

EMBnet.news

**MacOSX and
Bioinformatics (part I)
Focus on UMBER
Courses in Colombia
DNA array technology
Protein Spotlight**

**Volume 9 Nr. 1
April 2003**

Editorial

New design, new front cover, new articles, new editorial board, it seems that nothing was left over from the previous EMBnet.news!

Although we made a lot of changes, we tried to stick to the spirit of the journal. Since its beginnings in July 1994, our journal targets the EMBnet user community providing information about the nodes and descriptions of useful bioinformatics tools. In this issue and in the future, we will provide articles about EMBnet nodes (e.g. Focus on UMBER), technical tutorials (e.g. MacOSX and Bioinformatics), and more scientific articles (e.g. DNA Microarray technology). We are open to publish articles from our user community and encourage our readers to submit their manuscripts. Of course we will favor articles dealing with topics in bioinformatics or to the use of bioinformatics for successful science.

To broaden our scope and starting with this issue of EMBnet.news, we are happy to announce a new collaboration between EMBnet.news and the SWISS-PROT team publishing the "Protein Spotlight" review. This monthly review selects proteins of interest to the general public. From anthrax toxins to prion proteins and ricin, from antifreeze glycoprotein to defensins and cocaine esterase, all of these proteins are "famous" in the media, but not really well described by our newspapers. The goal of "Protein Spotlight" is to popularize science by providing people with short and accurate information about proteins. We hope that you will enjoy this valuable addition to our journal.

The editorial board:
Erik Bongcam-Rudloff
Laurent Falquet
Pedro Fernandes
Gonçalo Guimaraes Pereira

Cover picture: DNA sculpture Bror Marklund 1973, Linnaeus Centre, Uppsala, [© Erik Bongcam-Rudloff]

Contents

Editorial	2
Focus on UMBER	3
Courses in Colombia	8
DNA Array technology	10
Announcements	12
MacOSX and Bioinformatics (Part I)	13
Protein spotlight 32	17
Protein spotlight 33	19
National nodes	21
Specialist nodes	23

Publisher:

EMBnet Administration Office
c/o Jack Leunissen
Dept. of Genome Informatics
Wageningen UR, Dreijenlaan 3
6703 HA Wageningen
The Netherlands
Email: Jack.Leunissen@wur.nl
Tel: +31 317 484074

Editorial Board:

Erik Bongcam-Rudloff, "The Linnaeus Centre for Bioinformatics" LCB, SLU/UU, SE
Email: erik.bongcam@bmc.uu.se
Tel: +46-(0)18-4716696
Fax: +46-(0)18-4714525

Laurent Falquet, SIB, Lausanne, CH
Email: laurent.falquet@isb-sib.ch
Tel: +4121 6925954
Fax: +4121 6925945

Pedro Fernandes, Instituto Gulbenkian, PT
Email: pfern@igc.gulbenkian.pt
Tel: +315 214407912
Fax: +315 214407970

Gonçalo Guimaraes Pereira, UNICAMP, BR
Email: goncalo@unicamp.br
Tel: +55-19-37886237/6238
Fax: +55-19-37886235

Focus on UMBER University of Manchester Bioinformatics Education & Research



T.K. Attwood

School of Biological Sciences
& Department of Computer Science
University of Manchester
Oxford Road
Manchester M13 9PT, UK

Introduction

In March 1999, the Specialist Node at UCL transferred to the University of Manchester. Although that was only 5 minutes ago, incredibly, 4 years have passed in the meantime! Much has happened during this time, so what follows is an update on UMBER's bioinformatics research and education activities.

Our work has centred for some years around the provision of the PRINTS protein fingerprint database [1] and associated tools for searching and visualising protein sequence information: e.g., the Expert Analysis SYstem, EASY, which combines protein sequence and family database searches to provide a consensus diagnosis of the identity of a protein in the context of the family to which it belongs [2]; FingerPRINTScan, the custom-built tool for searches against fingerprints in the current version of PRINTS [3]; BLAST PRINTS, for searches of sequences matched in the current release [4]; and the Colour INTERactive Editor for Multiple Alignments, CINEMA [5]). Another major theme of the group has been the development of on-line materials for education, notably BioActivity, our interactive bioinformatics practical [6]. Some of these activities have been supported either directly from EMBnet's coffers, or through collaboration with other EMBnet Nodes. In the sections that follow, we outline our progress in these and related areas.

PRINTS and its automatic supplement, prePRINTS

PRINTS is a database of 'fingerprints' – i.e., groups of conserved motifs excised from

sequence alignments that provide diagnostic signatures for particular protein families. PRINTS can be used to make family and tentative functional assignments for newly determined sequences, offering greater diagnostic reliability than single-motif approaches, such as those embodied in PROSITE [7]. A limitation of using databases like PRINTS and PROSITE to infer function is that the success of a search depends on the database size – if a family isn't encoded in the resource, a search will never find it. Unfortunately, both PROSITE and PRINTS are compromised by the fact that they are maintained and annotated manually; they therefore lag behind resources that are generated by largely automatic means (e.g., Pfam).

To address this problem, we developed a suite of tools to reduce the manual burdens inherent in PRINTS' maintenance. In so doing, we hoped to be able to increase the rate of growth of the resource and hence improve its diagnostic utility. It was not our intention to try to automate PRINTS' maintenance entirely, as the amount of human knowledge required to produce complete, fully annotated fingerprints is considerable, and is not readily captured computationally. Nevertheless, we devised an automatic pipeline in which sequence alignments are generated, motifs are excised, the source sequence database (a Swiss-Prot/TrEMBL composite) is searched for related family members, the scanning process is repeated until no further family members are identified, and the results are collated into 'naked' templates. By applying this pipeline to a public source of seed alignments (ProDom), we were able to generate large numbers of naked fingerprints in batch mode; we then developed a database to store the results, which we termed prePRINTS [8].

The next challenge was to develop a tool capable of injecting annotation into the templates. For pragmatic reasons, we decided to use Swiss-Prot rather than the literature: by contrast with the free-texts available in, say, MEDLINE abstracts, Swiss-Prot annotation is semi-structured and has already been culled from the literature by

human curators. Accordingly, we designed a tool to produce protein family reports from information in Swiss-Prot (see below), which we then fed into the prePRINTS pipeline. Thus, prePRINTS now provides an automatic, annotated supplement to PRINTS, from which entries can be migrated into PRINTS itself with relative ease.

PRECIS

There have been several endeavours to address the problem of annotating sequence data computationally, but the task is non-trivial and few tools have emerged that gather useful information on a given sequence or set of sequences in a simple and convenient manner. As more genome projects bear fruit, the mass of uncharacterised sequence data accumulating in public repositories grows ever larger, and with it the need for tools to facilitate their analysis and annotation. With this in mind, we developed a tool that automatically creates Protein Reports Engineered from Concise Information in Swiss-Prot (PRECIS) [9,10]. PRECIS was initially developed to help with the tedious task of annotating PRINTS entries (hence, in part, its reliance on Swiss-Prot). However, to make the software more generally useful, we implemented a BLAST interface, which allows it to be run with a single starting sequence and an E-value cut-off, below which Swiss-Prot matches are to be allowed into the annotation-culling process.

PRECIS processes the Swiss-Prot entries of matched sequences using a number of rules and filters to determine whether a protein superfamily, a discrete functional family or domain family is being analysed. It then collates the results in the form of a structured report detailing (where known) the function and structure of the protein, the disease(s) with which it is associated, the family to which the protein belongs, database and literature references, and a list of keywords. Figure 1 illustrates the report generated for the rhodopsin family, having BLASTed Swiss-Prot with TrEMBL entry Q9PUZ5, a triggerfish rhodopsin. The search returned 107 significant matches (with E-values $<1e-60$), for each of which the full entry was retrieved from Swiss-Prot. The

complete data-set, occupying 204 pages of text, was analysed and processed, and the results distilled into a 1.5-page report (Fig. 1).

The principal limitation of PRECIS is that the current implementation is entirely dependent on the quality of pre-existing annotation in Swiss-Prot (if Swiss-Prot annotation is poor, incorrect or absent, PRECIS will inherit its defects); its value, however, is in moving beyond simply providing lists of keywords (as do tools such as the Protein Annotator's Assistant), endeavouring as it does to construct concise protein reports that are English-like. PRECIS is thus a useful annotation tool.

BioMinT

As mentioned above, using Swiss-Prot alone as a source of annotation is problematic, as any errors or deficiencies it has are simply inherited by PRECIS. To address this problem, we are re-visiting on-line texts, to try to refine our protein reports with information extracted from the literature – PRECIS was originally configured to collect MEDLINE abstracts referenced in each Swiss-Prot entry analysed, but we had neither the time nor in-house expertise to properly make use of this information. However, we now lead a European project, BioMinT, for biological text mining and database annotation, which involves 6 partners, including the Swiss Institute of Bioinformatics and the University of Geneva, through which we hope to develop the next generation of annotation tools for Swiss-Prot and PRINTS.

CINEMA/UTOPIA

At the end of its last grant period, EMBnet provided funds to extend the CINEMA alignment editor and render it compatible with EMBOSS (in fact, the project started the day before the EMBnet core grant terminated!). This project was fraught with difficulties from the outset, primarily as a result of problems acquiring and retaining suitable staff. A new version of CINEMA was eventually produced, termed CINEMA-MX [11], but the package was a poor replacement for the original package and lacked much of its functionality.

Rhodopsin

PRINTS; PR00237 GPCRRHODOPSN; PR00238 OPSIN; PR00579 RHODOPSIN
 PROSITE; PS00237 G_PROTEIN_RECEP_F1_1; PS00238 OPSIN; PS50262 G_PROTEIN_RECEP_F1_2
 PFAM; PF00001 7tm 1
 INTERPRO; IPR000276; IPR001760
 PDB; 1BOJ; 1BOK; 1EDS; 1EDV; 1EDW; 1EDX; 1F88; 1FDF
 SCOP; 1BOJ; 1BOK; 1EDS; 1EDV; 1EDW; 1EDX; 1F88; 1FDF
 CATH; 1BOJ; 1BOK; 1EDS; 1EDV; 1EDW; 1EDX; 1F88; 1FDF
 MIM; 163500; 180380; 268000

1. FYHRQUIST, N., DONNER, K., HARGRAVE, P.A., MCDOWELL, J.H., POPP, M.P. AND SMITH, W.C.
 Rhodopsins from three frog and toad species: sequences and functional comparisons.
 EXP.EYE RES. 66 295-305 (1998).

2. ARCHER, S.N. AND HIRANO, J.
 Opsin sequences of the rod visual pigments in two species of Poeciliid fish.
 J.FISH BIOL. 51 215-219 (1997).

3. HUNT, D.M., FITZGIBBON, J., SLOBODYANYUK, S.J., BOWMAKER, J.K. AND DULAI, K.S.
 Molecular evolution of the cottoid fish endemic to Lake Baikal deduced from nuclear DNA
 evidence.
 MOL.PHYLOGENET.EVOL. 8 415-422 (1997).

4. GALE, J.M., TOBEY, R.A., D'ANNA, A.
 Localization and DNA sequence of a replication origin in the rhodopsin gene locus of
 Chinese hamster cells.
 J.MOL.BIOL. 224 343-358 (1992).

5. PALCZEWSKI, K., KUMASAKA, T., HORI, T., BEHNKE, C.A., MOTOSHIMA, H., FOX, B.A., LE
 TRONG, I., TELLER, D.C., OKADA, T., STENKAMP, R.E., YAMAMOTO, M. AND MIYANO, M.
 Crystal structure of rhodopsin: a G protein-coupled receptor.
 SCIENCE 289 739-745 (2000).

6. YEAGLE, P.L., ALDERFER, J.L. AND ALBERT, A.D.
 Structure of the third cytoplasmic loop of bovine rhodopsin.
 BIOCHEMISTRY 34 14621-14625 (1995).

7. YEAGLE, P.L., SALLIUM, A., CHOPRA, A., BHAWSAR, N., ALI, L., KUZMANOVSKI, G.,
 ALDERFER, J.L. AND ALBERT, A.D.
 Structures of the intradiskal loops and amino terminus of the G-protein receptor,
 rhodopsin.
 J.PEPT.RES. 55 455-465 (2000).

Function:

Visual pigments are the light-absorbing molecules that mediate vision. They consist of an
 apoprotein, opsin, covalently linked to cis-retinal.

Additional Info:

Integral membrane protein.

Disease:

Defects in rho are one of the causes of autosomal dominant retinitis pigmentosa (adrp).
 Patients typically have night vision blindness and loss of midperipheral visual field; as
 their condition progresses, they lose their far peripheral visual field and eventually
 central vision as well. (OPSD_HUMAN).

Defects in rho are one of the causes of autosomal recessive retinitis pigmentosa (arrrp).
 (OPSD_HUMAN).

Defects in rho are also one of the causes of congenital stationary night blindness
 (csnb4). (OPSD_HUMAN).

Family and structural information:

The structure has been determined, e.g., "Crystal structure of rhodopsin: a G protein-
 coupled receptor" [5]; "Structure of the third cytoplasmic loop of bovine rhodopsin"
 [6]; "Structures of the intradiskal loops and amino terminus of the G-protein receptor,
 rhodopsin" [7].

Belongs to family 1 of g-protein coupled receptors. Opsin subfamily.

Keywords: Photoreceptor; Retinal protein; Transmembrane; Glycoprotein; Vision;
 Phosphorylation; Lipoprotein; Palmitate; G-protein coupled receptor; Acetylation; 3D-
 structure; Retinitis pigmentosa; Disease mutation.

Figure 1. Example PRECIS output, showing the report generated for the rhodopsin family using TrEMBL entry Q9PUZ5 (E-values <1e-60). From this single sequence, 107 significant matches were returned and distilled fully automatically into the concise report shown.

Determined not to give up, we embarked on a larger project funded by the e-Science Centre for the North West (ESNW), in which EMBnet is also a partner. Ambitiously referred to as UTOPIA (User-friendly Tools for OPERating Informatics Applications), this project aims to develop an easy-to-use bioinformatics analysis environment. It will provide a platform for building visualisation tools and applications capable of high performance data visualisation that is transparently accessible from Grid-enabled resources, and supporting real-time collaboration between geographically distant sites. It includes, at its heart, CINEMA3D (illustrated in Figure 2) and many additional sequence analysis, data-base search and visualisation tools. The first release of the CINEMA component is now available for download.

BioActivity/EMBER

BioActivity is an interactive bioinformatics Web practical with links to an introductory text-book [6, 12]. The book provides background material for the concepts, tools and resources used in the practical, setting the scene for one of the later chapters (a worked example through each of the practical steps); in turn, the practical provides on-line

excerpts from the book, supplemented with colour pictures where appropriate.

In order to extend this idea and, crucially, update the Website, in 2001, the European Commission funded the development of a new European Multimedia Bioinformatics Educational Resource, EMBER [13,14]. EMBER was conceived as a mechanism both to bridge the skills gap in bioinformatics and to foster academic and private-sector collaboration in the development of standardised, multimedia courses in bioinformatics. There are 10 consortium partners, all of which are EMBnet Nodes.

