a web page with information on the authors of the live CD, the link to the wEMBOSS login and a tutorial on how to use wEMBOSS. The full EMBOSS suite is available through the powerful wEMBOSS interface together with FASTA, BLAST, etc. which are also available via the wEMBOSS interface through the use of wrappers4EMBOSS. The version of the bioinformatics software installed is not the most up-to-date. Although the set of bioinformatics programs is not as rich as in the distros reviewed Part 1, it certainly helps that the available ones are all accessible through a web interface, which is simple, clear and well designed.

This system can be easily used by biologists, on the other hand making any modifications to the system on the fly might be difficult since the root account is locked.

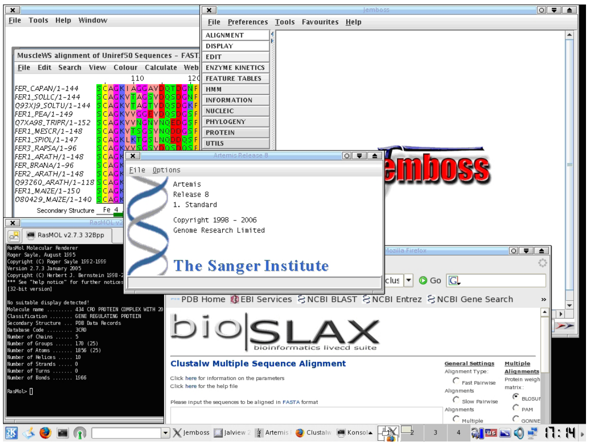
To facilitate saving the work done, the **Utilities** menu contains useful links to create boot flop-

[1] Disclaimer required under the terms and conditions of use of the Internet and electronic mail from Commission equipment:
"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."

pies, to install new software, the **Config** menu contains links to configure a printer or create a persistent KNOPPIX directory and save KNOPPIX configuration.

Home page: `http://www.ar.embnet.org/livecd.html`
Current version: 0.1
Base system: KNOPPIX 3.6
Media: LiveCD
Kernel version: 2.4.27
Wireless: not recognised

BioSLAX


File Tools Help Window
MuscleWS alignment of Uniref50 Sequences - FAST
File Edit Search View Colour Calculate Web
File Preferences Tools Favourites Help
ALIGNMENT
DISPLAY
EDIT
ENZYME KINETICS
FEATURE TABLES
HMM
INFORMATION
NUCLEIC
PHYLOGENY
PROTEIN
UTILS
emboss
Secondary Structure
RasMOL v2.7.3 32Bpp
Artemis Release 8
File Options
The Sanger Institute
PDB Home EBI Services NCBI BLAST NCBI Entrez NCBI Gene Search
bio SLAX bioinformatics livecd suite
Jemboss Jalview Artemis Clustalw Konsole

```
Artemis
Release 8
1. Standard

Copyright 1998 - 2006
Genome Research Limited
```

```
RasMol Molecular Renderer
Roger Sayle, August 1995
Copyright (C) Roger Sayle 1992-1999
Version 2.7.3 January 2005
Copyright (C) Herbert J. Bernstein 1998-2
*** See "help notice" for further notices
[32-bit version]

No suitable display detected!
Molecule name ......... 434 CRO PROTEIN COMPLEX WITH 20
Classification ........ GENE REGULATING PROTEIN
Secondary Structure ... PDB Data Records
Database Code ......... 3CRO
Number of Chains ...... 5
Number of Groups ...... 170 (25)
Number of Atoms ....... 1856 (25)
Number of Helices ..... 10
Number of Strands ..... 0
Number of Turns ....... 0
Number of Bonds ....... 1966

RasMol> 
```

Clustalw Multiple Sequence Alignment

Click here for information on the parameters
Click here for the help file

Please input the sequences to be aligned in FASTA format

Figure 2. BioSLAX.

BioSLAX has been released by the resource team of the BioInformatics Center (BIC), National University of Singapore (NUS). The current release is 5.1.8.1, dated April 2006. Unlike most of the distributions reviewed in this paper, it is not based on the popular KNOPPIX distribution but on Slax, a live distribution based on Linux Slackware. It provides a graphical desktop environment with higher resolution with KDE as window manager. An interesting feature of this distribution is the availability of additional modules with additional functionality (some wireless cards) and system utilities (such as NTFS-Fuse) which can be downloaded from the home page and added to the ISO image before burning the CD.

When the system starts the graphical window manager is automatically started and all bioinformatics applications can be started from the BioSLAX menu which is added to the main KDE menu. This menu groups the available tools in "**Console Application**" (Bioperl, BioGrep, BLAST, etc.), "**Desktop Apps**" (Artemis, ClustalX, JAlign, etcc) and "**Web Apps**" (Web BLAST, Web ClustalW, Web Phylip, etc.), making it easy to locate a chosen program.

This distribution contain also wEMBOSS, however there seems to be problems with this installation: some modifications to the configurations are necessary in order to use the application properly.

The documentation for this distribution is rather poor, very little is available on the web site and no tutorial is immediately available on the desktop of the system when it starts. However, this seems to be quite a new distribution and it will likely improve in its next releases.

Home page: `www.bioslax.com`
Current version: 5.1.8.1
Base system: Slax 5.1.8
Kernel version: kernel 2.6.16
Media: LiveCD
Wireless: device is not detected.

Package repositories

Package repositories, which distribute bioinformatics applications packaged for easy installation on the most commonly available Linux distributions, are an interesting alternative for those who have experience with Linux and possibly Linux systems already in place. Most of these focus on the Fedora Core distribution and use the rpm package management system. This provides an alternative to the KNOPPIX-Debian-derived distributions such as those examined in part 1 (EMBnet.news Volume 13 Nr. 2).

BioLinux

BioLinux is a repository of RPM packages which provides a number of widely used bio-informatics tools that can be installed on popular Linux distributions, such as Fedora Core, Red Hat and SuSE. Packages can be downloaded from the RPM repository which requires user registration. Available software include: rpms for EMBOSS, NCBI Tools, ClustalX/W, Phylip, Bioperl and others.

According to the information provided on the web site *"BioLinux does not aim to become Yet Another Linux Distribution, but to provide add-on RPM packages for existing Linux distributions"*.

The website has recently been updated using wiki technology in order to provide more interactivity between users and developers. It is possible to contribute to the repository by uploading packages in rpm format via anonymous ftp to ftp://rpm.biolinux.org.

Home page: `www.biolinux.org`
Last Update: 28 February 2007
Supported distributions: Red Hat 9, SuSE 9.1, Fedora Core 1-6.

BioRPMs

BioRPMs is a collection of software used in bioinformatics, targeted at the same type of Linux distributions as BioLinux, i.e. Fedora Core and Red Hat. The repository is hosted at the Bioinformatics and Expression Analysis facility located at the Department of biosciences and nutrition at Novum, Sweden. These packages are managed through a so called meta-installer, a tool that facilitates the installation of rpm packages using the capabilities provided by the Debian based APT package manager. Although these are RPM packages, their authors recommend the use of this meta-installer (AptRpm) for a clean and trouble-free installation. The motivations that lead its authors to create this repository are related to the difficulty in installing a tool from source code and to the frequent availability of a tool in form of a package but for older versions. By creating a repository, they strive to help the scientific user by avoiding them the hunt for "recent and decent packages" by *"visiting a fair number of diverse sites (thank you, Google) and resorting to the "scientific" method of trial-by-error"*.

The repository contains about one hundred packages with indication of the licence under which they are distributed and include some genome databases (e.g. acedb) and many packages from the R and BioConductor suites. However, the repository does not seem to be updated very frequently, the most recent package being dated Sept 2005.

Home page: `apt.bea.ki.se`
Last Update: Jul 21, 2004
Supported distributions: Red Hat 8-9, Fedora Core 1-2

Debian-Med

Debian-Med is a "Custom-Debian-Distribution", or, in other words, a Debian internal project which aims to cover the needs of special groups of users. In particular, this project is targeted at the medical community providing packages of medical and biological software. It is important to notice that this is not a separate distribution from the classic Debian GNU/Linux. It is made of two meta-packages: med-bio and med-bio-devel, which provide a collection of software tools for medical practice and patient management, medical research, hospital information systems, medical imaging, molecular biology and medical genetics. The bioinformatics tools are only a small fraction of the whole Debian-med collection.

These meta packages are available with the full Debian distribution and can be selected for installation similarly as for any other Debian package.

A more detailed description of Debian-med and an overview of the benefits of Debian GNU/Linux over the other Linux distributions is provided in [11].

The Debian community has not yet addressed the problem of incorporating biological databases into Debian, however it seems not likely that this will happen in the main Debian distribution.

Home page: `http://www.debian.org/devel/debian-med/index.html`
Last Update: 12/08/2007
Supported distributions: Debian

Debian-bioinformatics

Debian-bioinformatics is a different initiative from Debian-med. According to its authors at the University of South-Wales in Australia, Debian-bioinformatics is *"An attempt to build up an archive of easily installable biology/molecular genet-*

ics/bioinformatics-related software for Debian". However, attempts to download files from this repository were not successful.