Eventually, EMBER will offer a self-contained, interactive bioinformatics Web-tutorial, the equivalent stand-alone course on CD-ROM, and an accompanying introductory text-book. The decision to use conventional text, coupled with Web- and CD-based media, should ensure that students in European domains for whom Internet access is not optimal have access to the same educational materials. EMBER's path has not been smooth, and we hope to extend the project beyond its 2-year funding period to see it through to completion. Figure 3 illustrates the

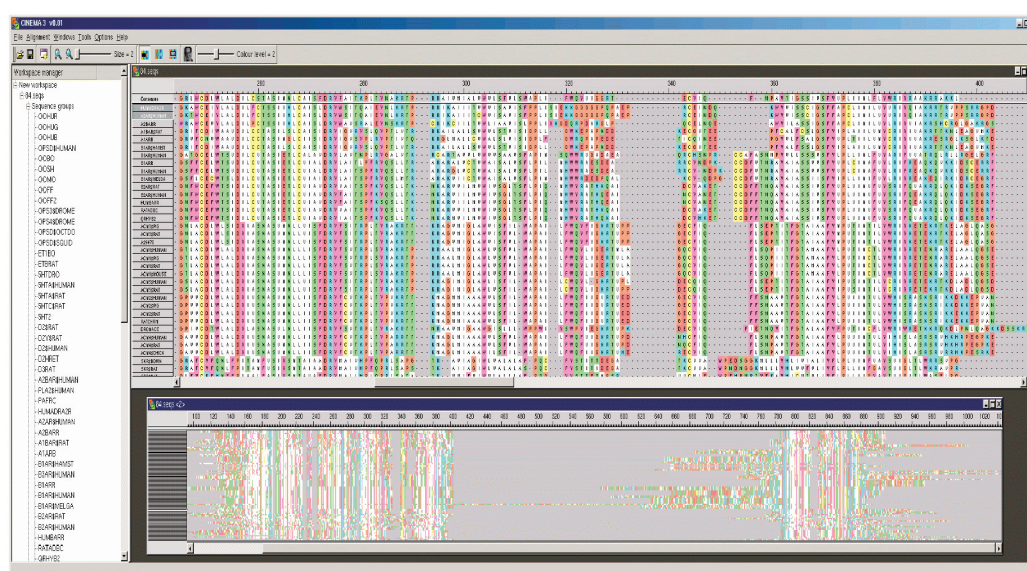
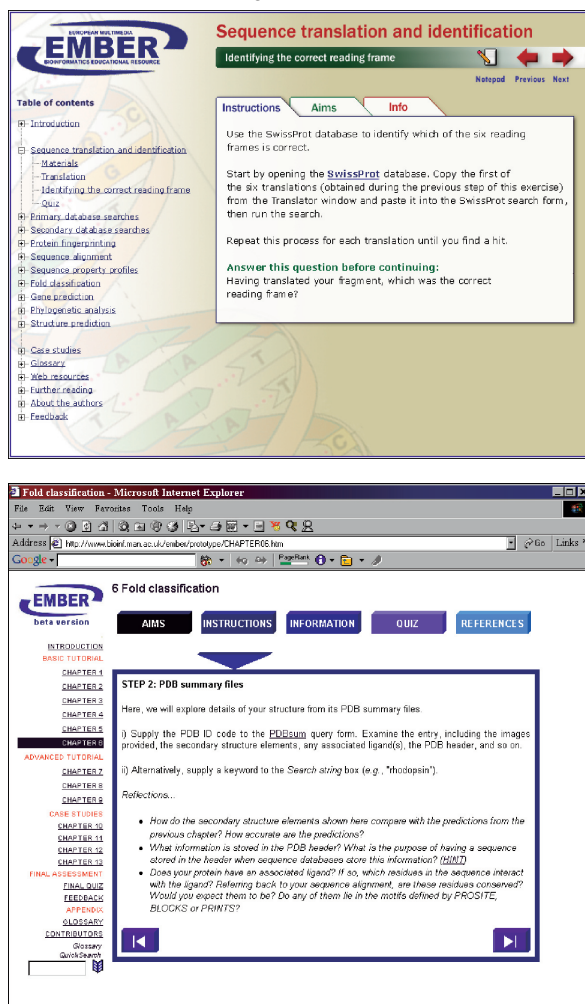


Figure 2. Screen-shot from the CINEMA3D module of UTOPIA illustrating part of a multiple alignment with an overview of the full alignment beneath.

prototype being developed at ETI, contrasting with the site available at UMBER for use in the meantime.

Figure 3. Screen-shots illustrating the look-&-feel of the ETI prototype Website (top), and our interim solution while the ETI site is being developed.



Book

Finally, we are negotiating with Prentice Hall to write a second edition of Introduction to Bioinformatics, first published in 1999 by Addison Wesley Longman [12]. Meanwhile, our colleagues at the Chinese National Node have translated the book into... Chinese (!); a Spanish translation has also just been published, and a low-cost version is also now available for the Indian market.

Web Addresses

UMBER

<http://www.bioinf.man.ac.uk/dbbrowser/PRINTS>

<http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/>

PrePRINTS

<http://www.bioinf.man.ac.uk/dbbrowser/preprints/>

PRECIS

http://www.bioinf.man.ac.uk/cgi-bin/db-browser/precis/blast_precis.cgi

BioMinT

<http://www.bioinf.man.ac.uk/dbbrowser/biomint/>

CINEMA

<http://www.bioinf.man.ac.uk/dbbrowser/CINEMA2.1/>

CINEMA3D

<http://aig.cs.man.ac.uk/utopia/>

BioActivity

<http://www.bioinf.man.ac.uk/dbbrowser/bioactivity/>

EMBER

<http://www.bioinf.man.ac.uk/dbbrowser/ember>

References

1. Attwood, TK, Beck, ME, Bleasby, AJ and Parry-Smith, DJ (1994). PRINTS - A database of protein motif fingerprints. *Nucleic Acids Research*, 22(17), 3590-3596.
2. Selley, JN and Attwood, TK (2001) EASY - an Expert Analysis SYstem. *Bioinformatics*, 17(1), 105-106.
3. Scordis, P, Flower, DR and Attwood, TK (1999). FingerPRINTScan: Intelligent searching of the PRINTS motif database. *Bioinformatics*, 15(10), 799-806.
4. Wright, W, Scordis, P and Attwood, TK (1999). BLAST PRINTS - An alternative perspective on sequence similarity. *Bioinformatics*, 15(6), 523-524.
5. Parry-Smith, DJ, Payne, AWR, Michie, AD and Attwood, TK (1998). CINEMA - A novel Colour Interactive Editor for Multiple Alignments. *Gene*, 221, GC57-63.
6. Attwood, TK (2000). An interactive practical at the interface of Web-based and conventional publishing. CAL-laborate. A collaborative publication on the use

- of Computer Aided Learning for tertiary level physical sciences, 4, 1–6.
7. Attwood, TK (2002) The PRINTS database, a resource for identification of protein families. *Briefings in Bioinformatics*, 3(3), 252-263.
 8. Attwood, TK, Bradley, P, Flower, DR, Gaulton, A, Maudling, N, Mitchell, AL, Moulton, G, Nordle, A, Paine, K, Taylor, P, Uddin, A and Zygori, C (2003) PRINTS and its automatic supplement, pre-PRINTS. *Nucleic Acids Research*, 31(1), 400-402.
 9. Reich, JR, Mitchell, A, Goble, CA and Attwood, TK (2001) PRECIS: Protein Reports Engineered from Concise Information in SWISS-PROT. *IEEE Intelligent Systems*, 16(6), 42-51.
 10. Mitchell, AL, Reich, JR, and Attwood, TK (2003) PRECIS: Protein Reports Engineered from Concise Information in SWISS-PROT. *Bioinformatics*, accepted.
 11. Lord, P and Attwood, TK (2002) CINEMA-MX – a modular multiple alignment editor. *Bioinformatics*, 18(10), 1402-1403.
 12. Attwood, T.K. and Parry-Smith, D.J. (1999) *Introduction to bioinformatics*. Addison Wesley Longman, Harlow, Essex CM20 2JE, UK. ISBN 0 582 327881
 13. Mabey, J.E. and Attwood, T.K. (2001) EMBER: a European Multimedia Bioinformatics Educational Resource. CAL-laborate. A collaborative publication on the use of Computer Aided Learning for tertiary level physical sciences, Vol. 6, pp.13-16.
 14. Attwood, TK (2002) EMBER – A European Multimedia Bioinformatics Educational Resource. *EMBnet News*, 8(1), 15-1

Bioinformatics courses in Colombia



Laurent Falquet

Swiss EMBnet node manager, Swiss Institute of Bioinformatics
155, ch. des Boveresses,
CH-1066 Epalinges,
SWITZERLAND
Laurent.Falquet@isb-sib.ch

Introduction

Over the past 4 years I had the opportunity to travel to Colombia to teach bioinformatics. Although this country is known mostly for its violence, guerrilla and drug production, it has a friendly population that is eager to learn and to have contacts with foreign people. This experience is an enrichment and proves, if necessary, the needs for high level courses in developing countries.

History

The first contact I had with Colombia was in Cali in 1999 where I was invited with two colleagues by Prof. Socrates Herrera (Univ. del Valle) to give an introductory course. During this course I met Prof. Emiliano Barreto (Univ. Nacional) who is now the Colombian EMBnet node manager¹ member of the Biotechnology Institute of the National University². We organised three courses in Bogota, following the first one: introductory courses in 2000 and 2002, and advanced course in sequence analysis and introduction to microarrays data analysis in 2001.

The students

Each course is run over one full week with both morning and afternoon divided in theoretical and practical sessions. The students usually 30 per course, originated from several regions of Colombia (Bogota, Medellin, Cali, Cartagena) and several universities (Nacional, Los Andes, La Javeriana, Militar, CIAT, etc.). Most of them were biologists, pharmaceutical chemists, or computer scientists, and even some

local professors! We greatly appreciated their contact, smart questions and most importantly, their enthusiasm and willing to learn behavior.



Courses in 2001 (top) and 2002 (bottom)

The infrastructure

The teaching and the computer rooms are simple and equipped with recent PCs under Windows.

The network is sufficiently fast, but the bandwidth limits somehow the flow of pages. Before the first course in Bogota we helped them to set up the servers, a BLAST machine (dual proc Pentium under Linux) and a file server with users accounts and software, mainly the EMBOSS package plus several other tools and the databases (EMBL, SwissProt, TrEMBL). Currently they continue to maintain and enhance the services with a web portal and other tools.

The culture

Well my first effort was to try learn a bit of Spanish to improve the communication and year after year I enjoy to practice the spanish language. My first discovery came from the food and especially the fruits. They have an incredibly rich collection of tropical fruits that we don't even know about here in Europe! The few tourist places that we visited are really astonishing (e.g., the Gold Museum in Bogota or the Salt Cathedral of Zipaquirá)^{3,4}, but the most interesting feature of the Colombian culture is the music and especially the Salsa dance, that you hear everywhere and that you can even learn in a moving bus "la chiva rumbera"! (see picture below)



Conclusion

Finally I can only encourage openminded people to visit this beautiful country, a paradise for biologists interested in biodiversity, and to meet fine colombian people. Of course you should listen to them and behave cautiously when they tell you.

References

1. Colombian EMBnet node (<http://bioinf.ibun.unal.edu.co>)
2. University Nacional (<http://www.unal.edu.co>)
3. El Museo del Oro (<http://www.banrep.gov.co/museo>)
4. Cathedral de sal (<http://www.catedraldesal.com>)

Computational gene expression analysis; correlating data and knowledge



Torgeir R. Hvidsten

The Linnaeus Centre for
Bioinformatics
Uppsala University
Husargatan 3
SE-751 24 Uppsala
SWEDEN
Torgeir.Hvidsten@lcb.uu.se

Introduction

Physiological and pathophysiological responses are associated with specific changes in cellular gene expression and gaining insight into these specific patterns

enable hypotheses about gene function, understanding of gene regulation and promising new possibilities for molecular medical diagnosis. Microarray technology is a fundamental research vehicle in this context, allowing simultaneous readouts of relative gene expression levels for thousands of pre-selected genes in a tissue sample (Figure 1). Conduction several experiment using samples from different time points, clinical states and/or different treatments constitute a powerful setup for studying the genome-wide changes in cellular gene expression. However, this setup also raises non-trivial computational challenges not only related to storage and analysis of data, but also to evaluation and interpretation of the results according to current biological knowledge.

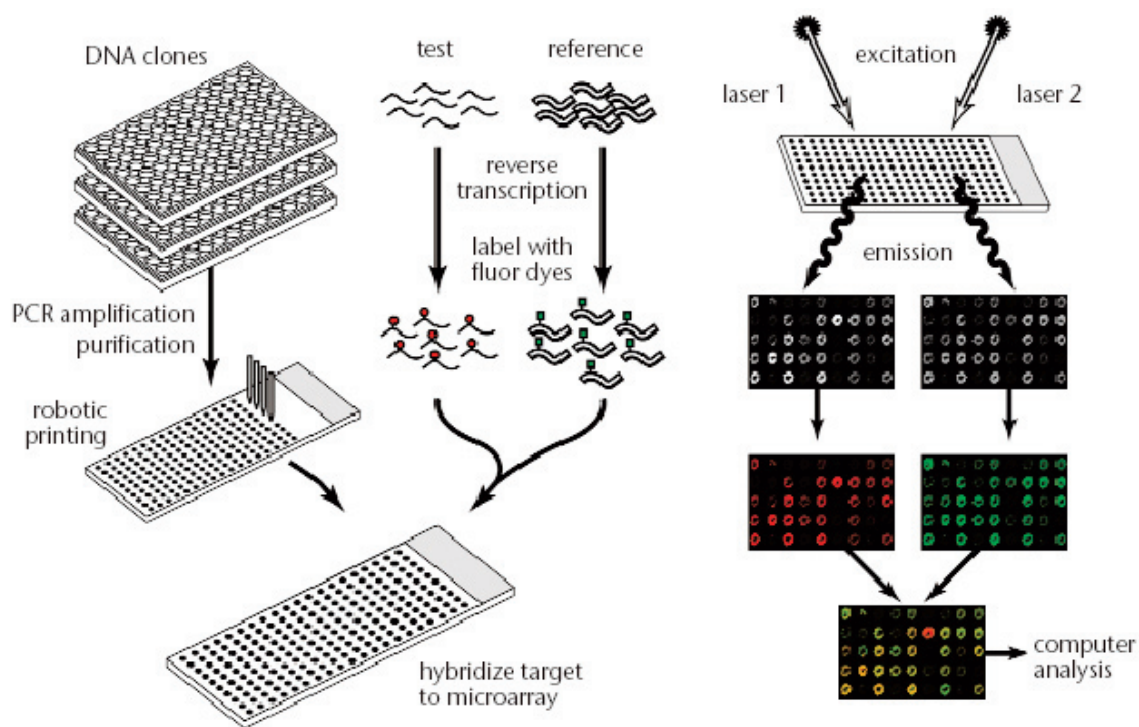


Figure 1 Microarray experiment. Gene probes of interest are printed on glass microscope slides. RNA from both test and reference sample are first fluorescently labeled with Cy3 (red) or Cy5 (green) and then hybridized to the clones on the array. The arrays are scanned and the resulting images are pseudo-colored and merged. Data from hybridization experiment is commonly viewed as normalized ratios Cy3/Cy5, indicating increased (>1), decreased (<1) or unchanged ($=1$) expression level in the test sample relative to the reference sample. Conduction several experiment using samples from different time points, clinical states and/or different treatments constitute a powerful setup for studying the genome-wide changes in cellular gene expression in living systems (D. J. Duggan et al., Nature Genetics 21, 10-14, 1999).

Gene expression data analysis

The typical intermediate result of a microarray study is the genes $\times$ samples – matrix where each row holds the expression profile of one gene over all samples and each column hold the expression profile of one sample over all genes. Common for most data analysis methods is that they are based on representing genes/samples as points in a multi-dimensional expression space spanned by the samples/genes (obviously, the exception is studies in which samples are taken from different time point and genes can be represented as functions of time in two dimensions).

Data analyses strategies typically try to reduce the complexity by assume that there are underlying, known or unknown, classes of genes or samples that corresponds to biological or medical phenomena. Such classes can, for examples, be groups of genes that code for proteins with the same molecular function or are involved in the same biological process, or groups of tissue samples that are taken from patients with the same disease or that underwent the same medical treatment. There are two conceptually different analysis strategies. The first strategy is called class discovery or clustering (1,2). Such methods try to discover biological related genes or samples by assuming that they will appear close to each other in the expression space. Hierarchical clustering is one commonly used clustering method in the context of gene expression analysis (Figure 2). The second strategy take advantage of already well studied examples and try to learn models that can be used both to gain insight into the parameters discriminating different classes and for prediction/diagnosis. These methods are called supervised learning or machine learning (3). Models can range from simply identifying genes that are differentially expressed in, for example, samples from different cancer types, through IF-THEN rules combining such discriminatory genes, to complex illegible models such as artificial neural networks and hyper-planes from support vector machines.

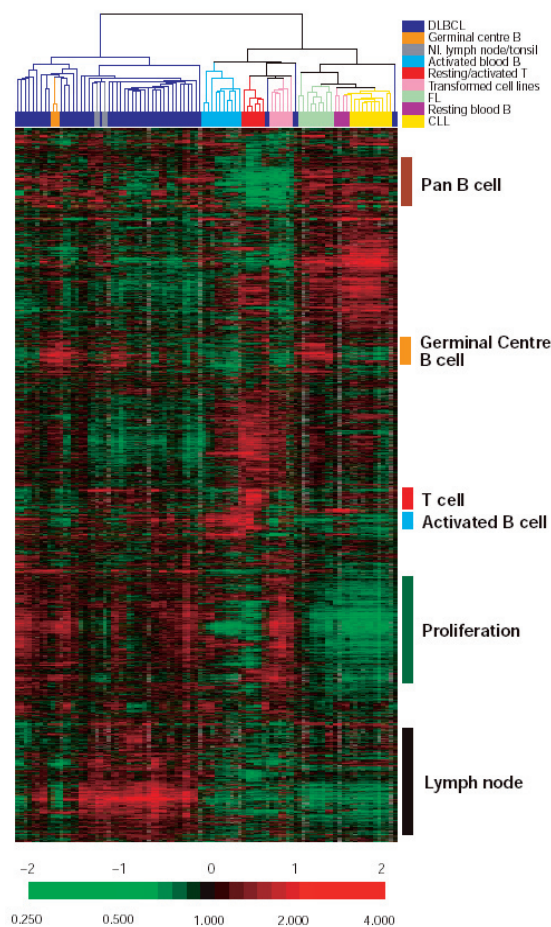


Figure 2 Hierarchical clustering. Hierarchical clustering organize samples in a tree that reflects their level of (expression) similarity. The figure shows the hierarchical clustering of 96 samples from normal and malignant lymphoma using expression profiles from almost 20 000 selected human genes. Diffuse large B-cell lymphoma (DLBCL) is known to be clinically heterogeneous in that only about 40% respond well to current treatment. The figure shows how this subtype is almost perfect clustered separately from the other samples. Furthermore, it was shown that the two sub-clusters of DLBCL (found using clustering) include patients with significantly different expected survival time indicating two distinct molecular types of DLBCL (Alizadeh et al., Nature 403, 503-511, 2000).

Common to both clustering and machine learning methods is the need to validate results in order to assess their biological or medical relevance. While early studies on the relation between gene function and gene expression only visually hinted on expression similarity within groups on functional related genes, newer studies typically test the

statistical significance of overlap between expression clusters and functional classes or even build models capable of hypothesizing novel functions for genes. Such computerized use of large amount of biological knowledge, in this case knowledge of gene function, requires structured databases and controlled vocabularies (6). Building such databases again requires reading thousands of articles in order to extract the relevant information. As a consequence, developing new tools for text-mining (automatic extraction of knowledge from text) has become one of the most important research areas in Bioinformatics (7).

Selected data analysis tools and further reading

The research community early recognized the need for standardization to facilitate data management, processing, transfer, publication and reproducibility of microarray data. This has resulted in Laboratory Information Management Systems (LIMS) sharing common data formats (MIAME: Minimum Information about a Microarray Experiment), and data warehouses such as ArrayExpress (4) and the Gene Expression Omnibus (5) publishing microarray studies. There is a wide variety of freely available systems and services applicable to gene expression data analysis on the web. Some recommended sites and directions for further reading are indicated here:

References

1. Cluster and TreeView: various clustering algorithms including a much used visualization tool for hierarchical clustering (<http://rana.lbl.gov/EisenSoftware.htm>).
2. J-Express: cluster and statistical analysis of gene expression data (<http://www.ii.uib.no/~bjarted/jexpress/>).
3. The ROSETTA system: a general toolkit for IF-THEN rule-based supervised learning (<http://www.idi.ntnu.no/~aleks/rosetta>). Case study using ROSETTA for functional gene expression analysis: <http://www.lcb.uu.se/~hvidsten/fibroblast/>.
4. ArrayExpress: public database for microarray studies by the European Bioinformatics

Institute (EBI) (<http://www.ebi.ac.uk/arrayexpress>).

5. Gene Expression Omnibus: public database for microarray studies by the National Center for Biotechnology Information at NIH (<http://www.ncbi.nlm.nih.gov/geo/>).
6. Gene Ontology: a dynamic controlled vocabulary of gene and protein roles in cells (<http://www.geneontology.org/>).
7. PubGene: gene networks automatically extracted from literature (<http://www.pubgene.uio.no/>).
8. The Chipping Forecast II, freely available supplement to Nature Genetics, 32: 461 – 552, 2002 (http://www.nature.com/ng/web_specials/)

Announcements

BEN - the Belgian EMBnet node (<http://www.be.embnet.org>) - has been invited to give a workshop in Bioinformatics, the 3 and 4 April 2003 at the "Université de la Méditerranée" (<http://mediterranee.univ-mrs.fr/>), in Marseille, France. Guy Bottu and Valérie Ledent ran the workshop in front of an attendance of 24 bioinformaticians and IT specialists.

This workshop (<http://www.esil.univ-mrs.fr/%7Eethieffry/BEN/index.html>) was organised in the frame of the DESS CCI (Compétences Complémentaires en Bioinformatique; <http://www.dil.univ-mrs.fr/desscci/>). It focused on the Software Suites Staden and EMBOSS, the EMBOSS web interface WEMBOSS developed by BEN and the Sequence Retrieval System SRS. The focus was on the overview of the functionalities of these software packages, as well as their installation and maintenance on Linux machines.

Other courses and meetings

See our web site:

http://www.embnet.org/activities/dyn_schedule.php

You can even use this page to announce your own courses and meetings (OBS: this service is moderated).

A guide on how to install Bioinformatics tools on MacOSX (Part I)



Erik Bongcam-Rudloff
Assistant Professor
Swedish University of
Agricultural Sciences,
The Linnaeus Centre for
Bioinformatics

MacOSX an introduction

In 1999 Apple Computer released its first version of Macintosh OS X Server. This next generation operating system was the precursor to their Macintosh OS X. This new member of the Unix family unifies the power of Unix with the ease-of-use and sophisticated applications of a consumer O/S.

Macintosh OS X is a combination of Apple's OS, NeXT Computer's OpenStep OS, and UNIX operating systems. The system has a Mach micro kernel; and a BSD layer on top of that. The kernel for MacOSX is UNIX based (Mach from Carnegie-Mellon University and 4.4BSD from the University of California at Berkeley). The kernel for MacOSX is called Darwin and is an open-source development¹.

Most command-line and X11² programs that can be compiled under Linux³ recompile under MacOSX without difficulty. It is often a function of the compiler (in this case gcc 3.1)⁴ and the available libraries (which are also typically the same open source ones on both systems). When installing OS X, Apple puts the command-line application in a less obvious place to reduce the chance that a novice would accidentally launch it, but it is definitely there. Go to /Applications/Utilities and you will find the "Terminal", put it on the Dock (just click on it and place it on the Dock, you will make heavy use of this terminal). MacOSX is designed so that an ordinary user never needs to know anything about Unix, but it keeps all the Unix power available to those that need/want it.

Elements of OS X

MacOSX⁵ is made up of four basic elements when viewed in the simplest way.



- a-The kernel (Darwin)
- b-The imaging layer (Quartz, OpenGL, QuickTime)
- c-The application environments (Classic, Carbon, Cocoa, Java)
- d-The graphical user interface, or GUI (Aqua)
(We will add X11 at this level!)

The imaging, application and GUI layers sit on top of the Darwin core. It is important to keep this in mind to understand why we install the X-windows system in parallel with Apples GUI (Aqua).

MacOSX and Bioinformatics

In the first article of this series you will learn an easy way to install some of the basic and most popular Bioinformatics packages on a Macintosh running MacOSX. In the next article we will discuss how to use the programs and how to expand this collection with other tools. After this simple step by step tutorial you will have on your hands a powerful workstation to accomplish most of the common bioinformatics tasks.

Prerequisites

A PowerPC computer running MacOSX 10.2.x. You should also install the freely available "Developers tools"⁶ (Do not install Apples latest X11 (as of 020403). Use the X11 provided in this tutorial. Apples X11 have some bugs that we fixed in our release.)

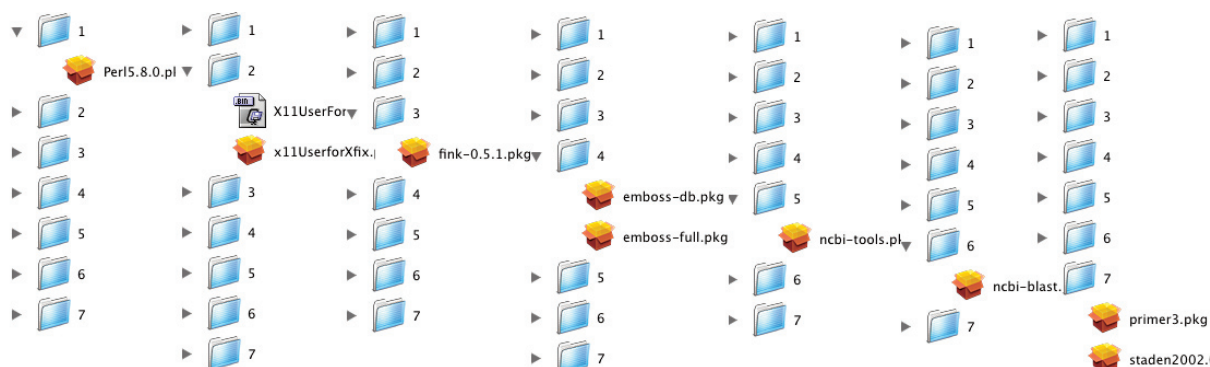
If you fulfill these prerequisites it is now time to start installing the packages as described in this tutorial.

Part 1

Go to <http://www.ebioinformatics.org>

Download all the packages on macosx/software/download.htm.

2.6.0: The European Molecular Biology Open Software Suite. EMBOSS is a new, free Open Source software analysis package specially developed for the needs



Please install the packages following exactly the indicated steps on the next figure (it is essential to do this in the same order as some packages depend on libraries and programs installed in the previous steps!):

The following packages will be installed:

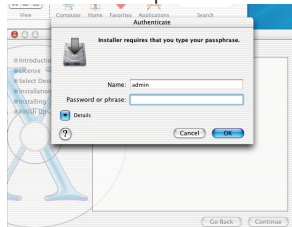
- 1- Perl 5.8.0⁷ (this is needed because MacOSX provides Perl 5.6)
- 2- X11, the X Window System that makes it possible to run X11-based applications in MacOSX. Based on the open source XFree86 project⁸. Add the fix provided, this will be changed in the future!
- 3- Package 3 installs Fink⁹ with some libraries and programs needed for the rest of this installation. The Fink project brings the full world of Unix Open Source¹⁰ software to Darwin and MacOSX. The people involved in the Fink project modify Unix software so that it compiles and runs on MacOSX ("port" it) and make it available for download as a coherent distribution. Fink uses Debian¹¹ tools like dpkg and apt-get to provide powerful binary package management. You can choose whether you want to download precompiled binary packages or build everything from source. Later on you will be able to use programs like "dselect" to install hundred of programs to your system. With our package we install Pymol¹² (among others) for demonstration purposes.
- 4- In this step we install EMBOSS¹³ version

of the molecular biology (e.g. EMBnet¹⁴) user community. Together with this installation you will also install additional packages (e.g., HMMER-2.1.1¹⁵, with Alivec-enhancements, Domainatrix¹³, Phylip-3.573c¹⁶). Add also the second package to install the necessary databases (e.g., Rebase: the restriction enzyme database maintained by New England Biolabs. Rebase is needed for programs such as remap and restrict.).

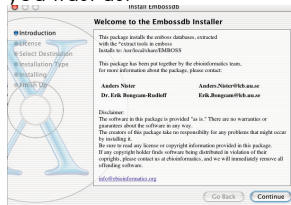
- 5- Now you can add the NCBI Toolbox¹⁷ i.e., the NCBI software tools for building bioinformatics resources. This toolbox is needed for adding the next package.
- 6- This package installs the most common tool in Bioinformatics: The BLAST® (Basic Local Alignment Search Tool) a set of similarity search programs. In this installation you will install the Alivec enhanced version¹⁸ (OBS. optimised only for blastn).
- 7- The last step. You will install Primer3¹⁹ and the Staden²⁰ package. The Primer3 program is used by the EMBOSS program eprimer. The Staden package is the version-2002.0. This is the first official release of the MacOSX version (10.2) of the package. This version requires X11 and provides a nice GUI (Graphical User Interface) to the EMBOSS package. (**OBS. The Staden package is only free for academics and you need to apply for a license**).

Part 2

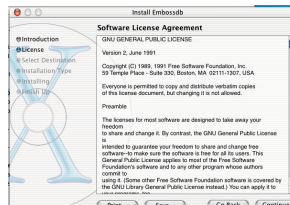
Step 1 Click twice on the first package. The first window will ask you for an administrator password.



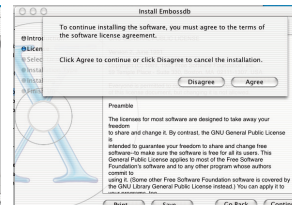
Step 2 You see a disclaimer with information about the people behind these packages. In this case Anders Nister and me. Continue if you trust us.



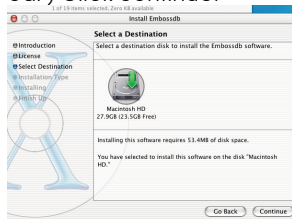
Step 3 Now you will read the license information related to each specific package



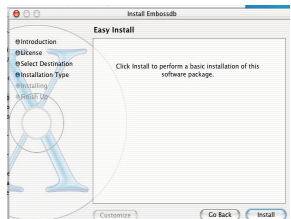
Step 4 Click agree to the license agreement (read it first!).