Home page: `debian.bioinformatics.unsw.edu.au`
Last Update: unknown
Supported distributions: Debian

Bio-nix

Bio-nix is a project listed on bioinformatics.org as *"a bold attempt to foster the use of linux among biologists working in wet-labs and in silico alike"*. Bio-nix uses yet another meta-installer called yum to manage the package installation of the most commonly available bioinformatics tools including: EMBOSS, Bioperl, BioRuby, BioJava, BLAST, ClustalX, Glimmer, GROMACS, HMMER and also SciLab, Octave and GNUplot.

However, the page which describes the packages contains links to the packages home page and the content of the "Yum-Repository" is still empty. Discussions related to this initiative are being held within a yahoo group, but it is not clear whether anything has been released yet.

Home page: `http://bioinformatics.org/BIO-NIX/mediawiki/index.php/Main _ Page`
Last Update: 2/7/2007
Supported distributions: Fedora Core 5

Biopackages.net

Biopackages is interesting among package repositories because it provides rpm packages for Apple Darwin in addition to the more common Fedora Core and CentOS Linux distributions. Like Bio-nix, it uses the yum meta-installer. Available packages include:

Bioperl, BioConductor, the Generic Genome Browser, Chado, Turnkey, GMODweb, Textpresso, BLAT, EMBOSS, HMMER, R, BioConductor, although not all of them are available for all distributions.

The repository consists of two branches: the stable and the testing branch. Packages are well documented and revisions are frequent. The repository is also well organized and it is certainly a bonus the fact that CentOS and Darwin are supported in addition to Fedora Core.

Home page: `biopackages.net`
Last Update: 25/04/2007
Supported distributions: Fedora Core 2, Fedora Core 5, CentOS 4

Complete systems

While the live CDs might be useful to become familiar with bioinformatics software and for training purposes, a full distribution complete with the necessary bioinformatics software which installs out of the box might be more useful for a production environment. We have found a few complete distributions which have been made available in the open source community that can be installed and work as the main (only) operating system. All of them where started around 2004. Unfortunately, from a recent review of their websites, they do not seem to be supported or even available any longer.

BioLand

BioLand is a Linux distribution based on Fedora Core 2 created by The Center for Bioinformatics of the Peking University. Downloading of the four CDs that constitute the BioLand distribution was not successful, therefore, it was not possible to hands-on test this distribution. According to its homepage (which was last updated in 2004), this distribution "is designed to provide an easy to use and configure computing environment for biologist" and includes ClustalW/ClustalX, various databases (e.g prosite, the Restriction Enzyme Database, etc), EMBOSS, Jemboss, HMMER, NCBI-Tools, njplot, phylip, pftools, etc. and uses wEMBOSS as the graphical interface to access most of these programs.

Home page: `bioland.cbi.pku.edu.cn` [2]
Last Update: 2004
Base distribution: Fedora Core 2

[2] At the time of reviewing the second part of this article, the website does not seem to be active any longer.

	NCBI Blast	Bioperl	ClustalX/ ClustalW	EMBOSS	Glimmer	HMMER	Phylip	primer3	T-Coffee	Gromacs
Bio-Linux	2.2.13	1.4	1.82-3	3.0.0	2.13	2.3.2	3.65	1.0.0	3.27	
DNAlinux	2.2.12		1.83	3.0.0		2.3.2		1.0.0	1.37	
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			3.573c			
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	I	I	1.83	4.0.0			I		4.9.3	
Biorpms			1.83	3.0.0		2.3.2		1.0.0		
Biopackages										
Biolinux[1]			1.83	2.9.0			3.61			
Rocks + Bio roll	2.2.12		1.83	3.0.0	3.02	2.3.2	3.65	1.0.0	3.84	3.3.1
BioSLAX	2.2.13	I	1.83	3.0.0	I	2.3.2	I	1.0	3.93	
AR.EMBNET	2.2.9		1.83	2.10.0						
Package Current Version*	2.2.16 Apr 2007	1.5.2 Dec 2006	1.83 2003	4.1.0 Mar 2007	3.0.2 May 2006	2.3.2	3.66	1.1.1 Mar 2006	5.05	3.3 2006

Table 1. Versions of the most popular bioinformatics programs as instlled in the reviewed distributions compared with the latest version available.
I - installed according to the information on the web site.)
* - Packages indicated are for Fedora Core 4 (directories for FC 5 and 6 are available but still empty)

BioLinuxBR

BioLinuxBR is listed in tucows (`www.tucows.com`, a well-known shareware and freeware hosting service) as Parallel BioLinux-BR 1.1 and BioLinux-BR 2.1 for download, submitted in Feb 2006. However all links appear to be broken (timeout). This distro is also mentioned in [8] and [10] as an interesting and promising Linux distribution for bioinformatics.

According to [12] "*BioLinux-BR is a project directed to the scientific community. The intention is to create a Linux distribution for people with little familiarity with the installation of the operational system and mainly for people that do not know how they must proceed to unpack a program, compile and install it correctly. For these reasons, this is a Linux system that aims to be easy to use and still offering packages that will be part of the BioLinux-BR.*"

No hand on testing was possible for this distribution as well because the download server does not seem to be alive.

Home page: `Biolinux.df.ibilce.unesp.br` [3]
Last Update: unknown
Base distribution: unknown

BioBrew

BioBrew [13] was officially launched by its author Glen Otero and Callident, a team of Linux professionals dedicated to high performance computing on Linux clusters, at ClusterWorld, June 24th-26th, 2003. It is an open source Linux distribution that installs a cluster for bioinformatics. It is based on the NPACI ROCKS cluster environment which automates installation of cluster components, includes all the management software a cluster requires, and contains a number of the most popular bioinformatics tools such as: the NCBI toolkit, BLAST, mpiBLAST, HMMER, ClustalW, GROMACS, WISE, and EMBOSS, among the others.

The latest version of the complete BioBrew distribution available as a single DVD including NPACI Rocks was 3.1.0, released in 2005. Subsequent releases of BioBrew consisted only of a CD containing the bioinformatics software to be added to the NPACI Rocks distribution (BioBrew roll). These releases can be installed after installing NPACI Rocks and following the instructions provided by this distribution. The latest BioBrew release is 4.1.2 which was released in 2006 in the form of a "roll" (additional CD), to be added to a standard NPACI Rocks installation.

Home page: `biobrew.bioinformatics.org`
Last Update: 2006
Base distribution: NPACI Rocks

Linux-based bioinformatics clusters

BioBrew and the related NPACI Rocks (see *NPACI Rocks with BioRoll* section) are particularly interesting among the install distributions reviewed so

[3] At the time of reviewing the second part of this article, the website does not seem to be active any longer.

far as they can be used to set up a computer cluster for bioinformatics.

Various attempts have been made at implementing parallel versions of popular bioinformatics programs which are aimed at taking advantage of extra processing power of computer clusters. mpiBlast is one of the best known of these programs.

Parallel implementation of the most popular bioinformatics programs including BLAST [7] [14] [15] [16] [17] [18] and EMBOSS [4], that splits sequences into smaller ones (with an overlap to allow examination of the area at the split), and distributes these sub-sequences to a number of nodes in a cluster have been described.

These attempts, aimed at achieving greater speeds and improving the performance of algorithms being run against rapidly growing databases, have been fostered by the availability at low costs of high performance computing solutions that were unthinkable of some years ago.

In fact, before Beowulf clusters, the available hardware-based acceleration method was symmetric multiprocessing (SMP), in which the computations are distributed on several CPUs connected tightly in the same physical computer, and sharing the same memory. However SMP servers are expensive, on the other hand Beowulf clusters can be assembled from individual commodity PCs and workstations connected by a fast network such as Myrinet, Fast Ethernet, and Gigabit Ethernet. Each node in the cluster has its own memory and local hard disk space. Beowulf clusters typically consist of a master (or "login" or "front-end") node that distributes the bioinformatic application among the other nodes (compute nodes). To the users, the cluster is a single logical entity that masquerades as a single computer, their jobs are accepted, scheduled, queued, and prioritized by a software-based distributed resource management (DRM) layer which coordinates the availability of resources needed by the jobs, monitor the jobs as they run, and clean up after job complete. The most commonly seen DRM software suites within the life sciences run Sun Grid Engine (SGE) (`gridengine.sunsource.net`) or the Portable Batch System (PBS).

Improvement of the performance of bioinformatics applications through parallelization on a cluster is obtained by modifying the code to split the problem across the nodes and collect the results. The most popular software libraries that can be used to assist in these modifications are the Message Passing Interface (MPI) and Parallel Virtual Machine (PVM) (an example of a modified algorithm which uses these libraries is mpiBLAST [6]). Different types of parallelisms are possible with bioinformatics applications [14][22] on clusters. Job Parallelism, in which each job is run on a different processor using a batch scheduler, is the easiest way to achieve parallelism in a multi-user environment. However, job parallelism is limited as it does not benefit single jobs from the availability of multiple processors. Database Parallelism is achieved by separating the search space in several independent segments, having different nodes search simultaneously and merging the results on the front-end node.

Examples of Beowulf style clusters used in bioinformatics can be found in [4] [5] [6] [7] [14] [15] [18] [19] [21] [22] [25].