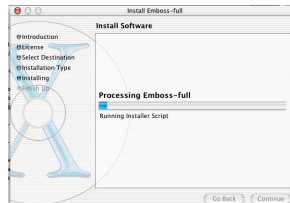
Step 5 Select the hard disk where you want the package to be installed (all packages install on /usr/local) Click continue.



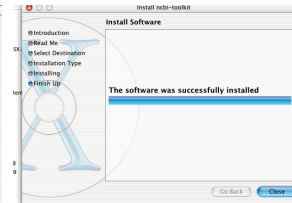
Step 6 Click on install to perform a basic installation. The packages are installed in /usr/local.



Step 7 The installation starts, wait a few seconds.



Step 8 This is the last step of each package. Repeat with all packages as indicated on figure 1.



Part 3

During installation you will have an automatic configuration of your environment parameters, this to tell the system that you have new programs and libraries installed. You can also adapt your new system in the way that suites your needs. I did one change for my computer (This is just an example!).

Open a new Terminal.app window and type "pico .cshrc". A text editor will pop up. Enter these lines:

```
source /sw/bin/init.csh
setenv PLPLOT_LIB /usr/local/share/emboss
set path=( /usr/local/bin /usr/local/blast/bin /usr/local/staden/macosx-bin ${path} )
setenv DISPLAY :0.0
```

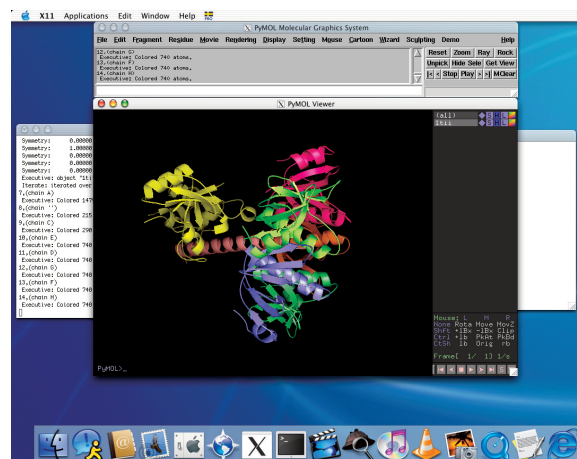
To get out of the editor, press control-O, return, control-X.

This works for me but if you are a user with some Unix knowledge you should adapt the system for your own needs!

The complete installation procedure takes less than one hour.

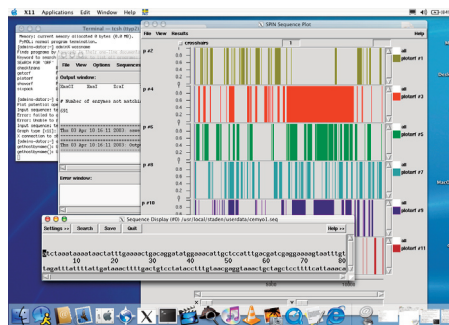
Now you are ready to test your new "Bioinformatics Workstation".

The first thing you do is to start the X11 environment, do this with the icon X11 in your Applications folder. Then open a new Terminal.app window and type: "pymol" You should see this when testing the demo files:



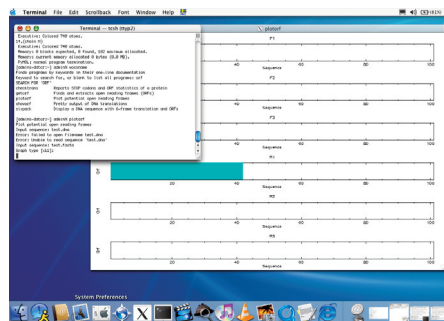
If you want to test the Staden package: Open a new Terminal.app window and type: "spin"

This starts spin, one of the Staden package modules. in Demo mode (Apply for a license!) but using the demo files and some testing you should be able to see this using plotorf on the EMBOSS menu:



The EMBOSS package can also be used directly on the command line (on the Terminal.app). Test a DNA sequence (in fasta format) using the EMBOSS program plotorf. You should see something similar to this if X11 is working correctly.

We installed the png²¹ (Portable Network Graphics) libraries with Fink and compiled EMBOSS with libpng so you are also able to produce png images.



I also strongly recommend to buy a three button mouse for your Mac. You will need it for most Bioinformatics applications under X11. MacOSX has support for it, the three button mice work without any configuration changes, just plug in!

The next article will discuss:

- a- how to install GUI- and web-interfaces to the command line programs,
- b- how to install other packages and
- c- how to build your own databases to use the blast programs

Acknowledgements:

Anders Nister graduate student at The Linnaeus Centre for Bioinformatics. Anders is the person responsible for the development of the MacOSX packages described in this article.

References

- 1- <http://developer.apple.com/darwin/>
- 2- <http://www.xdarwin.org/>
- 3- <http://www.linux.org/>
- 4- <http://gcc.gnu.org/>
- 5- <http://www.apple.com/macosx/>
- 6- <http://developer.apple.com/tools/macosxtools.html>
- 7- <http://www.perl.org/>
- 8- <http://www.xfree86.org/>
- 9- <http://fink.sourceforge.net/>
- 10- <http://www.opensource.org/>
- 11- <http://www.debian.org/>
- 12- <http://pymol.sourceforge.net/>
- 13- <http://www.emboss.org/>
- 14- <http://www.embnet.org>
- 15- <http://hmmer.wustl.edu/>
- 16- <http://evolution.genetics.washington.edu/phylip.html>
- 17- <http://www.ncbi.nlm.nih.gov/Ftp/index.html>
- 18- http://www.apple.com/downloads/macosx/math_science/applegenentechblast.html
- 19- http://www-genome.wi.mit.edu/genome_software/other/primer3.html
- 20- <http://www.mrc-lmb.cam.ac.uk/pubseq/>
- 21- <http://www.libpng.org/pub/png/>

Protein Spotlight

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.

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

We provide the EMBnet community with a printed version of issues 32 and 33.

Please let us know if you like this inclusion.

THE THINGS WE FORGET

By Vivienne Baillie Gerritsen

Why remember when it is so important to forget? That sounds promising in a day and age when techniques to enhance one's memory are as popular as a cup of tea. It is important not to forget your mother's birthday, your husband's cell phone number, the date you have to give in your income tax sheet but it is very desirable to forget all the items you float past in a supermarket, the names you scan through in a phone book, the sign posts you sweep past on your way home. We would like to think that we are the masters of what we choose to remember and what we choose to forget. No doubt, to a certain degree we are. But some initial sifting is done on the molecular level and for our own benefit. The understanding of the biochemistry involved in the process of learning, memory.....and forgetfulness is still very much in its infancy, but we do know of one protein, protein phosphatase 1, which certainly seems to be at the heart of this primordial sift.



Figure 1 Now, how did it go?

In humans, the necessity to forget is best portrayed by those who suffer from what has been called the 'Savant Syndrome'. This disorder was first described in the late 1900s by the English clinician Dr J. Langdon Down (1828-1896) – after whom by the way the Down Syndrome is named - who was then the Medical Superintendent of the Royal Earlswood Asylum for Idiots. In fact, the disorder was initially termed 'Idiot Savant' but with time the word 'idiot' was discarded because of its contemporary pejorative connotation. The Savant Syndrome was brought to public awareness thanks to Dustin Hoffman's performance in 'Rain Man' in the 1980s, where Hoffman played the part of an autistic savant based on the life of the 'real'

Rain Man, Kim Peek¹, born in 1951 and still alive today. While reading a book, Kim's parents would run his finger along each sentence and before the age of two, Kim was already memorizing every book that was read to him. In fact, he has read and can recite 7600 books - amongst many other forms of expertise.

Though this type of disorder is fascinating, it can be a bit of a burden on the normal psyche. It no doubt depicts less an overwhelming faculty to memorize things than an awesome inability to forget things. The mechanisms, underlying the process of learning and memory, were introduced by the American psychologist Donald Hebb (1904-1985) who developed the theory which states that synapses are the key elements of memory. The term 'synaptic plasticity' soon bounced into the playing field. It is believed that memory is the result of a synaptic change in size, shape and sensitivity brought on by the synthesis of various proteins. And this change is the work of complex biochemical pathways - one of which protein phosphatase 1 (PP1) belongs to.

What exactly does PP1 do? PP1 hinders the synthesis of proteins involved in synaptic plasticity, hence in the process of memory. And how does it do this? As its name suggests, it is a phosphatase and, as all phosphatases, it deprives another molecule of its phosphate thereby modifying its function. In this event, PP1 binds to the gene transcription factor CREB (cyclic AMP-dependent response element binding) and dephosphorylates it. CREB is subsequently inactivated and the transcription of a number of 'memory' genes checked.

When though does PP1 pick its moment to inhibit CREB? Obviously, PP1 should not always be active otherwise we would remember nothing at all. The German psychologist Hermann Ebbinghaus (1850-1909) was the first person to study memory scientifically and showed very elegantly –

¹

<http://www.wisconsinmedicalsociety.org/savants/kimpeek.cfm>

what we know instinctively – that one memorizes information which is given in little spurts (distributed) much better than information given in one large gush (massed). This may remind some (or not, precisely) of wearisome, unending, informative lectures and brain-cramming sessions before exams, which in the long run have not left much of an impression on our memories.

When PP1 is inhibited in mice, there is no difference between distributed learning and massed learning. This means that the gene transcription factor CREB is left to get on with its job unimpeded whatever the situation. It is thought that in the event of massed learning, the synthesis of PP1 increases, thereby checking CREB function and putting a constraint on what we learn and remember. In the event of distributed learning, PP1 synthesis does not have the chance to increase so sharply and its effect on CREB is weaker. What is more, the longer the interval between two sessions, the better the imprint of the memory, probably because the level of PP1 decreases and CREB is given time to settle and function as it should. So, in effect, our memorizing faculties depend on a certain PP1 threshold.

More to the point, when PP1 is inhibited *after* the memory process, the memory is long-lasting, which means that in its natural state, PP1 has a direct effect on forgetfulness.

The findings are interesting. Memory decline as we get older is probably due, not to the wear and tear of our brain as we could easily imagine or, for that matter, to a lack of space, but to an increasing amount of PP1. It shouldn't be too difficult to conjure up some memory enhancing therapy then. But it is. PP1 is 'just' a phosphatase. Phosphorylation and dephosphorylation are among the most common biochemical processes and are used as on/off switches in many biological instances. Inhibiting PP1 could have drastic effects – its clients are not only the gene transcription factor CREB. However, the knowledge of PP1 function does shed light on the very complex path of learning, memory and forgetfulness and could pave the way to subtle therapies in the event of Alzheimer's disease for instance, brain traumas and indeed the withering of our memories due to the simple passing of time.

Cross-references to Swiss-Prot

P08129: human, *Mus musculus* (mouse), *Rattus norvegicus* (rat) and *Oryctolagus cuniculus* (rabbit) protein phosphatase PP1-alpha
P37140: human protein phosphatase PP1-beta
P36873: human protein phosphatase PP1-gamma
P37139: *Mus musculus* (mouse) and *Rattus norvegicus* (rat) protein phosphatase PP1-gamma

References

1. Silva A.J., Josselyn S.A.
The molecules of forgetfulness
Nature 418:929-930(2002)
PMID: 12198533
2. Sweatt J.D.
Memory mechanisms: the yin and yang of protein phosphorylation
Curr. Biol. 11:R391-R394(2001)
PMID: 11378403
3. Gross M.
Mysteries of the memory molecules
http://www.chemsoc.org/chembytes/ezine/2002/gross_dec02.htm

MOODY WALLPAPER

By Vivienne Baillie Gerritsen

How would you like to have wallpaper that changes color according to the seasons? Imagine a warm ochre hue for the cold winter months and a fresh yellowy tint for the hot summer months. Fantasizing? No, not really. Researchers in Madagascar recently discovered a very peculiar protein from the epidermis of a rare species of chameleon, *Chamaeleo differensis*, which is in fact a pigment and takes on different hues depending on various environmental criteria. The protein was baptized, in an outburst of scientific originality, chameleonin.

Chameleons are found mainly in Africa, Europe and Asia but over 50% of the world's chameleon species live in Madagascar. The first description of a chameleon dates back to the late 18th century. It was made by Jeanne Baret, the traveling companion of the renowned botanist Philibert Commerson (1728-1773), whom she accompanied in his numerous travels, amongst which a visit to the island of Madagascar. She remarked the peculiarity of the independently revolving eyes, the curled prehensile tail, the swiftness of the chameleon's tongue and, naturally, its capacity to change color. In her diaries, Jeanne Baret also mentions the use of dried chameleon flesh in folk medicine. Still today in Northern Africa, women who fear that their men have strayed a little add parts of chameleon to their food believing that this restores their fidelity.

Why is it that chameleons change color in the first place? It is the popular belief that they change for camouflage purposes and that they can produce any color on the market. This is not the case. The change of color in chameleons – which can occur in only 20 seconds – is an answer to the surrounding temperature, light and mood. An exasperated chameleon does not portray the same colors as a composed one; a warm chameleon does not look like a cool one; a sexually aroused chameleon bears no resemblance to one that shows no lust. And all this is due to a number of pigment cells, or chromophores, which lie under the surface of the skin. Hormonal, light or temperature differences affect the distribution of the pigments and the light

which is reflected off the epidermis offers a panel of colors, which range from browns to greens to blues, reds, yellows, black and white.



Figure 1 *Chamaeleo differensis* above 18° Celsius.

Chameleonin, however, does not function in the same way at all. Unlike the other pigments, it is not found in the chromophores under the skin but on the skin's surface itself and seemingly only in *Chamaeleo differensis*. Chameleonin is not distributed regularly on the chameleon's skin but appears in patches along the animal's spine. In fact, it is thanks to this odd distribution that the pigment was discovered in the first place. Talisker and his research team were studying the distribution of color on *Chamaeleo differensis*' body under different conditions. To speed things up, part of the team were carrying out their research in the main lab while the others had moved down to the basement. In time, they noticed that despite carrying out identical tests on the chameleons, those in the basement sported a row of very small bright orange spots down their spine. In a true rational procedure, the chameleons from the basement were moved upstairs and vice-versa. And, lo and behold, the ex-lab chameleons also developed the spinal row of bright orange spots.

The next question was what was it that caused this orange eruption? After much investigation, it became apparent that it was the difference of temperature between the lab and the basement that was responsible for the color change. A closer inspection of the chameleons' skin actually revealed the same row of spots at any given temperature. However, they remain

invisible unless the temperature is quite low, because they adopt the natural yellowish/beige hue of the chameleon. Chameleonin is a globular protein with an end which protrudes like an antenna. Below 18°C the protein's globular entity is deformed; the 'antenna' undergoes structural stress and its shape is modified accordingly. It is this change of shape which results in a change of color along the chameleon's spine, from the yellowish-brown hue to a bright orange. This phenomenon is similar to that which exists in lobsters, whose shell turns from blue to red when boiled*.

It is not clear why *Chameleo differensis* portrays this colorful display in the first place. Since all the individuals studied were in fact males and females don't seem to sprout orange patches, it is thought that it may be part of the mating game. *Chameleo differensis* does

indeed mate only at night when temperatures are relatively low and the bright orange spots are clearly visible in the dark. One surprising development is the very recent use of chameleonin in interior decoration. Wallpaper, smeared with the protein trapped in a solution, is now available on the market and all the rage in Paris, London and New York. Above 18°C, the wallpaper glows yellow; below 18°, a bright orange appears. The results are so encouraging that they will be stretched to the world of fashion design, materials and makeup. The next step, naturally, is to modify chameleonin's antenna to produce a wide range of colors. However, as Talisker remarks, 'a sitting-room is rarely below 18°C and if it is, no-one would be sitting there, what's more there are perhaps more pressing issues to address than the color of your wallpaper and evening dress.'

* Read Spotlight issue 26.

Cross-references to Swiss-Prot

P99834: *Chameleo differensis* chameleonine

References

1. Talisker J., Oban T., Scapa W.
Chameleonin: a chromomodulated protein from *Chameleo differensis*
Helv. J. Zool. 271:45-32(2003).
2. Dimaki M., Valakos E.D., Legakis A.
Variation in body temperatures of the African Chameleon *Chamaeleo africanus* Laurenti, 1768 and the Common Chameleon *Chamaeleo chamaeleon* (Linnaeus, 1758)
Belg. J. Zool. 130:87-91(2000).
3. Raxworthy C.J.
A truly bizarre lizard
<http://www.pbs.org/edens/madagascar/creature3.htm>

National Nodes

Argentina

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

Australia

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

Austria

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

Belgium

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

Brasil

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

Canada

Laura Brown
Canadian Bioinformatics Resource, National
Research Council Canada, Institute for
Marine Biosciences,
Email: manager@cbr.nrc.ca
Tel: +1-902-426 7310 Fax: +1-902-426 9413
<http://www.ca.embnet.org>

Chile

Dr. Ricardo Baeza-Yates
Dept. of Computer Science, Santiago,
Email: rbaeza@dcc.uchile.cl
Tel: N/A
<http://www.embnet.cl>

China

Jingchu Luo
Room 303, Exchange Centre, Peking Univer-
sity
Email: luojc@cbi.pku.edu.cn
Tel: +86-10-6275 9001
<http://www.cbi.pku.edu.cn>

Colombia

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

Cuba

Ricardo Bringas
Centro de Ingeniería Genética y Biotecnol-
gía, La Habana,
Email: bringas@cigb.edu.cu
Tel: +53 7 218200
<http://www.cu.embnet.org>

Denmark

Hans Ullitz-Moeller
BioBase, University of Aarhus
Email: hum@biobase.dk
Tel: +45-86-13 9788
<http://www.dk.embnet.org>

Finland

Kimmo Mattila
CSC, Espoo
Email: Kimmo.Mattila@csc.fi
Tel: +358 9 457 2708
<http://www.fi.embnet.org>

France

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

Germany

Sandor Suhai
EMBnet node at the German Cancer
Research Center
Department of Molecular Biophysics (H0200)
Email: genome@dkfz.de
Tel: +49-6221-422 342 Fax: +49-6221-422 333
<http://www.de.embnet.org>

Greece

Babis Savakis
Institute of Molecular Biology and Biotech-
nology
Heraklion, Crete
Email: savakis@nefeli.imbb.forth.gr
Tel: +30-81-391 114 Fax: +30-81-391 104
<http://www.imbb.forth.gr>

Hungary

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

India

H.A.Nagarajaram
Laboratory of Computational Biology &
Bioinformatics
facility, Centre for DNA Fingerprinting and
Diagnostics
(CDFD), Hyderabad
Email: han@www.cdfd.org.in
Tel: +91 40 7155607 / 7151344 ext:1206
Fax : +9140 7155479
<http://www.in.embnet.org>

Israel

Eitan Rubin
INN (Israeli National Node) Weizmann
Institute of Science
Department of Biological Services, Biologi-

cal Computing Unit, Rehovot
Email: Eitan.Rubin@weizmann.ac.il
Tel: +972- 8-934 3456
<http://www.il.embnet.org>

Italy

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

Mexico

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

The Netherlands

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

Norway

Rune Groven
The Norwegian EMBnet Node
The Biotechnology Centre of Oslo
Email: admin@embnet.uio.no
Tel: +47 22 84 0535
<http://www.no.embnet.org>

Poland

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

Portugal

Pedro Fernandes
Instituto Gulbenkian de Ciencia
Unidade de Bioinformatica
2781-901 OEIRAS
Email: pfern@igc.gulbenkian.pt
Tel: +351 214407912 Fax: +351 2144079070
<http://www.pt.embnet.org>

Russia

Sergei Spirin
Biocomputing Group, Belozersky Institute
Moscow
Email: sas@genebee.msu.su
Tel: +7-095-9328825
<http://www.genebee.msu.su>

Slovakia

Lubos Klucar
Institute of Molecular Biology SAS Bratislava
Email: klucar@embnet.sk
Tel: +421 7 5941 2284
<http://www.sk.embnet.org>

South Africa

Ruediger Braeuning
SANBI, University of the Western Cape,
Bellville
Email: ruediger@sanbi.ac.za
Tel: +27 (0)21 9593645
<http://www.za.embnet.org>

Spain

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

Sweden

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

Switzerland

Laurent Falquet
Swiss Institute of Bioinformatics, CH-1066
Epalinges/Lausanne
Email: Laurent.Falquet@isb-sib.ch
Tel: +41 (21) 692 5954 Fax: +41 (21) 692 5945
<http://www.ch.embnet.org>

United Kingdom

Alan Bleasby
UK MRC HGMP Resource Centre, Hinxton,
Cambridge
Email: ableasby@embnet.org
Tel: +44 (0) 1223 494535
<http://www.uk.embnet.org>

Specialist Nodes

EBI

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

ETI

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

EU

Dr Bernard Mulligan
DG Research - European Commission
Brussels BELGIUM
Email: bernard.mulligan@cec.eu.int

Hoffmann-La Roche

Martin Ebeling
F. Hoffmann-La Roche Ltd, Basel, Switzerland
Email: martin.ebeling@roche.com
Phone: +41-61-68-71398
<http://www.roche.com>

ICGEB

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

LION Bioscience

Peter Rice
LION Bioscience AG, Heidelberg, Germany
Email: Peter.Rice@uk.lionbioscience.com
Phone: +44 1223 224700
<http://www.lionbioscience.com>

MIPS

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

Pharmacia - Biovitrum

Timothy Wood
Biovitrum, Stockholm, Sweden
Email: timothy.wood@eu.pnu.com
Phone: +46 (8) 695 9134
<http://www.pharmacia.se>

Sanger Institute

Michelle Clamp
Wellcome Trust Sanger Institute, Wellcome
Trust Genome
Campus, Hinxton Hall, Hinxton CB10 1SD,
Cambridge, United Kingdom
Email: michele@sanger.ac.uk
Phone: +44-1223-494 967
Fax: +44-1223-494 919
<http://www.sanger.ac.uk>

UMBER

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

TECH-MGR

Email: tech-mgr@embnet.org
The team gives support to EMBnet nodes
and helps them with maintenance and
troubleshooting.
The team is formed of experienced system
administrators and programmers who ensure
the availability of local services for all EMBnet
users.

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
Internet E-mail.

Past issues of embnet.news are available as
PostScript or PDF files (ISSN 1023-4144). You
can get them by anonymous ftp from:
the EMBnet organisation Web site
[http://www.embnet.org/download/embnet-](http://www.embnet.org/download/embnet-news)
news
the Belgian EMBnet node
<ftp://ftp.be.embnet.org/pub/embnet.news>
the UK EMBnet node
<ftp://ftp.uk.embnet.org/pub/embnet.news>
the EBI EMBnet node
<ftp://ftp.ebi.ac.uk/pub/embnet.news>

Submission deadline for next
issues:

June 30, 2003
October 31, 2003

EMBnet.news



Volume 9 Nr. 2
September 2003

**MacOSX and
Bioinformatics (part 2)**
SARS report
Pftools version 2.3
**Pattern discovery in
regulatory sequences**

Editorial

The SARS (Severe Acute Respiratory Syndrome) took us by surprise in March and the whole world was shocked and scared by this emerging threat. Thankfully the scientific community and the political authorities reacted fast enough to circumvent the disease.

The Bioinformatics community also was rapidly involved in the fight with the identification of the virus by BLASTing against the databases of known virus sequences and the creation of dedicated web sites. The Chinese EMBnet node was the first to set up such a web site followed by big members of the field like the NCBI, the EBI, and the Swiss-Prot database.

In this issue Prof. Luo gives us a short report on the topic.

Continuing the MacOSX and Bioinformatics section, we provide our readers with a full explanation on how to install a local Blast server, accessible both from the command-line and the default web server found in every MacOSX. This section also explains the tips and tricks to obtain different interfaces to the EMBOSS package, and the next issue will give an overview of the performance gain from Altivec optimised programs versus other platforms. Many thanks to Erik Bongcam and Anders Nister!

Finally I encourage all readers to have a deep look at two software packages RSAT (regulatory sequence analysis tool) from Dr. van Helden and the long awaited Pftools release 2.3 from Dr. Philipp Bucher and Volker Flegel.

The editorial board:

Erik Bongcam-Rudloff, Laurent Falquet
Pedro Fernandes, Gonçalo Guimaraes Pereira

Protein Spotlight

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

Contents

Editorial	2
Pattern discovery in regulatory sequences	3
Anti-SARS web site	10
Pftools version 2.3	12
MacOSX and Bioinformatics (Part II)	16
Protein spotlight 36	23
National nodes	25
Specialist nodes	27

Publisher:

EMBnet Administration Office
c/o Jack Leunissen
Dept. of Genome Informatics
Wageningen UR, Dreijenlaan 3
6703 HA Wageningen
The Netherlands
Email: Jack.Leunissen@wur.nl
Tel: +31 317 484074

Editorial Board:

Erik Bongcam-Rudloff, The Linnaeus Centre for Bioinformatics, SLU/UU. SE
Email: erik.bongcam@bmc.uu.se
Tel: +46-18-4716696
Fax: +46-18-4714525

Laurent Falquet, SIB, Lausanne. CH
Email: laurent.falquet@isb-sib.ch
Tel: +41-216925954
Fax: +41-216925945

Pedro Fernandes, Instituto Gulbenkian. PT
Email: pfern@igc.gulbenkian.pt
Tel: +315-214407912
Fax: +315-214407970

Gonçalo Guimaraes Pereira, UNICAMP. BR
Email: goncalo@unicamp.br
Tel: +55-19-37886237/6238
Fax: +55-19-37886235

Cover picture: Col de Fenêtre, near Grand-St-Bernard pass, Swiss Alps June 2002 [© Dmitri Kuznetov]

Pattern discovery in regulatory sequences : how to interpret the results ?



Jacques van Helden
Service de Conformation des
Macromolécules Biologiques
et de Bioinformatique.
Université Libre de Bruxelles,
Campus Plaine, CP 263.
Bld du Triomphe.
B-1050 Bruxelles, Belgium.
e-mail: Jacques.van.Helden
@ulb.ac.be

Abstract

In this paper, we illustrate the utilisation of two pattern discovery programs from the Regulatory Sequence Analysis Tools (<http://rsat.ulb.ac.be/rsat/>), for the prediction of regulatory elements in a set of co-expressed genes. The motifs discovered are then compared to known binding site stored in the SCPD database (<http://casiqma.cshl.org/jian>).

Introduction

With the advent of microarray technology, biologists are now able to detect groups of genes characterized by a common response to some environmental (culture medium, drug, ...), or genetical (gene mutation) condition. The question is then to detect the cis-acting elements which mediate this co-expression. Several web services allow to predict regulatory motifs from sets of co-expressed genes:

<http://rsat.ulb.ac.be/rsat/>;
<http://www.lsi.upc.es/~alggen>;
<http://cbcsrv.watson.ibm.com/Tspd.html>;
<http://bighost.area.ba.cnr.it/BIG/PatSearch/>;
<http://bio.cs.washington.edu/software.html>;
<http://bayesweb.wadsworth.org/gibbs/gibbs.html>;
<http://bussemaker.bio.columbia.edu/reduce/>;
<http://www.esat.kuleuven.ac.be/inclusive>;
 We will illustrate here some functionalities of the regulatory Sequence Analysis Tools (<http://rsat.ulb.ac.be/rsat/>) on the basis of a concrete example. Starting from a set of co-regulated genes, we will apply several pattern discovery methods, and check whether the programs are able to return relevant motifs.

Selection of a cluster of co-expressed genes

Expression data about genes responding to

various carbon sources was imported from http://www-genome.stanford.edu/yeast_stress (Gasch et al., 2000). A chip-per-chip normalization was performed (median-centring, quartile-based estimation of dispersion), and genes were filtered with a lower Z-score threshold of 3.6 (corresponding to an E-value of 1). The remaining genes were clustered with Michael Eisen's Cluster software (Eisen et al., 1998) (<http://rana.lbl.gov/>), using average-linkage hierarchical clustering. The resulting tree was displayed with Treeview (part of Eisen's package), and a few groups of co-expressed genes were selected manually. Among them, we selected a group of 58 genes which are repressed by ethanol, and activated by all other carbon sources (Figure 1). This cluster will be used as study case to introduce the concepts of pattern discovery.

Sequence retrieval

The first access to Regulatory Sequence Analysis Tools (RSAT) is usually through the sequence retrieval program. More than 130 organisms are currently supported, and a default upstream size has been defined for each of them, on the basis of prior knowledge about regulation. We selected the organism *Saccharomyces cerevisiae*, entered the 58 gene names selected above, and obtained their upstream sequence over 800bp (the default size for the yeast). In addition

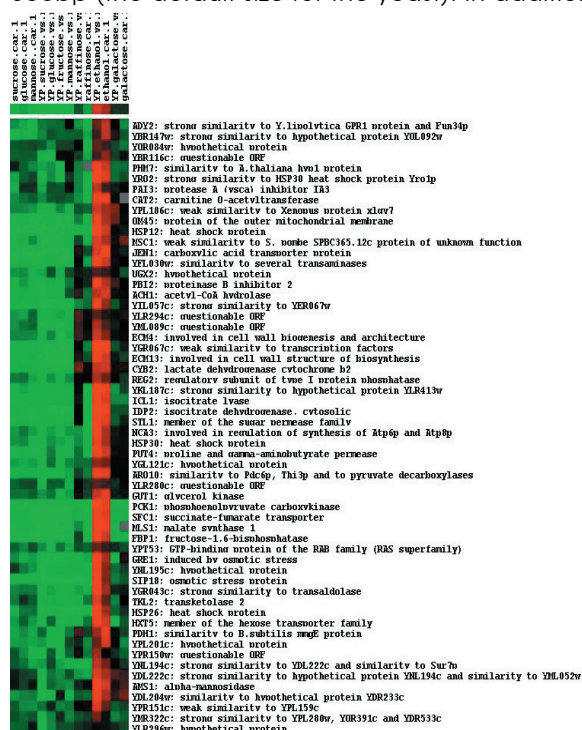


Figure 1: expression profiles of the 58 genes selected with Eisen's Cluster program.

to the default options, for the specific case of *S.cerevisiae*, we allowed overlap with upstream ORFs, because the annotations of *S.cerevisiae* contain around 1,000 false genes, located in intergenic regions, which leads to truncate many upstream sequences erroneously. Upstream ORFs should however be clipped for other organisms, especially for bacteria, where many genes are located inside operons.