An alternative solution to Beowulf style clusters is openMosix [20], a kernel extension which allows multiple uniprocessors and symmetric multiprocessors (SMP) nodes running the same Linux kernel to work in close cooperation. OpenMosix creates a cluster environment that simulates a large SMP machine so that applications can run unmodified.

It has no central control or master/slave relationship between nodes: each node works as an autonomous system which is capable of migrating processes to peer nodes when the requirements for resources exceed some threshold levels. Users can run parallel applications by starting multiple processes in one node and when more resources are required, the local node can migrate some processes to other nodes to take advantage of available remote resources.

The configuration design of this type of clusters is dynamic, nodes may join or leave the network with minimal disruptions, process dispatching is managed directly by the Linux kernel, transparent to users and does not involve parallelization of the application (bioinformatics tools). This type of cluster is supported by the Quantian distribution reviewed in Part 1 of this article.

BioBrew

BioBrew is the first full Linux distributions that we became aware of, which was aimed at building a cluster for bioinformatics. The project is listed in bioinformatics.org and is reported to have been installed with a high number of nodes and in many places. However, as already indicated in the *BioBrew* section, this distribution is no longer released as a unique DVD containing both the OS and required tools to manage the cluster and the bioinformatics tools, but only as a roll to be added to the usual NPACI Rocks installation.

In addition the roll version for the current release of NPACI Rocks (4.3) has not been released yet. This, together with the release of a bio-roll directly from NPACI seems to indicate that this distribution might no longer be supported in the future.

NPACI Rocks with BioRoll

While a BioBrew release which supports the new version of Rocks is not available yet, NPACI Rocks, an open source, Linux-based software stack for building and maintaining Beowulf high-performance computing (HPC) clusters distributes, together with the main installation DVD, a "bio-roll" containing the most popular bioinformatics tools.

NPACI Rocks, which is available for download under the open-source Berkeley Software Distribution (BSD) and GNU General Public License (GPL) licenses at `www.rocksclusters.org`, is built on standard and mostly open source components, to make computer clusters easy to deploy, operate and maintain.

While based on RedHat in its first releases, the current version of Rocks (4.3) is based on the CentOS 4 update 5 and adds to the base operating system a collection of software components that can be used to, maintain, monitor and operate the cluster: the MySQL DBMS to build and maintain a database of cluster-wide information, MPICH2 libraries (or PVM) to provide the tools to build parallel programs to run on the cluster, the Ganglia Cluster Toolkit (`ganglia.sourceforge.net`), an open source, realtime monitoring tool which is also able to create a remote execution environment for users, to monitor the status of the cluster, a DRM (SGE is the current default) to manage the cluster resources all integrated, packaged and with detailed documentation.

Installing Rocks is as easy as installing a single computer OS, provided the cluster components (front-end and nodes) comply with the hardware requirements. The nodes are installed from the front-end using the Preboot Execution Environment (PXE), which allows nodes to obtain the OS via a network boot. Installation of the bioinformatics tools can be done at the same time with the installation of the OS or in a second time. The packages included in the bio roll are currently: HMMER: v2.3.2, NCBI BLAST v2.2.12, mpiBLAST v1.4.0-May 2006, Biopython v1.41, ClustalW v1.83, MrBayes v3.1.2, T-Coffee v3.84, EMBOSS v4.0.0, Phylip v3.65, FASTA v3.4, Glimmer v3.02, Gromacs v3.3.1, Perl-Bioperl v1.5.1. All these applications can be run on the cluster using SGE commands to dispatch jobs on the various nodes. Details on how to use them on the cluster via the SGE scheduler are provided in the BioRoll user guide. Unfortunately, these tools, with the exception of BLAST and T _ Coffe can be run only on the command line logging in on the front-end.

OSCAR

OSCAR (Open Source Cluster Application Resources) was developed by the Open Cluster Group to provide users with the best practices for installing, programming and maintaining HPC (High Performance Computing) clusters by providing all the necessary software to create a Linux cluster in one package. OSCAR targets mid-size clusters (50+ node clusters). Unlike NPACI rocks, it does not come bundled with a Linux distribution, but it provides a framework for the easy installation of a cluster on a Linux distribution of choice, provided it is among those supported (currently: Red Hat Enterprise Linux 4, Fedora Core 4 and 5, Mandriva Linux 2006, SuSE Linux 10.0). OSCAR helps automate the installation, maintenance and even the use of cluster software. A graphical user interface provides a step-by-step installation guide and also functions as a graphical maintenance tool.

The main components of OSCAR are:

- System Installation Suite (SIS), Cluster Command and Control (C3) and OPIUM (user management) for cluster administration;
- MPICH, LAM/MPI and PVM; OpenPBS/MAUI, Torque and SGE; Ganglia and Clumon; for jobs scheduling, resource management, cluster monitoring and parallel programming libraries;
- OSCAR Database (ODA) and OSCAR Package Downloader (OPD) for Core infrastructure/management.

A description of the use of OSCAR for bioinformatics can be found in [27].

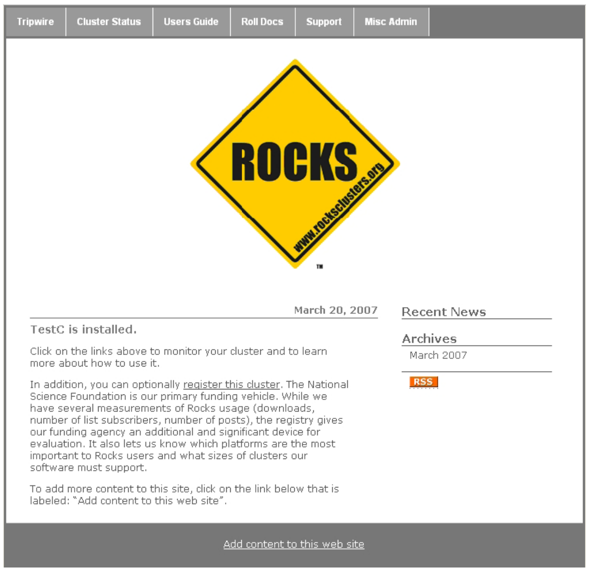
Setting up a Linux cluster for bioinformatics

We have used some of the open source solutions mentioned in this article to set up a test Linux cluster for bioinformatics.

The operating system software used to set up the basic environment is NPACI (National Partnership for Advanced Computational Infrastructure) Rocks 4.2 available at `www.rocksclusters.org/`. This software is distributed with a set of interesting features, collected in what they call roll, including "Bioinformatics utilities". The Bioinformatics Roll is a collection of some of the most common bioinformatics tools that are being used by the bioinformatics community today, which includes:

- NCBI BLAST
- mpiBLAST
- EMBOSS
- FASTA
- ClustalW
- Etc.

Two of these tools, BLAST and T _ Coffee, are distributed with a web interface (WWW BLAST and T _ Coffee). These are simple interfaces which allow user to access remotely the programs with all their options but only on the front-end. The aim of the web interface is to mimic the software in a web browser, but no features are available to organize and manage projects and related files.



March 20, 2007

TestC is installed.

Click on the links above to monitor your cluster and to learn more about how to use it.

In addition, you can optionally register this cluster. The National Science Foundation is our primary funding vehicle. While we have several measurements of Rocks usage (downloads, number of list subscribers, number of posts), the registry gives our funding agency an additional and significant device for evaluation. It also lets us know which platforms are the most important to Rocks users and what sizes of clusters our software must support.

To add more content to this site, click on the link below that is labeled: "Add content to this web site".