Pattern discovery

The sequence retrieval result page contains a list of buttons, which allow to send the sequence to several pattern discovery programs.

oligo-analysis

We will start by applying the simplest approach : oligo-analysis (van Helden et al., 1998). This program counts the number of occurrences of all oligonucleotides of a specified size (hexanucleotides by default), and selects those which are significantly over-represented in the submitted sequences, as compared to the random expectation (i.e. if genes had been selected at random in the genome). Hexanucleotides are regrouped by pairs of reverse complements, because most yeast regulatory elements are strand-insensitive. In the following text, we refer to a pair of reverse complementary hexanucleotides as a pattern. Among the 2080 considered patterns, no more than 13 are significantly over-represented (Figure 2). The most significant hexanucleotide is CCCTTA|TAAGGG, found in 45 occurrences. The column 'expected occurrences' indicates that a random selection of 58 yeast upstream sequences would contain on the average 16.86 occurrences of this pattern. The probability to observe at least 45 occurrences when expecting 16.86 is $1e-08$. However, this probability can be misleading, since each analysis involves simultaneous tests of 2080 patterns. To correct for this, the program returns an E-value (expected value), which indicates how many patterns would be expected at random, with such a level of over-representation. The pattern CCCTTA|TAAGGG has an E-value of $2.1e-05$, which means that such a result would be expected by chance once per 50,000 trials (sequence sets). This pattern is thus highly significant.

Among the 13 selected patterns, some are mutually overlapping, and can be aligned to form larger motifs, as shown at the bottom of the result page (Figure 2). For example, the hexanucleotides gCGGCT and CGGCTa form a heptanucleotide gCGGCTa.

dyad-analysis

A more exhaustive analysis is performed by the program dyad-analysis (van Helden et al., 2000). This is an extension of oligo-analysis, which detects over-represented dyads. A dyad is defined here as a pair of short oligonucleotides (e.g. trinucleotides) separated by a spacing of fixed width but variable content. By default, all spacing values between 0 and 20 are considered. Among the 43,680 possible dyads, 35 are considered as significantly over-represented in the upstream sequences of the 58 ethanol-repressed genes (Figure 3). Several of these dyads are mutually overlapping, and can be assembled to form 7 larger patterns.

The most significant dyad, CCCN{0}TTA|TAAN{0}GGG, is actually identical to the first hexanucleotide detected by oligo-analysis. This is not surprising: dyad-analysis somehow encompasses oligo-analysis, since a hexanucleotide is nothing else than a pair of trinucleotides with a spacing of 0. On the contrary, the second dyad, CCGN{5}CCG|CGGN{5}CGG, could not be detected by oligo-analysis, due to the 5 nucleotide spacing between the two conserved parts. This dyad is found in 21 occurrences, whereas 4.51 would be expected at random. The E-value of $7.1e-04$ indicates a high significance level.

Displaying pattern matches

The patterns detected by oligo-analysis or dyad-analysis can be searched in the input sequences, and their positions displayed on a feature-map (Figure 4). Each pattern is represented as a coloured box, with a thickness proportional to its significance.

Matching discovered patterns against SCPD

One would of course like to know which factors are likely to bind the discovered patterns. A very valuable resource for the yeast community is the *Saccharomyces Cerevisiae* Promoter Database (SCPD; <http://cgsigma.cshl.org/jian>).

The most significant motif, CCCTTA, matches two entries in SCPD : an UASCAR element found upstream the gene CAR2, and an element called 'heat_shock_not_HSE' in DDR2. These two factors are documented by a little number of entries in the database, and it is hard to define a consistent consensus for them. The matching of the discovered pattern with these two elements might thus reflect the binding of one of them, but could also be spurious.

A more interesting result is obtained with the motif GGAN{5}GGA|TCCN{5}TCC, which matches 6

RSA-tools - oligo-analysis result

Information

One or several button(s) will appear at the bottom of this page, allowing you to send the result as input for a subsequent query.

Result

```
; oligo-analysis -sort -format fasta -tho 1 -thosig 0 -return occ,rank,proba -2str -noov -v -seqtype dna -l 6 -bg upstream
-org Saccharomyces_cerevisiae -pseudo 0.05
; Citation: van Helden et al. (1998). J Mol Biol 281(5), 827-42.
; Detection of over-represented words
; Oligomer length      6
; Input format         fasta
; Discard overlapping matches
; Counted on both strands
; grouped by pairs of reverse complements
; Background model     upstream
; Organism              Saccharomyces_cerevisiae
; Method                Frequency file
; Expected frequency file
data/genomes/Saccharomyces_cerevisiae/oligo-frequencies/6nt_upstream_Saccharomyces_
cerevisiae-lstr.freq
; Pseudo weight        0.05
; Pseudo frequency     2.40384615384615e-05
; Sequence type        DNA
; Nb of sequences      58
; Sum of sequence lengths 45097
; discarded occurrences 0 (contain other letters than ACGT)
; nb possible positions 44807
; total oligo occurrences 44807
; total overlapping occurrences 728
; total non overlapping occ 44079
; alphabet size        4
; nb possible oligomers 2080
; threshold on occurrences 1
; threshold on occ sig 0
; threshold on occ proba 0.000480769230769231
; Sequences:
; ADY2_0_800          800
; (...)
; YLR296w_653_147 147
;
; column headers
; 1 seq oligomer sequence
; 2 identifier oligomer identifier
; 3 expected_freq expected relative frequency
; 4 occ observed occurrences
; 5 exp_occ expected occurrences
; 6 occ_P occurrence probability (binomial)
; 7 occ_E E-value for occurrences (binomial)
; 8 occ_sig occurrence significance (binomial)
; 9 ovl_occ number of overlapping occurrences (discarded from the count)
; 10 rank rank
seq identifier expected_freq occ exp_occ occ_P occ_E occ_sig ovl_occ rank
ccctta ccctta|taagg 0.0003781427313 45 16.86 1e-08 2.1e-05 4.68 0 1
ggggta ggggta|tacc 0.0002019404467 27 9.02 1e-06 2.1e-03 2.68 0 2
ccgaaa ccgaaa|tttcg 0.0004096816183 42 18.27 1.4e-06 2.9e-03 2.54 0 3
ccgcgc ccgcgc|gcgcg 0.0001371629722 21 6.13 2e-06 4.2e-03 2.38 2 4
aggggg aggggg|ccccc 0.0001821106075 24 8.14 4.9e-06 1.0e-02 1.99 0 5
agccgc agccgc|gcgct 0.0002587859854 28 11.56 3e-05 6.2e-02 1.21 0 6
cccgcc ccggcg|cgccg 0.0001163888551 17 5.21 3.2e-05 6.8e-02 1.17 0 7
acaagg acaagg|cctgt 0.0005099639475 43 22.74 9.8e-05 2.0e-01 0.69 0 8
cggcta cggcta|tagcg 0.0002219591414 24 9.92 0.00011 2.2e-01 0.65 0 9
aaggga aaggga|tccct 0.0006546273451 51 29.17 0.00016 3.3e-01 0.49 0 10
gatccc gatccc|gggatc 0.0001889094095 21 8.44 0.00019 4.0e-01 0.40 1 11
agggga agggga|tccct 0.0003209194813 30 14.33 0.00020 4.2e-01 0.38 0 12
aagggg aagggg|cccct 0.0003630342824 32 16.21 0.00034 7.1e-01 0.15 0 13
catccc catccc|gggatg 0.0002567085738 25 11.47 0.00036 7.6e-01 0.12 0 14
cgggcc cgggcc|ggccg 0.0001305530259 16 5.84 0.00038 7.9e-01 0.10 2 15
aacaag aacaag|cttgt 0.0009889017758 68 43.97 0.00046 9.7e-01 0.02 0 16
; Job started 12/07/03 22:50:37 CEST
; Job done 12/07/03 22:51:06 CEST
```

Figure 2: pattern discovery result of oligo-analysis

sites in SCPD, 4 of which are bound by the Carbon Source Response Element (CSRE). Another discovered motif, GGGGTA|TACCCC matches two binding sites of Mig1p, the glucose repressor.

Checking the rate of false positive

An important advantage of the methods

presented above is that they return a very low rate of false positive. Indeed, over-representation is assessed by applying a statistical test on each considered pattern independently. Thus, in principle, when a sequence set does not contain any over-represented pattern, oligo-analysis and dyad-analysis are able to return a negative answer. In theory, a pattern with an E-value ≤ 1 is expected once per trial (random sequence set),

```

Pattern assembly
; pattern-assembly -v 1 -subst 1 -2str -i public_html/tmp/
oligo-analysis.2003_07_12.225035.res
; Input file      public_html/tmp/oligo-analysis.2003_07_
12.225035.res
; score column 8
; two strand assembly
; max flanking bases      1
; max substitutions      1
; number of patterns      16
;

;cluster # 1  seed: ccctta  4 words      length
;align      rev_cpl      score
tccctt.      .aaggga      0.49
tccctt.      .agggga      0.38
cccctt.      .aaggga      0.15
.cccctta      taaggga      4.68
tccctta      taaggga      4.68      best consensus

;cluster # 2  seed: ccgcgc  2 words      length 7
;align      rev_cpl      score
ccgcgc.      .cgcgga      1.17
.ccgcgcc      gcgcgc.      2.38
cccgcgc      gcgcgc.      2.38      best consensus

;cluster # 3  seed: aggggg  3 words      length 7
;align      rev_cpl      score
aggggg      cccctt      1.99
agggga      tccctt      0.38
aaggga      cccctt      0.15
aggggg      cccctt      1.99      best consensus

;cluster # 4  seed: agccgc  2 words      length 6
;align      rev_cpl      score
tagccg.      .cggcta      0.65
.agccgc      gcggct.      1.21
tagccgc      gcggcta      1.21      best consensus

;cluster # 5  seed: acaagg  2 words      length 7
;align      rev_cpl      score
aacaag.      .cttggt      0.02
.acaagg      ctttgt.      0.69
aacaagg      ctttgtt      0.69      best consensus

;cluster # 6  seed: gatccc  2 words      length 7
;align      rev_cpl      score
gatccc      gggatc      0.40
catccc      ggcatg      0.12
gatccc      gggatc      0.40      best consensus

; Isolated patterns: 3
;align      rev_cpl      score
ggggta      taccct      2.68      isol
ccgaaa      ttccgg      2.54      isol
cgggcc      ggcccg      0.10      isol

;Job started 12/07/03 22:51:07 CEST
;Job done    12/07/03 22:51:08 CEST

```

Figure 2: (end)

and a pattern with E-value $\leq 1e-3$ every 1,000 trials.

This theoretical model can be challenged either with randomly generated sequences, or with a random selection of real upstream sequences from the organism of interest. Both facilities are provided on the RSAT web site. The program random-seq generates random sequences, and allows to specify different residue probabilities (e.g. to generate AT-rich sequences). More biological-like sequences can be generated with Markov chain models, which mimic not only the residue frequencies, but also the oligonucleotide composition of biological sequences. The program allows for example to generate a random sequences with hexanucleotide

frequencies calibrated on the full set of yeast intergenic sequences. Randomly generated sequences are useful for testing the accuracy of a program, but they rely on some predefined probabilistic model.

A more realistic test can be performed by submitting a random selection of genes from the chosen organism. The program random-genes allows to select a set of random genes, which can then be sent to sequence retrieval and pattern discovery programs.

Discussion

The two approaches presented here are based on simple statistics (binomial test) on simple motifs (non-degenerated hexanucleotides, or spaced pairs of trinucleotides). Despite their simplicity, these approaches present several advantages over other pattern discovery tools, namely exhaustivity and low rate of false positive.

The programs oligo-analysis and dyad-analysis are exhaustive in the sense that, within the class of considered patterns, all possibilities are tested, and each significant result is returned. A single analysis is thus able to return binding sites for several transcription factors. In our test example, we identified putative binding sites for the carbon source response element, and for Mig1p, both of which are consistent with the conditions of expression (ethanol-repressed genes).

The low rate of false positive comes from the fact that the detection methods rely on a statistical test, which allows to select a threshold of significance. This is particularly important for the interpretation of gene expression data, since clustering methods are generally tricky, and the resulting clusters may often be artifactual. A program which would systematically return the 'best' motifs, without assessing whether they are or not significant, would thus predict one or several patterns in each of these artifactual clusters, whereas oligo-analysis and dyad-analysis will generally return no prediction.

Finally, it is important to realize that these tools (as most other pattern discovery tools) have until now been tested and used for microbial organisms. Their extension to higher organisms is far from trivial. Our preliminary tests indicate that many more false positives are returned for human than for yeast or bacterial sequences. This probably comes from the heterogeneity of human chromosomes, the presence of repeated elements, and the larger extensions of the sequences to analyse. We are currently developing more elaborate statistical models in order to address the complexity of transcriptional regulation in higher organisms.

RSA-tools - dyad-analysis result

Information

One or several button(s) will appear at the bottom of this page, allowing you to send the result as input for a subsequent query.

Result

```
; dyad-analysis -v 1 -sort -return proba,rank -timeout 3600 -format fasta -type any -2str -noov -org Saccharomyces_
cerevisiae -thosig 0 -l 3 -spacing 0-20 -bg upstream
; Citation: van Helden et al. (2000). Nucleic Acids Res. 28(8):1808-18.
; Sequence type          DNA
; Nb of sequences        58
; Sum of sequence lengths 45097
; default return values   proba,occ
; return values           proba,occ,exp_freq,exp_occ,rank
; Monad parameters
;     monad size          3
;     monad positions     89962
;     valid               89962
;     discarded           0 (contain other letters than ACGT)
;     distinct monads    64
; Dyad parameters
;     dyad type           any dyad
;     minimal spacing     0
;     maximal spacing     20
;     dyad positions      89614
;     valid               89614
;     discarded           0 (contain other letters than ACGT)
;     distinct dyads     43680
; Threshold on sig       0
; Estimation of expected dyad frequencies
;     Background model
;     organism            Saccharomyces cerevisiae
;     sequence type       upstream
;     exp. freq. file     data/genomes/Saccharomyces_cerevisiae/oligo-frequencies/dyads_3nt_sp0-20_
upstream_Saccharomyces_cerevisiae-1str.freq
; Sequences:
;     ADY2_0_800         800
; (...)
;     YLR296w_653_147 147
; column headers
; 1      dyad_sequence   dyad sequence
; 2      dyad_identifier dyad identifier
; 3      expected_freq
; 4      obs_occ         observed occurrences
; 5      exp_occ         expected occurrences
; 6      occ_P           occurrence probability (binomial)
; 7      occ_E           E-value for occurrences (binomial)
; 8      occ_sig         occurrence significance (binomial)
; 9      ovl_occ         number of overlapping occurrences
; 10     all_occ         number of non-overlapping + overlapping occurrences
; 11     rank            rank
; 12     ov_coef         overlap coefficient
; 13     remark         remark
dyad_sequence dyad_identifier expected_freq obs_occ exp_occ occ_P occ_E occ_sig ovl_occ all_occ rank ov_coef remark
cccn{0}tta cccn{0}tta taan{0}ggg 0.0001785182508 45 15.80 2.2e-09 9.5e-05 4.02 0 45 1 1.0039
ccgn{5}ccg ccgn{5}ccg ccgn{5}cgg 0.0000506136859 21 4.51 1.6e-08 7.1e-04 3.15 5 26 2 1.0195 dir_rep
ggan{5}gga ggan{5}gga tccn{5}tcc 0.0001508407873 38 13.37 3.7e-08 1.6e-03 2.79 1 39 3 1.0156 dir_rep
tccn{6}cca tccn{6}cca tggg{6}gga 0.0001574418937 39 13.96 3.8e-08 1.7e-03 2.78 1 40 4 1.0000
gggn{0}gta gggn{0}gta tacn{0}ccc 0.0000892591254 27 7.94 1e-07 4.5e-03 2.35 0 27 5 1.0039
ctcn{10}ccc cccn{10}ccc gggn{10}gag 0.0000974431639 28 8.66 1.7e-07 7.4e-03 2.13 0 28 6 1.0010
ccgn{0}cgc ccgn{0}cgc gcn{0}cgg 0.0000592410231 21 5.28 2.1e-07 9.3e-03 2.03 2 23 7 1.0674
ccgn{6}cgc ccgn{6}cgc gcn{6}cgg 0.0000502772460 19 4.48 3.1e-07 1.3e-02 1.87 1 20 8 1.0674
cccn{8}gtc cccn{8}gtc gacn{8}ggg 0.0000716916704 23 6.38 3.2e-07 1.4e-02 1.85 0 23 9 1.0010
aggn{0}ggg aggn{0}ggg cccn{0}cct 0.0000813073102 24 7.24 7.8e-07 3.4e-02 1.47 0 24 10 1.0000
cctn{4}ccg cctn{4}ccg ccgn{4}agg 0.0000759244097 23 6.76 8.5e-07 3.7e-02 1.43 0 23 11 1.0039
cgcn{14}ccg cgcn{14}ccg ccgn{14}gcy 0.0000613152051 20 5.46 1.4e-06 6.2e-02 1.20 1 21 12 1.0078
cggn{4}ccg cggn{4}ccg ccgn{4}ccg 0.0000461540490 24 4.11 1.9e-06 8.4e-02 1.08 2 26 13 2.0078 inv_rep
ccgn{5}tgc ccgn{5}tgc gcan{5}cgg 0.0000858232066 24 7.64 1.9e-06 8.5e-02 1.07 0 24 14 1.0049
ccgn{0}aaa cccn{0}aaa tttt{0}cgg 0.0002077411716 42 18.40 2.2e-06 9.6e-02 1.02 0 42 15 1.0000
cccn{1}gcy cccn{1}gcy ccgn{1}ggg 0.0000632965815 20 5.64 2.3e-06 1.0e-01 1.00 1 21 16 1.0664
gggn{5}gga gggn{5}gga tccn{5}ccc 0.0000934252622 25 8.31 2.6e-06 1.1e-01 0.95 2 27 17 1.0000
cccn{0}gcy cccn{0}gcy ccgn{0}ggg 0.0000475120957 17 4.24 2.6e-06 1.1e-01 0.95 0 17 18 1.0664
ccgn{17}tgc ccgn{17}tgc gcan{17}cgg 0.0000771935983 22 6.88 3.8e-06 1.6e-01 0.78 0 22 19 1.0049
cggn{6}ggc cggn{6}ggc gcn{6}ccg 0.0000602926336 19 5.37 4.2e-06 1.8e-01 0.74 2 21 20 1.0049
cggn{7}gcc cggn{7}gcc gcn{7}ccg 0.0000738068363 21 6.58 6.3e-06 2.8e-01 0.56 3 24 21 1.0635
ccgn{4}ccc ccgn{4}ccc gggn{4}cgg 0.0000567435062 18 5.06 6.8e-06 3.0e-01 0.53 1 19 22 1.0049
ccgn{3}gtc ccgn{3}gtc gacn{3}cgg 0.0000522817596 17 4.66 8.8e-06 3.8e-01 0.42 0 17 23 1.0010
cgcn{4}cgc cgcn{4}cgc ccgn{4}gcy 0.0000523478824 17 4.67 8.9e-06 3.9e-01 0.41 0 17 24 1.0078
atan{9}ggg atan{9}ggg cccn{9}tat 0.0001628688074 34 14.46 1e-05 4.4e-01 0.36 0 34 25 1.0000
cggn{6}gga cggn{6}gga tccn{6}ccg 0.0000829274097 22 7.39 1.1e-05 4.9e-01 0.31 1 23 26 1.0000
atan{1}ggg atan{1}ggg cccn{1}tat 0.0001576443162 33 14.00 1.3e-05 5.6e-01 0.26 0 33 27 1.0000
cccn{5}aag cccn{5}aag cctt{5}ggg 0.0001382373794 30 12.28 1.6e-05 7.0e-01 0.16 1 31 28 1.0000
agcn{0}cgc agcn{0}cgc gcn{0}gct 0.0001248434983 28 11.10 1.7e-05 7.4e-01 0.13 0 28 29 1.0039
cgcn{4}tcc cgcn{4}tcc gggn{4}cgg 0.0000791212269 21 7.05 1.7e-05 7.6e-01 0.12 1 22 30 1.0049
cgcn{5}gcc cgcn{5}gcc ggcn{5}cgc 0.0000670181217 19 5.97 1.8e-05 7.8e-01 0.11 2 21 31 1.0625
atgn{18}ggg atgn{18}ggg cccn{18}cat 0.0001124816378 26 10.01 2e-05 8.8e-01 0.06 0 26 32 1.0000
cccn{3}aag cccn{3}aag cctt{3}ggg 0.0001400831879 30 12.45 2e-05 8.8e-01 0.05 0 30 33 1.0000
cgcn{12}gtc cgcn{12}gtc gacn{12}gcy 0.0000559105979 17 4.99 2e-05 8.9e-01 0.05 0 17 34 1.0010
cccn{1}taa cccn{1}taa ttan{1}ggg 0.0001548576743 32 13.75 2.2e-05 9.5e-01 0.02 0 32 35 1.0000
;Job started 12/07/03 23:06:56 CEST
;Job done 12/07/03 23:09:31 CEST
```

Pattern assembly

```

; pattern-assembly -v 1 -subst 0 -2str -maxfl 1 -subst 0 -i public_html/tmp/dyad-analysis.2003_07_12.230654.res
; Input file      public_html/tmp/dyad-analysis.2003_07_12.230654.res
; score column 8
; two strand assembly
; max flanking bases      1
; max substitutions      0
; number of patterns      35
;

;cluster # 1          seed: ccctta          2 words          length
;align               rev_cpl          score
ccctta.              .taaggg          4.02
cccntat              atanggg          0.26
cccttat              ataaggg          4.02          best consensus

;cluster # 2          seed: cggnnnnnccg      9 words          length 7
;align               rev_cpl          score
ggcnnnnnnnccg...    ...cggnnnnnnngcc      0.56
.gccnnnnnnccg...    ...cggnnnnnnnggc      0.74
..ccgnnnnnnccg...   ..cggnnnnnnccgg..    3.15
..ccgnnnnnnccgc...  ..gcnnnnnnnccgg..    1.87
..ccgnnnnncc...     ...gggnnnnccg...     0.53
..cggnnnnccg...     ...cggnnnnccg...     1.08
.....ccnccg...      .cgcnngg.....        1.00
.....cccgcg...      .cgcggg.....        0.95
.....cccgcg...      .cgcggg.....        2.03
ggccggnncccgcgcc    ggcggggggnccggcc     3.15          best consensus

;cluster # 3          seed: ggannnnngga      4 words          length 16
;align               rev_cpl          score
tggnnnnnngga        tcnnnnnnncca          2.78
.ggannnnngga        tcnnnnntcc..          2.79
..ggannnnccg..      .ccgnnntcc..          0.12
..gacnnncgg..       .ccgnnngtc..          0.42
tggaennncgga        tcggnngtcca          2.79          best consensus

;cluster # 4          seed: gggnnnnngga      3 words          length 12
;align               rev_cpl          score
tggnnnnnngga        tcnnnnnnncca          2.78
.gggnnnnngga        tcnnnnnccc..          0.95
.gggnnnnccg..       .ccgnnnnccc..          0.53
tgggnnnnccgga       tcggnnncccca          2.78          best consensus

;cluster # 5          seed: cgcnnnnccg      8 words          length 12
;align               rev_cpl          score
ggcnnnnnnnccg..     ..cggnnnnnnngcc      0.56
.gccnnnnnnccg..     ..cggnnnnnnnggc      0.74
..ccgnnnnnnccg..    ..cggnnnnnnccgg..    3.15
..ccgnnnnnnccgc...  ..gcnnnnnnnccgg..    1.87
..ccgnnnnncc...     ...gggnnnnccg...     0.53
..cgcnnnnccg...     ..cggnnnnngcg...     0.41
.....cccnccg...      .cgcnngg.....        1.00
.....cccgcg...      .cgcggg.....        0.95
ggccgcnncccgcg      cgcggggngcgcc        3.15          best consensus

;cluster # 6          seed: cggnnnnnnngga      7 words          length 15
;align               rev_cpl          score
gcgnnnnnnccg..      .ccgnnnnnnccgc      1.87
.cggnnnnnnccg..     .ccgnnnnnnccg..      3.15
.cggnnnnccg..       .cggnnnnccg..        1.08
.cggnnnnnnngga      tcnnnnnnnccg..       0.31
..ggannnnngga       tcnnnnntcc..         2.79
..ggannnnccg..      .ccgnnntcc..         0.12
..gacnnncgg..       .ccgnnngtc..         0.42
gcggacnnccgga       tcggnngtccg          3.15          best consensus

;cluster # 7          seed: ccntaa          2 words          length 13
;align               rev_cpl          score
ccctta.              .taaggg          4.02
cccntaa              ttanggg          0.02
cccttaa              ttaaggg          4.02          best consensus

; Isolated patterns: 16
;align               rev_cpl          score
ggggta               tacccc          2.35          isol
ctcnnnnnnnnnncccc   gggnnnnnnnnnnngag    2.13          isol
cccnnnnnnnngtc      gacnnnnnnnnnggg      1.85          isol
aggggg              cccccc          1.47          isol
cctnnnnccg          cggnnnnnagg         1.43          isol
cgcnnnnnnnnnnnnnccg cggnnnnnnnnnnnnngcg  1.20          isol
ccgnnnnntgc         gcannnnccg          1.07          isol
ccgaaa              ttccgg           1.02          isol
cgcnnnnnnnnnnnnnntgc gcannnnnnnnnnnnnnccg  0.78          isol
atannnnnnnnnggg     cccnnnnnnnnntat      0.36          isol
cccnnnnnnaag         cttnnnnnggg         0.16          isol
agccgc              gcggct           0.13          isol
gcgnnnnngcc         ggcnnnnnccg         0.11          isol
atgnnnnnnnnnnnnnnnggg cccnnnnnnnnnnnnnnccat 0.06          isol
cgcnnnnnnnnnnnngtc  gacnnnnnnnnnnngcg    0.05          isol
cccnnnnaag          cttnnngg           0.05          isol

;Job started 12/07/03 23:09:31 CEST
;Job done 12/07/03 23:09:44 CEST

```

Figure 3 (pages 7 & 8) Pattern discovery result of dyad-analysis

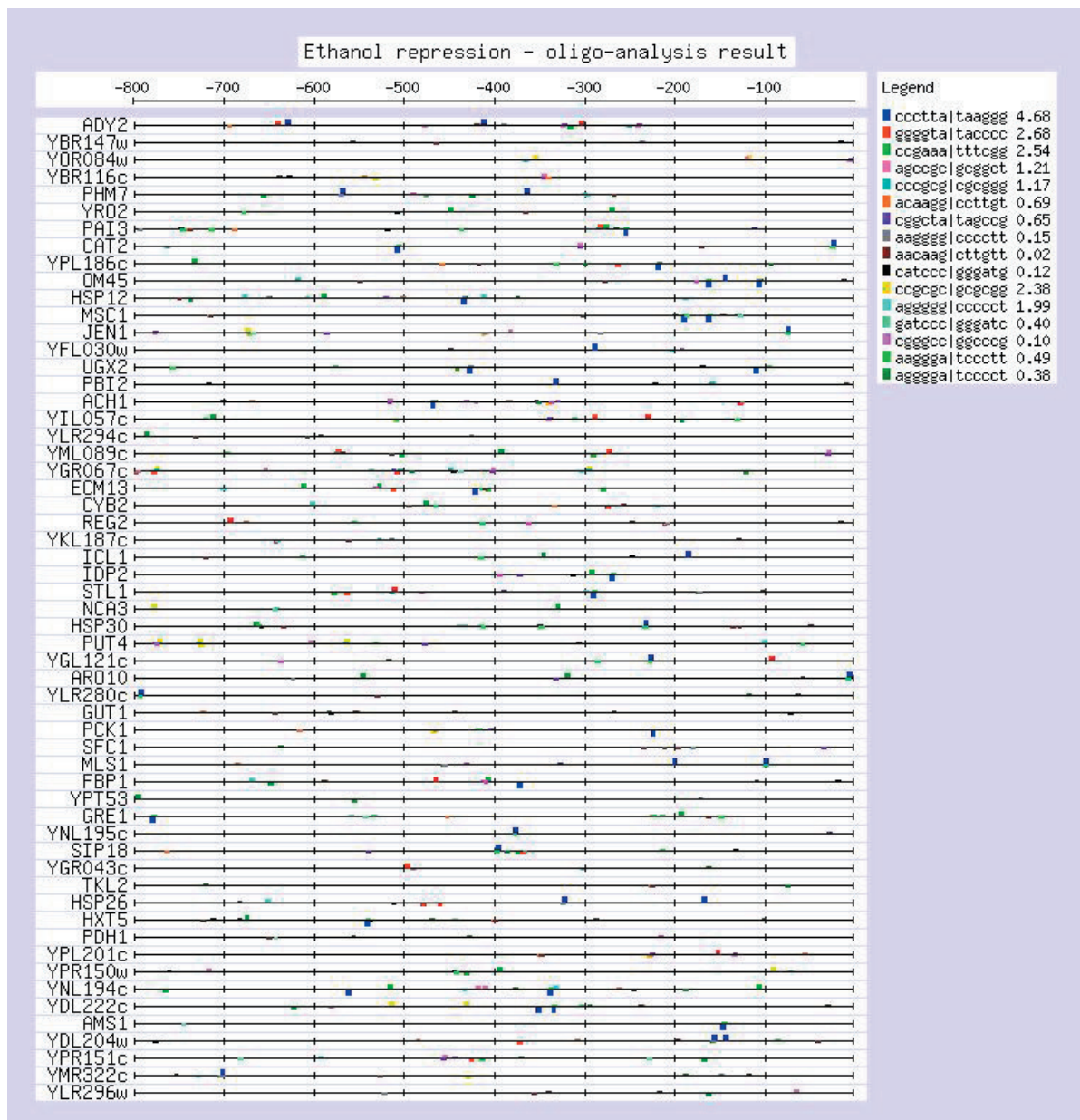


Figure 4: feature-map of the patterns discovered with oligo-analysis

References

1. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95(25), 14863-8.
2. Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D. & Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* 11(12), 4241-57.
3. van Helden, J., André, B. & Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* 281(5), 827-42.
4. van Helden, J., Rios, A. F. & Collado-Vides, J. (2000). Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res* 28(8), 1808-18.