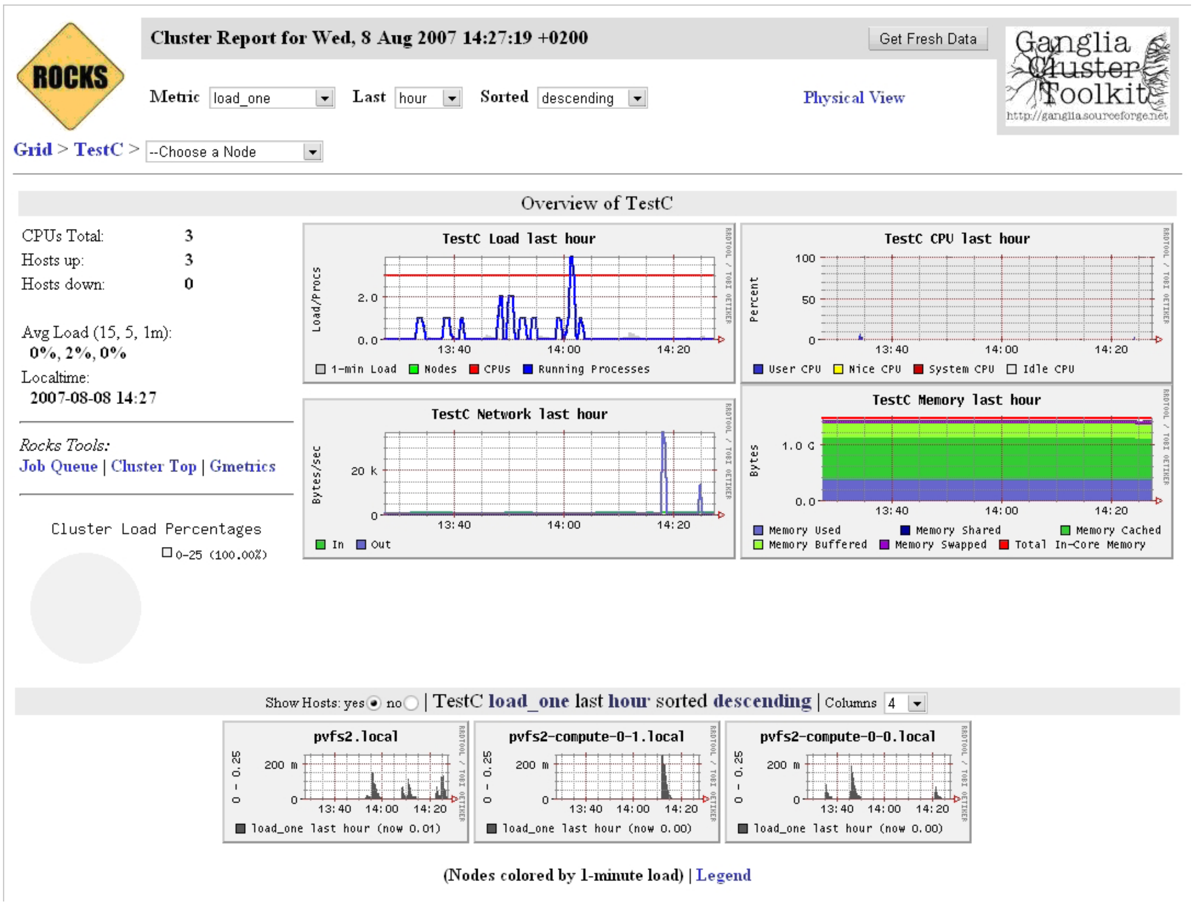
Add content to this web site

Figure 3. the home page of our test NPACI Rocks cluster.

These tools are developed to be used on a single machine, optionally with multiple CPU, for these reason even the web interfaces doesn't allow the software to exploit the full power of the cluster.

The installation of the cluster is extremely simple and well documented. This distribution allows the creation of a cluster from two to thousand nodes with little effort and in a short amount of time. The most expensive thing is the hardware, but for a test one can use spare old PCs, that meets the minimum requirements for the installation (20GB HDD, 1GB RAM and 1 Ethernet for node and 20GB HDD, 1GB RAM and 2 Ethernet for front-end). The ISO image of the full distribution can be downloaded from `www.rocksclusters.org`. The ISO image must be burned on a DVD which is used to boot and setup the front-end and eventually the backend if the nodes don't support the network boot (PXE).

When the installation is complete the cluster is ready to be used with bio-informatics tools from a CLI (command line interface). Rocks uses SGE (SUN Grid Engine) as batch system to submit jobs to the cluster. In order to use the cluster to its full potential it is necessary to submit tasks to SGE as batch programs. The Rocks Bio-Roll user guide contain some examples on how to create a batch program including, for instance, parallel execution of BLAST searches. More examples on more sophisticated ways to use SGE are available at `http://www.sun.com/software/gridware/`.



Cluster Report for Wed, 8 Aug 2007 14:27:19 +0200

Metric load_one Last hour Sorted descending

Grid > TestC > --Choose a Node

Overview of TestC

CPUs Total: 3
Hosts up: 3
Hosts down: 0

Avg Load (15, 5, 1m):
0%, 2%, 0%
Localtime:
2007-08-08 14:27

Rocks Tools:
Job Queue | Cluster Top | Gmetrics

Cluster Load Percentages
□ 0-25 (100.00%)

Show Hosts: yes ● no ○ | TestC load_one last hour sorted descending | Columns 4

(Nodes colored by 1-minute load) | Legend

ROCKS
Get Fresh Data
Physical View
Ganglia Cluster Toolkit
http://ganglia.sourceforge.net
TestC Load last hour
Load/Procs
2.0
0.0
13:40
14:00
14:20
□ 1-min Load 🟩 Nodes 🟥 CPUs 🟦 Running Processes
TestC CPU last hour
Percent
100
50
0
🟦 User CPU 🟨 Nice CPU 🟥 System CPU □ Idle CPU
TestC Network last hour
Bytes/sec
20 k
0
🟩 In 🟦 Out
TestC Memory last hour
Bytes
1.0 G
0.0
🟦 Memory Used 🟦 Memory Shared 🟩 Memory Cached
🟩 Memory Buffered 🟪 Memory Swapped 🟥 Total In-Core Memory
RRDTOOL / TOBI OETIKER
pvfs2.local
pvfs2-compute-0-1.local
pvfs2-compute-0-0.local
0 - 0.25
200 m
■ load_one last hour (now 0.01)
■ load_one last hour (now 0.00)
■ load_one last hour (now 0.00)

Figure 4. A sample screen of the current status of our test NPACI Rocks cluster using Ganglia.

Jobs submitted from users are queued on the cluster and their executions order depends on different factors, some of these factors are:

- System load
- Resource availability
- Job priority

To distribute information between nodes there are different ways, the two most common are:

- using a shared file systems to allow all the node to access the same repository;
- using the cluster management commands to copy the desired resource in every node of the system.

During set up, Rocks configures the shared file system for the user home directory and the directory containing the bioinformatics tools.

The cluster distribution enforces by default a number of security measures including the use of a firewall (`iptables`) to protect the cluster through its front-end from the network and a file integrity scanner (tripwire) to check that no critical file is maliciously modified.

In addition, during installation, a web server is installed which provide remote access to the Rocks documentation, the status of the cluster and grants access to the tripwire reports.

Since the bioinformatics tools included in the BioRoll can only be used via the command line, we decided to install and test wEMBOSS (`www.wemboss.org/`) in order to add a web interface to EMBOSS and other bioinformatics tools installed with BioRoll

At the moment of writing this article the latest stable release of wEMBOSS is 1.7.1. This release is distributed with wrappers4EMBOSS 1.5.1, a package that integrates a number of popular bioinformatics software to the EMBOSS installation. BLAST, FASTA and muscle are included among others.

The installation of the wEMBOSS package is simple. This software is based on C and perl programs that are personalized and compiled during the installation procedure. The only pre requisite is the availability of the perl module "Mail::Mailer" (`http://search.cpan.org/CPAN/authors/id/M/MA/MARKOV/MailTools-1.77.tar.gz`), ps2pdf (`http://www.cs.wisc.edu/~ghost/`) and of course a C compiler and perl interpreter.

The installation is performed using a perl script that asks some questions and uses the answers to personalize the C and perl programs to the actual system characteristics. A point to be made on the installation procedure is that the script doesn't modify the configuration file of the web server and required this task to be performed manually. However the procedure to modify the Apache HTTPD web server configuration file is provided in the INSTALLATION file available with the distribution.

The installation instructions suggest to modify the /etc/shadow file as useful step to facilitate the user management in the software. However from a security point of view to modify the security measures on the /etc/shadow file is not an advisable thing to do.

The instructions provide, in effect, for a workaround to this. However, the requirement that users in the wEMBOSS system are listed as Unix users and not simply users of the web service remains. Separating the management of the wEMBOSS users for the users of the operating system would make the system more flexible and more secure.

Once the installation procedure is completed wEMBOSS is available from the cluster front-end using any web browser.

Using wrappers4EMBOSS it is possible to add extra programs, as BLAST, Clustal, etc. to the wEMBOSS interface. The procedure to install a new wrapper requires that the source code of EMBOSS is available on the computer. The installation procedure copies the desired files inside the source distribution of EMBOSS and then recompiles and reinstalls everything. This is done for every new wrapper to be added.

For every program available through wEMBOSS a simple interface and a full options interface is available, allowing even a beginner to start using these tools immediately.

wEMBOSS organizes work into projects. Users can create as many projects and subprojects as they want; they can also rename and move projects to different locations. This allows work to be logically organized facilitating the review of all project related data.

Although wEMBOSS + wrappers4EMBOSS is a powerful solution to use remotely and via web interface a number of the most popular bioinformatics tools, it doesn't allow to exploit the full potential of a cluster without interfacing a job scheduler that can dispatch and thus parallelize programs executions to the various nodes of the cluster.

The possibility to use a job scheduler on a cluster and still access bioinformatics tools via a web interface is offered by PISE (Pasteur Institute Software Environment) [3]. In fact PISE is a tool to generate Web interfaces for Molecular Biology programs that is a commonly used component in web based cluster solution for bioinformatics ([24] and [21]

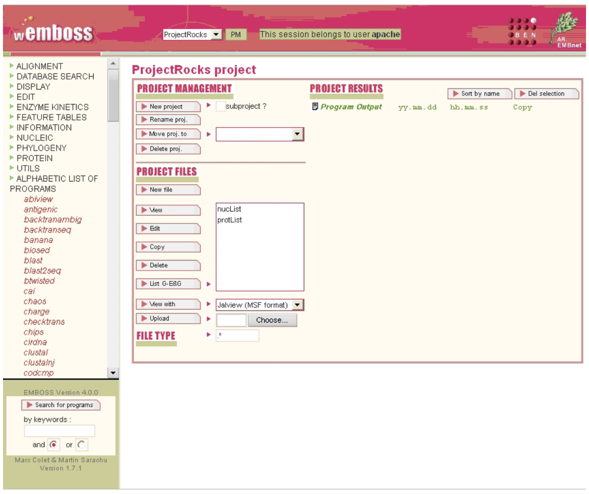


Figure 5. the wEMBOSS entry point for the installation of our cluster.

At the moment of writing this article the latest stable release is the 5.a.

This software is based on perl programs that are personalized and executed during the installation procedure. This program needs some packages to be available on the system and these are: some perl modules (CGI.pm, BSD::Resource, MIME::Lite etc.), rxp program (`http://www.cogsci.ed.ac.uk/~richard/rxp.html`), seqio (`ftp://ftp.pasteur.fr/pub/GenSoft/unix/programming/seqio-1.2.2.tar.gz`) and a perl interpreter. These requests are documented in the PISE web page.

After all required software is installed, only few files need to be personalized to meet the installed system configuration as reported in the documentation. After that, the commands to execute to install everything are "`make applis`", "`make install _ applis`" and to add the whole EMBOSS suite "`make emboss`" and "`make install _ emboss`". Now hundreds of bio-tools are available trough a web interface.

PISE supports by default PBS (Portable Batch System) as the batch system to distribute jobs on the cluster node, while Rocks install SGE by default. So in order to use PISE on our test cluster we had two options:

1. either to add PBS to our cluster or
2. to modify PISE to use SGE

Both tasks are very simple to perform and are briefly explained here.

In case 1 we need to download the PBS roll for the installed version of Rock from `http://www.rocksclusters.org/wordpress/?p=43`, add the roll to the front-end and then distribute the roll to the nodes. This step requires that all the nodes must be reinstalled but this is a very trivial and fast task to perform (it only needs the execution of one command). The last step is to modify the user environment under which the web server that provide the PISE interface runs to indicate that it must use PBS i.e.: add "`PATH=/opt/torque/bin/:$PATH; export PATH`" at the end of the file `.bash _ profile`.

In case 2 it is sufficient to change few lines in the following files: **pbs.pl**, **wrapper.pl** and **make-cgi.pl.in** i.e.:

- In pbs.pl substitute the line with the qsub command with:

  ```
  my $cmd = "qsub -cwd -m n -N
  $jobname";
  ```

- In wrapper.pl adjust the first line with the correct path to the perl interpreter:

  ```
  #! /usr/bin/perl
  ```

- In make-cgi.pl.in locate the section related to the PBS instructions and add just before the line that execute the batch command (`print SCRIPT \"open (BSUB, \\\"|\$batch _ command\\\");\\n\";`) the following lines:

  ```
  print SCRIPT \"     my \\\$sge_
  root=\\\"/opt/gridengine\\\";\";
  print SCRIPT \"     \\\$ENV{SGE_ROOT}
  =  \\\$sge_root;\\n\";
  print SCRIPT \"     my \\\$sge_
  qmaster_port=\\\"536\\\";\";
  print SCRIPT \"     \\\$ENV{SGE_
  QMASTER_PORT} =  \\\$sge_qmaster_
  port;\\n\";
  ```

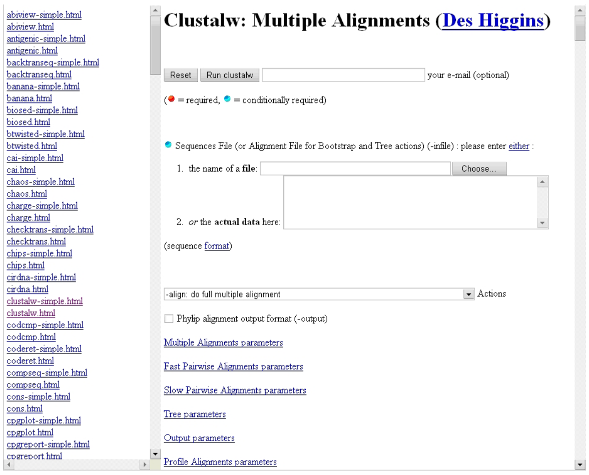
and then follow the usual steps. The resulting patch files are:

```
Index: lib/pbs.pl
=======================================
--- lib/pbs.pl      (revision 1)
+++ lib/pbs.pl      (working copy)
@@ -7,7 +7,7 @@
   my ($command, $queue, $jobname) =
@_;

   ## Command init
-  my $cmd = "qsub -m n -N $jobname";
+  my $cmd = "qsub -cwd -m n -N
$jobname";
   ## Set stdout & stderr output files
   $cmd .= " -o $command.out -e
$command.err";
   ## Set queue
Index: lib/wrapper.pl
=======================================
--- lib/wrapper.pl  (revision 1)
+++ lib/wrapper.pl  (working copy)
@@ -1,4 +1,4 @@
-#! /local/bin/perl
```

```
+#! /usr/bin/perl

 use Fcntl ,:flock';

Index: Maker/make-cgi.pl.in
=======================================
--- Maker/make-cgi.pl.in   (revision 1)
+++ Maker/make-cgi.pl.in   (working
copy)
@@ -1302,6 +1302,13 @@

 print SCRIPT \"open (STDOUT, \\\">
\\\$SCRATCH_DIR/.qsub\\\");\\n\";

+\#\#\#\#\#\# SGE - START
+print SCRIPT \"     my \\\$sge_
root=\\\"/opt/gridengine\\\";\";
+print SCRIPT \"     \\\$ENV{SGE_ROOT}
=  \\\$sge_root;\\n\";
+print SCRIPT \"     my \\\$sge_
qmaster_port=\\\"536\\\";\";
+print SCRIPT \"     \\\$ENV{SGE_
QMASTER_PORT} =  \\\$sge_qmaster_
port;\\n\";
+\#\#\#\#\#\# SGE - END
+
 print SCRIPT \"open (BSUB, \\\"|
\$batch_command\\\");\\n\";
```

Figure 6. SGE patch file.

In our tests we used option 2 and verified that PISE work well with SGE.



abiview-simple.html
abiview.html
antigenic-simple.html
antigenic.html
backtranseq-simple.html
backtranseq.html
banana-simple.html
banana.html
biosed-simple.html
biosed.html
btwisted-simple.html
btwisted.html
cai-simple.html
cai.html
chaos-simple.html
chaos.html
charge-simple.html
charge.html
checktrans-simple.html
checktrans.html
chips-simple.html
chips.html
cirdna-simple.html
cirdna.html
clustalw-simple.html
clustalw.html
codcmp-simple.html
codcmp.html
coderet-simple.html
coderet.html
compseq-simple.html
compseq.html
cons-simple.html
cons.html
cpgplot-simple.html
cpgplot.html
cpgreport-simple.html
cpgreport.html

Clustalw: Multiple Alignments (Des Higgins)

Reset | Run clustalw | your e-mail (optional)

(● = required, ● = conditionally required)

● Sequences File (or Alignment File for Bootstrap and Tree actions) (-infile) : please enter either :

1. the name of a **file**: Choose...
2. *or* the **actual data** here:

(sequence format)

-align: do full multiple alignment Actions

□ Phylip alignment output format (-output)

Multiple Alignments parameters

Fast Pairwise Alignments parameters

Slow Pairwise Alignments parameters

Tree parameters

Output parameters

Profile Alignments parameters

Figure 7. An example of the PISE web interface generated on our cluster. PISE automatically queues the processes on SGE in a transparent way for users.

An interesting solution which can overcome the issues web access highlighted with in our test is found in [21]. Developed taking into account the more demanding requirements of a grid environment, the Bioportal can be usefully employed to support a local cluster with the necessary management tools, bioinformatics software and user centered web based access. BioPortal is based on Rocks for clustering, on the Open Grid Computing Environment (OGCE, `www.collab-ogce.org`), a toolkit of reusable portal components that can be combined to form a common portal container and on the notion of a "portlet," a portal component that allows management of user configurable blocks in the user authenticated component of the portal. It has as its main components the PISE web interface generator for molecular biology applications, cluster management middleware from the Globus toolkit (www.globus.org), common bioinformatics software (the current deployment supports roughly 140 bioinformatics applications, including EMBOSS, GLIMMER, HMMER, NCBI tools, PHYLIP, ClustalW and FASTA), about 300G of databases that are regularly updated, and workflow extensions based on Taverna which allows users to compose the available applications as a workflow.

The Bioportal provides user based access and security at multiple levels. User communication with the portal server can also be encrypted with the Secure Socket Layer (SSL). The Bioportal software is available for download to no profit organizations at `http://www.ncbioportal.org/downloads/downloads.php`.

Conclusions

We have presented in this article and in its part 1 a number of Linux distributions tailored for bioinformatics and molecular biologists users. Live distributions, which facilitate learning and training without requiring installations, were covered in the first part, packages repositories which provide ready-to-install versions of the most popular bioinformatics tools for the most popular Linux distribution and thus are useful in environments where Linux systems are already installed and complete systems comprising the OS and the bioinformatics tools in this part. Finally we covered some Linux based solutions for bioinformatics cluster providing an example using a slight modification of commonly used open source components.

Bibliography

[1] Luethy, R, Hoove, C.: Hardware and software systems for accelerating common bioinformatics sequence analysis algorithms Drug Discovery Today. BIOSILICO,Volume 2, Issue 1,1 January 2004,Pages 12-17

[2] Rieffel, M. A., Gill, G. T., White, W. R. Bioinformatics clusters in action, www.paracel.com/pdfs/clusters-in-action.pdf

[3] Letondal, C. PISE: A Web interface generator for molecular biology programs in Unix, Bioinformatics Vol. 17 no. 1 2001, 73-82

[4] Podesta, K., Crane, M., Ruskin, H. J.: A Sequence-Focused Parallelisation of EMBOSS on a Cluster of Workstations. Computational Science and Its Applications – ICCSA 2004, 473-480, Lecture Notes in Computer Science, Volume 3045/2004

[5] Chiou-Nan Chen, Kuan-Ching Li, Chuan Yi Tang, Yaw-Lin Lin, Hsiao-Hsi Wang, Tsung-Ying Wu, On Design and Implementation of a Bioinformatics Portal in Cluster and Grid Environments, http://vecpar.fe.up.pt/2006/programme/papers/44.pdf

[6] Darling, A.E., Carey, L., Feng, W.: The Design, Implementation, and Evaluation of mpiBLAST. ClusterWorld Conference & Expo and the 4th International Conference on Linux Clusters: The HPC Revolution (2003)

[7] Hokamp, K., Shields, D.C., Wolfe, K.H., Ca _ rey, D.R.: Wrapping up BLAST and other applications for use on Unix clusters. Bioinformatics, Vol. 19 (2003) 441-442

[8] Field D, Tiwari B, Booth T, Houten S, Swan D, Bertrand N, Thurston M: Open Software for Biologists: from famine to feast. Nature Biotechnology 2006, 24:801-803

[9] Tiwari, B and Field, D. The Bioinformatics Playground. Linux User and Developer. 2005. Issue 46. pp. 50-56

[10] Gearing Up for Bioinformatics, Canadian Bioinformatics Helpdesk Newsletter, March 3, 2005, http://gchelpdesk.ualberta.ca/news/03mar05/cbhd _ news _ 03mar05.php

[11] Tille, A., Moller, S. Free software in biology using Debian-Med: A resource for information Agents and Computational Grids, http://people.debian.org/~tille/debian-med/talks/200507 _ biomed/debian-med _ handout.pdf

[12] “[Fsf-friends] GNU/Linus distros for life sciences/bioinformatics researchers”, http://gnu.org.in/pipermail/fsf-friends/2005-January/002743.html

[13] Otero, G., Linux, Optimized for Science, The Scientist 2004, 18(1):33

[14] Braun, R.C., Pedretti, K.T., Casavant, T.L., Scheetz, T.E., Birkett, C.L., Roberts, C.A.: Parallelization of local BLAST service on workstation clusters. Future Generation Computer Systems, Vol. 17 (2001) 745-754

[15] Hokamp, K., Shields, D.C., Wolfe, K.H., Ca _ rey, D.R.: Wrapping up BLAST and other applications for use on Unix clusters. Bioinformatics, Vol. 19 (2003) 441-442

[16] Grant, J.D., Dunbrack, R.L., Mahon, F.J., Ochs, M.F.: BeoBLAST: distributed BLAST and PSI-BLAST on a Beowulf cluster. Bioinformatics, Vol. 18, No. 5 (2002) 765-766

[17] Mathog, D.R.: Parallel BLAST on split databases. Bioinformatics, Vol. 19, No. 14 (2003) 1865-1866

[18] Darling, A.E., Carey, L., Feng, W.: The Design, Implementation, and Evaluation of mpiBLAST. ClusterWorld Conference & Expo and the 4th International Conference on Linux Clusters: The HPC Revolution (2003)

[19] Dagdigian, C.: Building and managing production bioclusters Drug Discovery Today. BIOSILICO, Volume 2, Issue 5, September 2004, Pages 208-213

[20] Bar, M.: OPEN MOSIX. http://openmosix.sourceforge.net/linux-kongress _ 2003 _ openMosix.pdf

[21] Lavanya R, Mark SCR, Jeffrey LT, Daniel AR.: Grid Portals for Bioinformatics. Second International Workshop on Grid Computing Environments, 2006

[22] Rieffel, M. A., Gill, T. G., White, W. R.: Bioinformatics Clusters in Action. http://www.paracel.com/pdfs/clusters-in-action.pdf

[23] Gupta, R., Fang, Y., Hussain, M.: Streamlining Beowulf Cluster Deployment with NPACI Rocks. www.dell.com/downloads/global/power/ps1q05-20040176-Gupta.pdf

[24] The BioTeam: iNquiry Administrator’s Guide. http://bioteam.net/inquiry/InquiryAdmin-Guide.pdf

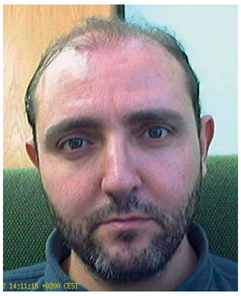
[25] Stocker, G., Rieder, D. Trajanoski, Z.: ClusterControl: a web interface for distributing and monitoring bioinformatics applications on a Linux cluster. Bioinformatics, Vol. 20 no. 5 2004, pages 805–807

[26] Papadopoulos, P. M., Katz, M. J., Bruno, G.: NPACI Rocks: Tools and Techniques for Easily Deploying Manageable Linux Clusters. Concurrency Comput. Pract. Exper. 2002;00:1-20

[27] Li, B. OSCAR and Bioinformatics, 2004, http://www.linuxjournal.com/article/7462

Keeping your data up to date

Part II: Indexing and formatting



José R. Valverde

EMBnet/CNB, Centro
Nacional de Biotecnología,
CSIC Campus Univ.
Autónoma Cantoblanco,
Madrid 28049, Spain

Most often, it is of little interest to just keep a copy of the databases as distributed (normally compressed), as this is usually not the best way to work with them. Normally we will need to pre-process the data before using it.

Maintaining processed data and indexes may become a daunting task, so labor intensive that it might easily become practically unmanageable: as a matter of fact, building all EMBL indexes may take longer than one day which is unpractical if we want to keep the database updated everyday by downloading the data during the night and making it available in the morning.

Reducing the work load

There are many tricks we have used. One of these tricks relies on the use of 'make' to compare the dates of creation of the files we have downloaded with the dates of our indexes, so that only those files that have changed are processed, saving thus the work of needless indexing of unchanged data.

This is an approach that gives rather good results. But it is certainly subject to enhancements. An additional step might involve the use of tricks to save space: for example, EMBL is a huge database; if we download it compressed for use with blast, we need to create first a working version in uncompressed format, and then reprocess this one to create the blast database. Working this way we will need a huge amount of space only for temporary use as we are only interested in the final results and not in the intermediate files.

Reducing space

We my reduce our space needs by using pipes. The trick consists in making a special 'pipe' or 'fifo' file and dump over it the data with a background process. This process will not actually write the data on the file, instead it will stop and wait until there is someone ready to read the data from the other end of the pipe. When a second process now tries to read from the pipe (e.g. to uncompress or format the data), it will start reading and the first process will wake up and begin writing as the reader reclaims data. This way we'll have saved disk space without any sacrifice in efficiency. Note that we may chain as many processes as needed, for instance:

```
1. mknod database.gz p
2. wget  -N  -nh  ftp://server.example.net/
   pub/database.gz &
3. mknod database p
4. zcat database.gz > database &
5. mknod formatted-database p
6. convert-to-fasta database > formatted-
   database &
7. make-blast-db formatted-database
```

In this example, lines 1, 3 and 5 create the auxiliary pipes. Line 2 downloads the compressed database from the network and writes it on one end of the first pipe. We run this process in the background (it will wait for another process to read its data) so we may continue issuing commands. In line 4 we read the compressed data from the other end of the first pipe as it is being downloaded and uncompresses the data over a second pipe. Line 6 runs a program to read the uncompressed database in its original format and convert it to fasta format writing the data on a last pipe. Finally, on line 7 we read the resulting fasta-formatted database and convert it to its definitive blast format.

In practice it would be as if we had typed:

```
wget ftp://server.example.net/pub/
database.gz -o log -O - |\
zcat | convert-to-fasta | make-blast-db
```

Such a command line would allow us to update a blast database from the Net without taking any additional disk space at all. Of course, we all know that one may download the standard databases pre-formatted as blast from the Net, but this helps us illustrate our point, solves our problem when we are interested in a non-standard blast dataset, and saves us downloading a blast-formatted database.

Some times it will be possible to issue such a command line. Other times we will need to resort to the multi-step process described initially. It will all depend on whether the programs we are using allow us to concatenate their input and output easily or not.

Avoiding stepping on each other

The other problem we have is -as we said- the run time for the formatting or indexing processes. Some times we can't avoid this time becoming too big. For example: we may reduce EMBL update time by indexing only nightly updates, but every so a new release will come out of the EBI factory and we'll need to index it anyway.

If our indexing/formatting process is run every day and the execution time exceeds at any time its periodicity (as it often happens whenever a new EMBL release is rolled out) we risk that a new processing job is started before the previous one has finished. This new job will see the indexes out-of-date and delete the "old" files (the ones the previous job is still trying to build) and start over again. The old job will continue till it finishes, but its job will have been lost, overwritten by the new one. Next day comes, and the process repeats itself as the last job hasn't finished yet and a new one is launched.

We need some way to prevent this from happening. A simple way is to avoid launching any new processing job over a database while there is a previous one running. Only if the previous job has finished will we allow a new update to start. To achieve this we need some element that will act as a 'mark' or 'signal' to know when an update is already running. We might use 'ps' and look for the update program, but this approach is highly error-prone (what if we use the same program for a different task at the same time?). A better solution is to use a file that indicates the presence of an update job.

Using a lock file works in the following way: before we start an update we check if the file exists. If it does, then it means that there is already an update under way and running a new one will only wreak havoc, and so, if it exists, we just give up. If the lock file does not exist, then there is no other update running and we may go ahead: our first step must be to flag our presence creating the lock file, so we preclude other updates running before we are done. Once we finish, we'll simply remove the lock file:

```
if [ -e running ] ; then
     echo "An update is already running"
else
     touch running
     do_update
     rm running
fi
```

A real life example

We provide as an example a 'Makefile' that we have been using to automatically update database indexes. This Makefile doesn't stretch the tricks described to the extreme as we were interested -for a variety of reasons- in keeping copies of the databases in a number of formats (in spite of the disk space consumed). You may actually be in a similar situation if you use a wide number of software packages each requiring one of these formats or if you have to provide network access to data in various formats to other people.

You can download this example from the following URL:

http://www.es.embnet.org/~jr/embnet/scripts/make _ update.tgz

In a simpler environment (e. g. one that is only interested in EMBOSS) it would be possible to extensively apply the tricks described for additional optimizations. Indeed, you may find that some times it makes more sense to substitute some of the productions used by a call to an auxiliary script that isolates the processing and allows for greater control and versatility without cluttering the Makefile.

This is a very long Makefile, and for you convenience we are making it available over the WWW. The Makefile contains comments that explain the tricks used. It displays examples of the methods described: avoiding overstepping on ourselves, indexing only changed files and use of pipes to save temporary space.

Finally, please note this is only an (outdated) example and if you want to use it you would need to check it thoroughly and maintain it as locations, databases and software needs tend to change from time to time.

Following we'll include some excerpts to illustrate the concepts presented. Let us start with the beginning: we *start by defining some variables to customize our setup*, mainly source and destination data directories:

```
#
# Makefile to automatically re-build
# databases and indexes.
#
# (C) Jose R. Valverde, May-2002
#     EMBnet/CNB
#
# $Id: Makefile,v 1.12 2005/04/11
# 14:33:59 genadmin Exp $
#
###### C O N F I G U R A T I O N ######
# Where are we located
MY_DIR=/u/sysadmin/genadmin/make_update
#
# First let us define some commodity
# general directories
# This is not strictly needed but
# simplifies definition of subsequent
# variables.
#
ORIG_DB_DIR=/data/ftp/pub/databases
DEST_DB_DIR=/data/gen
EMBOSS_INDEX_DIR=/data/gen/emboss
FASTA_DB_DIR=/data/gen/fastadb
BLAST_DB_DIR=/data/gen/blastdb
#-------------------------------------
--
# The location of database files:
#       - compressed distribution files
#       - expanded flat files
#       - emboss index files
#       - main database release
#--------------------------------------
#
# EMBL
#
ORIG_EMBL_DIR=$(ORIG_DB_DIR)/embl/
release
DEST_EMBL_DIR=$(DEST_DB_DIR)/embl
INDEX_EMBL_DIR=$(EMBOSS_INDEX_DIR)/embl
```

The main dependency rule is a trivial dummy one that builds all release databases (everything) and all new data since last release (new) if needed:

```
#=====================================
# MAIN DEPENDENCY RULE
#=====================================
all: everything new
        echo "Done."
```

Let us now see how do we *get a master, uncompressed, flat-file database*. At EMBnet/CNB we must keep an early access EMBL release in compressed format for others to download as well. Thus, we'll use this compressed version as the master copy: we depend on every single file and when invoked expand only those files that have changed. More complex dependency rules can be defined (see, e. g., the rules for REBASE in the Makefile). Just in case, we guard against running twice using a lock file.

```
# EMBL: FLAT DATABASE
# Rule for EMBL: uncompress only
#   changed files and ONLY if
#   no other update process is
#   already running.
#
#   COMMENT: first check if a lock file
#   exists flagging the existence of an
#   already running update. If none
#   exists then create one and go on.
#   For each changed/updated database
#   section ($?) we'll uncompress it
#   (zcat).
#   Of course, once done, remove the
#   lock.
#
embl: $(ORIG_EMBL_DIR)/*.dat.gz
  @if [ ! -e embl.lck ] ; then          \
     ( echo "Updating EMBL";            \
     touch embl.lck ;                   \
     cd $(DEST_EMBL_DIR) ;              \
     for i in $? ; do                   \
       ( echo "processing $$i" ;        \
       zcat $$i > `basename $$i gz` )   \
     done ;                             \
     cd - ;                             \
     mv embl.lck embl ;                 \
     echo "Updated EMBL" ) ;            \
  else                                  \
    echo "An EMBL update is already
running" ;                              \
  fi
```

This example fails if the number of dependencies is exceedingly long, as it may happen when processing PDB (currently more than 40 thousand files): Make detects all dependencies and passes it along to the shell, which can not handle such a long list of arguments and aborts. This needs *special handling with* `xargs` or `find`.

```
pdb_new:
  @if [ ! -e pdb.lck ] ; then                     \
     ( echo “Updating PDB” ;                      \
    touch pdb.lck ;                               \
    cd $(DEST_PDB_DIR) ;                          \
    find $(ORIG_PDB_DIR) -name ‘*.ent.
Z’ -follow                                        \
      -cnewer $(MY_DIR)/pdb -exec                 \
      $(MY_DIR)/uncomp.sh {} \; ;                 \
    cd - ;                                        \
    mv pdb.lck pdb ;                              \
    echo “Updated PDB” ) ;                        \
  else                                            \
    echo “A PDB update is already
running” ;                                        \
  fi
```

Building EMBOSS indexes now is a simple extension of the above examples. In our case we have hard-coded some localities (a script to initialize EMBOSS and the path to it), but it should be trivial to adapt to your needs. The major trick here is finding out what the current release number is. This will need to be tailored for each specific database (see the complete Makefile for other examples).

```
#---------------------------------------
# Dependency rules for EMBOSS
#---------------------------------------
#
# Index EMBL
#   Only if an update or indexing
#   process is not already running
#   Release number is computed from
#   “Release_*” ancillary file in EMBL
#   release distribution directory.
#
# COMMENT: After checking for a lock
# file, we start by initializing
# EMBOSS (/opt/molbio/emboss/SetUp.sh).
# Then go to the directory, extract the
# release number for later use and run
# ‘dbiflat’ to build the indexes in the
# target directory.
#
embl_index: embl $(DEST_EMBL_DIR)/*.dat
  if [ ! -e embl.lck ] && [ ! -e embl_
index.lck ] ; then  \
   ( echo “Indexing EMBL” ;               \
   touch embl_index.lck ;                 \
   . /opt/molbio/emboss/SetUp.sh ;        \
   cd $(INDEX_EMBL_DIR) ;                 \
   release=`basename $(ORIG_EMBL_DIR)/
Release_* | cut -d _ -f2` ;               \
   dbiflat -idformat EMBL -dbname embl\
     -directory $(DEST_EMBL_DIR)          \
     -filenames ‘*.dat’                   \
     -release $$release -date `date
“+%D”`                                    \
     -indexoutdir $(INDEX_EMBL_DIR)       \
   cd - ;                                 \
   mv embl_index.lck embl_index ;         \
   echo “Indexed EMBL” ) ;                \
  else                                    \
   echo “EMBL indexing is already
running” ;                                \
        fi
```

Finally, we include an example rule to build BLAST databases using the pipe trick to save space. Again you can do much better than this as shown above, but since we need to maintain the compressed and uncompressed source data anyway, in our case we can do with a simpler rule.

```
#---------------------------------------
# Dependency rules for BLAST
#---------------------------------------
#
# EMBL is generated from the FASTA
# files
#
#   COMMENT: We use a clever trick here
#   to save space: instead of copying
#   all files into one huge source
#   database flat file to be processed
#   later, we create a pipe (with
#   ‘mknod’) and send the files to
#   it in the background. Then we start
#   ‘formatdb’ reading from the pipe
#   and so no additional temporary and
#   huge disk space is consumed.
#
embl_blast: embl_fasta
  @if [ ! -e embl_blast.lck ] ; then  \
    ( echo “Making EMBL blast files”; \
    touch embl_blast.lck ;            \
    cd $(BLAST_DB_DIR) ;              \
    if [ ! -p embl ] ;                \
      then mknod embl p ;             \
    fi ;                              \
    cat $(FASTA_DB_DIR)/*.fasta > embl
& ;                                    \
    /opt/molbio/bin/formatdb -t embl -i
embl                                   \
      -p F -o T -l embl.log ;          \
    cd - ;                             \
    mv embl_blast.lck embl_blast ;     \
    echo “EMBL blast files updated”);  \
  else                                 \
    echo “EMBL blast files are already
being updated” ;                       \
  fi
```

Finally we'll demonstrate the summary dependency rules invoked originally by 'all': as you can see the sample Makefile provides support a rather comprehensive list of databases and formats. We simply create two rules: one for building the base releases (everything) and the other for building updates (new). We can now use the invocations 'make new' or 'make all' in a cron job everyday after updating the databases, and the commands 'make all' or 'make everything' each time we get a new major release of the databases. Of course we can also invoke database-specific rules as needed when we update a single specific database.

```
#----------- -------------------------
# Summary Dependency rules
#-------------------------------------
EMBL = embl embl_index embl_new embl_
new_index
SWISSPROT = swissprot swissprot_index
swiss_new swiss_new_index
TREMBL = trembl trembl_index
OTHER = prints unigene domo rebase
pdbfinder omim uniprot
FASTA = embl_fasta swiss_fasta trembl_
fasta pdb_fasta nr_fasta
FASTA_NEW = emblnew_fasta swissnew_
fasta
BLAST = embl_blast swiss_blast trembl_
blast
new: embl_new_index swiss_new_index
$(FASTA_NEW) pdb_new
  @ echo “Update finished” `date “+%D”`
everything: $(EMBL) $(SWISSPROT)
$(TREMBL) $(FASTA) $(BLAST) $(OTHER)
  @ echo “Done” `date “+%D”`
```

When this is not enough

Sometimes this is not enough. For example, to use a relational database and since we have no way to know which new cross-references may appear in new entries, we may need to rebuild **all** the cross-indexes among **all** the databases every time.

This is, for instance, what happens with MRS. The problem is that with the current database sizes simply building the cross-indexes may exceed greatly the computation capacity of most computers. MRS allows you to build several indexes in parallel, and so if we had a big multiprocessor we might perhaps build them in time, but this increases considerably the cost of building indexes locally.

The solution in these cases is to talk with your colleagues and look for more people with the same problem and willing to share the work load to achieve your goals using many machines in a distributed approach. In this scenario each node will take responsibility for building a manageable subset of indexes and then exchange the with the others. This approach allows you as well to distribute searches to increase lookup speed in huge databases.

The problem, however, of distributing indexing and lookup processes over a distributed infrastructure is not trivial. In the case of MRS this approach is being developed by a group of EMBnet nodes on a project known as the MRS Federation (which, by the way, you can join). There are other similar projects for other database systems on distributed or Grid environments under way, being this an active research area nowadays.

Red velvet

Vivienne Baillie Gerritsen

Autumn has come. So have the hunters. And stags have finished fashioning their antlers in their quest to seduce a partner and fight off rivals. Besides copulation, antlers are one of Nature's many wonders. Not only are they beautiful and sculptural but they are a rare example of an organ which regenerates, rapidly and on a yearly basis. Consequently, it is hardly surprising that scientists are spending a lot of time trying to unravel the underlying mechanisms which participate in the growth of an antler. Annexin II is just one of the proteins involved in antler regeneration, and more specifically in cartilage mineralization.

Red Stag, Robert Fuller

http://www.robertefuller.com/

Antlers are not an uncommon sight these days. If you are lucky enough to live on the outskirts of a forest, there is a great chance that you will spy an antler or two, usually at dusk. Antlers are made out of bone. They grow from pedicles that form at puberty and which, in time, become permanent protuberances from which antlers bud and are cast seasonally. They can grow at the amazing rate of two centimetres a day and represent the only example of both irrigated and innervated cartilage in the animal kingdom.

Once antlers are cast, the next generation initiates immediately thanks to resident stem cells on the permanent pedicles. These cells differentiate first into chondroblasts and then into chondrocytes, which are associated with the formation of cartilage. Mineral deposition then occurs on the cartilage scaffolding, and bone is formed. This stage gradually gets rid of any blood supply which is made to the antlers. As a consequence, they stop growing and their metamorphosis to bone is completed. The final touch comes with the shedding of a thin film of velvet skin that coats the appendages, and the antlers are then ready for the rutting season.

Naturally, the regeneration of an organ demands the existence of a complex network of proteins. Annexin II is just one of these proteins but an essential one, since it seems to be directly involved in bone formation. More specifically, annexin II seems to be involved in the formation of calcium channels and mineralisation in the environment of chondrocytes and osteoblasts. Annexin II is also involved in many other mechanisms and it is now becoming obvious that it is a multifunctional protein. It assembles into a heterotetramer and belongs to the very large annexin family which is characterised by the existence, in their sequence, of an annexin domain – a 70 amino acid domain – of which each type of annexin has a defined number. This particular domain binds calcium – a feature characteristic to all annexins.

Annexins are distributed both in the animal and the plant kingdom; from humans to guinea pigs, frogs, flies, zebra fish, worms, moulds and mouse-ear cress. They are found in many different tissues and participate in a variety of physiological processes. All are calcium-dependent and bind to phospholipids, and are usually found in the periphery of the plasma membrane, either intracellular or extracellular. One form even seems to be soluble. This was a surprising discovery because annexins do not have a signal peptide, so they must follow a route other than the customary endoplasmic reticulum secretory pathway prior to secretion.

Besides antler formation, annexin II is involved in a host of other processes such as fibrinolysis, cell proliferation and differentiation as in bone formation, endo- and exocytosis, cell migration, cell shape and even immunity. As a consequence, it is expressed in many different tissue types such as the central nervous system, the cardiovascular system, bone marrow and the small intestine. In fact, annexin II's tissue and function versatility is at the origin of an equally versatile nomenclature – as is frequently the case. And, with the years, it has been given a variety of names such as p39, calpactin I heavy chain, protein I, chromobindin 8 and lipocortin II…

When a protein is multifunctional in this way, there is a good chance that it is involved in just as many diseases. So far, annexin II is known to be over-expressed in patients suffering from acute promyelocytic leukemia, which results in excessive fibrinolysis. On the brighter side of things, annexin II's involvement in fibrinolysis could be promising for the design of anti-coagulant drugs for those suffering from cardiovascular complications. Furthermore, though it is not clear why, annexin II also seems to have a suppressive effect on tumour malignancy – which is an excellent reason to get to know it better.

Cross-references to Swiss-Prot

Annexin A2, *Cervus elaphus* (Red deer) : Q2Q1M6

References

1. Molnar A., Gyurjan I., Korpos E., Borsy A., Steger V., Buzas Z., Kiss I., Zomborszky Z., Papp P., Deak F., Orosz L.
   Identification of differentially expressed genes in the developing antler of red deer Cervus elaphus
   Mol. Genet. Genomics 277:237-248(2007)
   PMID: 11108960

2. Rand J.H.
   The annexinopathies: a new category of diseases
   Biochim. Biophys. Acta 1498:169-173(2000)
   PMID: 17131158

National Nodes

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

Australia
RMC Gunn Building B19, University of Sydney, Sydney

Austria
Vienna Bio Center, University of Vienna, Vienna

Belgium
BEN ULB Campus Plaine CP 257, Brussels

Brazil
Embrapa Informatica Agropecuaria, UNICAMP-CP, CampinasT

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

China
Centre of Bioinformatics, Peking University, Beijing

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

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

Cuba
Centro de Ingeniería Genética y Biotecnolgía, La Habana

Finland
CSC, Espoo

France
ReNaBi, French bioinformatics platforms network

Greece
Biomedical Research Foundation of the Academy of Athens, Athens

Hungary
Agricultural Biotechnology Center, Godollo

India
Centre for DNA Fingerprinting and Diagnostics (CDFD), Hyderabad

Israel
Weizmann Institute of Science, Department of Biological Services, Rehovot

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

Mexico
Nodo Nacional EMBnet, Centro de Investigación sobre Fijación de Nitrógeno, Cuernavaca, Morelos

The Netherlands
Dept. of Genome Informatics, Wageningen UR

Norway
The Norwegian EMBnet Node, The Biotechnology Centre of Oslo

Pakistan
COMSATS Institute of Information Technology, Chak Shahzaad, Islamabad

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

Portugal
Instituto Gulbenkian de Ciencia, Unidade 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

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

LION Bioscience
LION Bioscience AG, Heidelberg, Germany

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 print then please let us know. Please send your contributions to one of the editors. You may also submit material by e-mail.

Past issues of EMBnet.news are available as PostScript or PDF files. You can get them from the EMBnet organization Web site:
http://www.embnet.org/download/embnetnews

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

Submission deadline for the next issue:

November 20, 2007

EMBnet.news is an official publication of the EMBnet organisation
www.embnet.org