A color portrait photograph of Dr. Y. H. Zhang, a middle-aged man with dark hair, wearing glasses, a light blue shirt, a dark tie, and a grey suit jacket. He is smiling and standing in front of a window with a decorative border.

anti-sars - Microsoft Internet Explorer

文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)

地址(D) http://antisars.cbi.pku.edu.cn/phylogeny.jsp

Virus database
Other databases

Analysis

MultiSeqAlign
Pro2DStruc
Pro3DModel
Phylogeny

Tools

Literature

PubMed
Science free
Free book
NEJM Online

Links

SARS Related
Bioinformatics Resources
Bioinformatics Tools
Online Journals
NCBI SARS Page
EBI SARS Page
TUBIC SARS Page

A description of the method and a figure of phylogenetic tree of 18 coronavirus are listed [here](#).
Submitted by Ji Qi, PhD student of Prof Bailin Hao, Inst of Theoretical Physics, CAS.

Phylogenetic tree of 18 coronaviruses. The tree is rooted at the center and branches out to various virus names. The names are color-coded: blue for MHVM, MHVC, MHV2, MHV, MHVP, BCoVQ, BCoVM, BCoV, BCoVVE, BCovL, IBV, TGE, PEDV, 229E, BCCA, CDCP, BJ01, and CUHK. The tree is divided into four main groups by brackets: G2 (MHVM, MHVC, MHV2, MHV, MHVP), SARS-CoV (BJ01, CUHK, BCCA, CDCP), G3 (BCovQ, BCoVM, BCoV, BCoVVE, BCovL), and G1 (IBV, TGE, PEDV, 229E).

[1] Ji Qi, Bin Wang and Bai-lin Hao, Whole Proteome Prokaryote Phylogeny without Sequence Alignment: A K-String Composition Approach. Accepted by Journal molecular Evolution
[2] Ka Hou Chu, Ji Qi, Zu-Guo Yu and Vo Anh, Origin and Phylogeny of Chloroplasts Revealed by a Simple Correlation Analysis of Complete Genomes. Under revision for Molecular Biology and Evolution
[3] Paul A. Rota, Characterization of a Novel Coronavirus Associated with Severe Acute

Figure 1. Example of the web server, Phylogeny diagram of the Corona viruses

Table of 42 mismatches among 12 complete SARS corona virus genome sequences

No	Pos	Freq	Sequence*												Codon	protein	AA
			CA1	BJ1	US1	HK1	HK2	HK3	SG1	SG2	SG3	SG4	SG5	TW1			
1	1476	1	a	a	a	a	a	a	a	a	a	a	G	a	AGA-AGG	P65-225	R R
2	2601	1	t	t	t	t	C	t	t	t	t	t	t	t	GTT-GTC	P65-600	V V
3	3165	1	a	a	a	a	a	a	a	a	a	a	a	G	TCA-TCG	Nsp1-149	S S
4	7746	1	g	g	g	T	g	g	g	g	g	g	g	g	CCG-CCT	Nsp1-1676	P P
5	7919	1	c	c	T	c	c	c	c	c	c	c	c	c	GCT-GTT	Nsp1-1734	A-V
6	7930	1	g	g	g	g	A	g	g	g	g	g	g	g	GAC-AAC	Nsp1-1738	D-N
7	8387	1	g	g	g	g	C	g	g	g	g	g	g	g	AGT-ACT	Nsp1-1890	S-T
8	8417	1	g	g	g	g	C	g	g	g	g	g	g	g	AGA-ACA	Nsp1-1900	R-T
9	8572	1	g	T	g	g	g	g	g	g	g	g	g	g	GTA-TTA	Nsp1-1952	V-L
10	9404	2	t	C	t	C	t	t	t	t	t	t	t	t	GTT-GCT	Nsp1-2229	V-A
11	9479	1	t	t	t	C	t	t	t	t	t	t	t	t	GTA-GCA	Nsp1-2254	V-A
12	9854	1	c	T	c	c	c	c	c	c	c	c	c	c	GCC-GTC	Nsp1-2379	A-V
13	10587	1	a	C	a	a	a	a	a	a	a	a	a	a	ACA-ACC	3cl-201	T T
14	13494	1	g	g	g	g	A	g	g	g	g	g	g	g	GTT-AGT	RdRp-42	V-S
15	13495	1	t	t	t	t	G	t	t	t	t	t	t	t	GTT-AGT	RdRp-42	V-S
16	16622	1	c	c	T	c	c	c	c	c	c	c	c	c	GCC-GCT	Nsp10-152	A A
17	17564	2	t	G	t	G	t	t	t	t	t	t	t	t	GAT-GAG	Nsp10-466	D-E
18	17846	2	c	c	c	T	c	T	c	c	c	c	c	c	CGC-CGT	Nsp10-560	R R
19	18065	1	g	g	g	g	A	g	g	g	g	g	g	g	AAG-AAA	Nsp11-32	K K
20	18282	1	c	c	c	c	c	c	A	c	c	c	c	c	CTA-ATA	Nsp11-105	L-I
21	18965	1	t	t	t	t	t	t	t	t	t	A	t	t	ATT-ATA	Nsp11-332	I I
22	19064	2	a	a	G	G	a	a	a	a	a	a	a	a	GAA-GAG	Nsp11-365	E E
23	19084	4	c	c	c	c	c	c	c	T	T	T	T	c	ACA-ATA	Nsp11-372	T I
24	19838	1	a	G	a	a	a	a	a	a	a	a	a	a	GTA-GTG	Nsp12-96	V V
25	21721	2	g	A	g	A	g	g	g	g	g	g	g	g	GGC-GAC	Spike-77	G-D
26	22222	2	t	C	t	C	t	t	t	t	t	t	t	t	ATT-ACT	Spike-244	I T
27	23174	1	c	c	c	c	c	c	T	c	c	c	c	c	TCC-TCT	Spike-561	S S
28	23220	1	G	t	t	t	t	t	t	t	t	t	t	t	TCT-GCT	Spike-577	S A
29	23792	1	c	c	c	c	c	c	c	c	c	T	c	c	GTC-GTT	Spike-767	V V
30	24872	1	t	t	C	t	t	t	t	t	t	t	t	t	CTT-CTC	Spike-1127	L L
31	25298	1	A	g		g	g	g	g	g	g	g	g	g	GGA-AGA	Sars3a-11	G-R
32	25569	1	t	t	t	t	A	t	t	t	t	t	t	t	ATG-AAG	Sars3a-101	M-K
33	25673	1	a	C	a	a	a	a	a	a	a	a	a	a	AAG-CAG	Sars3a-136	K Q
34	26050	1	a	C	a	a	a	a	a	a	a	a	a	a	CCA-CCC	Sars3b-261	P P
35	26428	1	g	g	g	g	g	g	g	g	A	g	g	g	GAG-AAG	Membrane-11	E-K
36	26477	1	t	t	t	t	G	t	t	t	t	t	t	t	TTC-TGC	Membrane-27	F-C
37	26600	1	c	c	c	c	T	c	c	c	c	c	c	c	GCT-GTT	Membrane-68	A-V
38	26857	1	t	t	C	t	t	t	t	t	t	t	t	t	TCC-CCC	Membrane-154	S P
39	27111	1	a	a	a	a	a	a	a	a	a	a	G	a	GAG-GGG	Sars6-13	E-G
40	27243	1	c	T	c	c	c	c	c	c	c	c	c	c	CCT-CTT	Sars6-57	P-L
41	27827	2	t	C	t	C	t	t	t	t	t	t	t	t	TGC-CGC	Sars8a-17	C-R
42	28696	1	g	g	g	g	g	T	g	g	g	g	g	g	GGT-TGT	Nucleocapsid-193	G-C
Total			52	2	12	5	9	9	3	2	1	2	3	3	1		

*CA1= TOR2, BJ1=BJ01, US1=Urbani, HK1=CUHK-W1, HK2=HKU-39849, HK3=CUHKSu10, SG1=SIN2679, SG2=SIN2748, SG3=SIN2500, SG4=SIN2774, SG5=SIN2677, TW1-TW01. Colour code for amino acid change: green - same, yellow - similar, red - dissimilar. Colour code for number of changes: dark blue - 4, light blue - 2, light green - 1.

service, we have made bioinformatics analysis on the SARS-CoVs. Together with NCBI and EBI colleagues, we have published a paper «Initial analysis on the genome sequence of the SARS coronavirus» (3).

References

1. Chinese EMBnet node (<http://www.cbi.pku.edu.cn>)

2. Anti-SARS web site (<http://antisars.cbi.pku.edu.cn>)

3. Chen YJ, Gao G, Bao YM, Lopez R, Wu JM, Cai T, Ye ZQ, Gu XC, Luo JC (2003) Initial analysis on the genome sequence of the SARS coronavirus. Acta Genetica Sinica 30(6): 493~500.

Pftools version 2.3



Volker Flegel
 PROSITE group
 Swiss Institute of
 Bioinformatics
 155, ch. des Boveresses
 CH-1066 Epalinges,
 Switzerland
 Volker.Flegel@isb-sib.ch

Introduction

Pairwise sequence comparison has become a fundamental paradigm in biological sequence analysis. Many different computational tools have been developed in order to search for sequences in databases which display significant similarity between each other. Nevertheless, during the course of sequence evolution a great deal of variability has often been introduced in sequences belonging to the same functional or structural family. This renders pairwise sequence comparison inefficient, especially if only some functional residues (*i.e.* active site) are conserved. A simple way to tackle this problem is the use of *position specific scoring matrices* (PSSM). These matrices are derived from multiple sequence alignments and therefore include information about the variability or conservation of all residues at a given position of the alignment. This technique was further improved by the use of probabilistic models like *hidden Markov models* (HMM) and generalised profiles. These biological motif descriptors give a probability (or a score, for the generalised profiles) to each residue at a given position of the motif as well as a position specific probability for insertions and deletions in the alignment. Hence these motif descriptors allow the biologist to push his search further back on the evolutionary time scale, because they are able to find very distantly related members of a sequence family. In this short article we will describe a set of tools which allow the construction and use of generalised profiles as found in the PROSITE database.

Generalised profiles

These motif descriptors combine many aspects of PSSMs and certain types of HMMs. They can be used to search for promoter elements or other motifs on DNA sequences but they are more commonly applied to the identification of protein domains or families. The main repository for protein profiles is the PROSITE database maintained at the Swiss Institute of Bioinformatics.

Generalised profiles (hereafter simply called profiles) are composed of alternating match and insert positions (see figure 1). With each position is associated a set of scores. The match positions give a residue specific match extension score as well as a deletion extension score. When aligning a sequence to a profile, these scores are used when either a sequence residue is considered to match the given profile position or this particular position seems to be deleted in the sequence. The insert positions supply residue specific insert extension scores, as well as 16 values describing all the different transitions from one state to another (*begin*, *match*, *delete*, *insert* or *end*). The two begin and end states present at each insert position allow the profile to function in either a global alignment mode or several kinds of local alignment modes. The insert positions provide scores to residues inserted into the matching sequence as compared to the profile. They can also allow or penalise transitions between the states of the alignment (*i.e.* forbidding insertions at a specific position, by giving a negative score to match to insert transitions). Further information about generalised profiles can be found in Bucher 1994 and Bucher 1996.

For each alignment between a sequence and a profile, a score can be computed by summing all

Simplified profile states and transitions

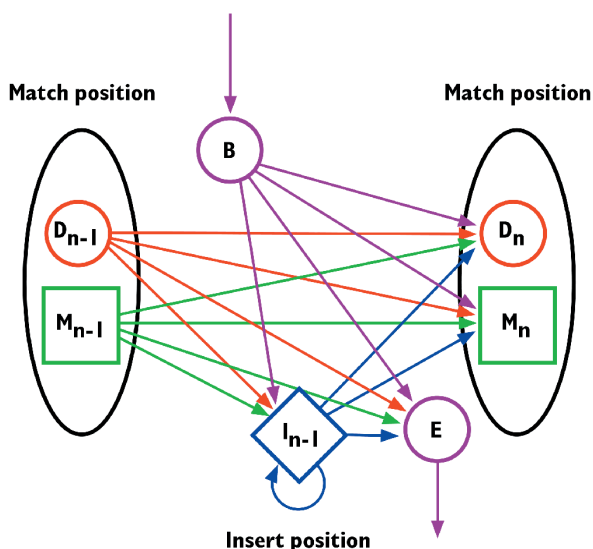


Figure 1. This diagram shows the basic building blocks of a PROSITE profile. The match positions are composed of a match (green) and a delete (red) state. An insert position is found between two successive match positions. The match and insert states provide scores for the complete residue alphabet, whereas only one score exists for the delete state. The transition scores are defined at the insert position. The begin (B) and end (E) states allow internal initiation or termination of the alignment.

the different state scores as well as the different state transitions. The goal being to find the best scoring alignment(s). A set of programs to build and search with generalised profiles is provided by the **pftools** package, which we will describe in the following paragraphs.

Pftools package description

The **pftools** package was originally developed by Philipp Bucher and is currently maintained at the Swiss Institute of Bioinformatics, Lausanne. These tools implement profiles in a machine-readable text file format as used by the PROSITE protein motif database and detailed at <http://www.expasy.org/txt/profile.txt>.

The package contains a set of programs to create, convert and search with profiles. It is distributed as FORTRAN sources which should readily compile on Unix like systems providing a FORTRAN compiler. Several precompiled binary packages will also be made available. The sources can be downloaded via anonymous ftp at: <ftp://ftp.isrec.isb-sib.ch/pub/software/unix/pftools>. A web page providing support and tutorials will be made available as soon as possible at the URL: <http://www.isrec.isb-sib.ch/pftools>.

The distribution contains 12 tools which are briefly described below.

2ft 2-frame interleaved translation of a DNA sequence into protein.
This tool translates both strands of a DNA sequence into protein by interleaving, for each strand, the translated residues of each reading frame. It is used to translate genomic or EST sequences to search against frame-search profiles generated by **ptof** (see below), thus allowing to easily identify coding regions or frame-shifts.

6ft 6-frame translation of a DNA sequence into protein.
As opposed to the **2ft** program, **6ft** does not generate interleaved protein sequences, but simply translates all 6 reading frames of a DNA sequence into protein.

gtop convert a profile in GCG format into PROSITE format.
Profiles created with the GCG program **ProfileMake** can be converted into the PROSITE profile format without data loss.

htop

convert a HMMER ASCII-formatted HMM into an equivalent PROSITE profile.

This tool implements the method described by Bucher 1996 to convert a HMMER HMM motif descriptor into an equivalent PROSITE profile. Both HMMER1 and HMMER2 ASCII HMMs are supported. This conversion is useful to bridge the gap between the Pfam (relying on HMMs) and PROSITE motif databases.

pfmake

generate a profile from a multiple sequence alignment.

This central program of the pftools package reads a weighted multiple sequence alignment in either MSF or MSA format, and computes a PROSITE profile according to the method described in Bucher 1994. Several parameters control the type of profile created (i.e. linear, circular, global-, local-alignment mode, etc.).

pfscale

scale the parameters of a profile's normalisation function.

Given a sorted score distribution obtained by searching a sequence database with a profile, this program computes the parameters of an extreme-value distribution and rescales the profile's normalisation function accordingly.

pfscan

scan a protein or DNA sequence against a profile library.

This is the first of two profile searching tools. It scans an input sequence for the occurrence of any profile stored in a given profile library. Matches are reported if the raw or normalised score exceeds a given cut-off level. The description of the alignment varies according to several command line parameters.

pfsearch

search with a profile against a protein or DNA sequence library.

This second search tool finds the occurrences of an input profile in the sequences of a protein or DNA library. Similar to pfscan, matches are reported if their score exceeds a given cut-off value. The format of the output depends on several command line parameters.

- pfw** weight sequences of a multiple alignment.
This tool computes new weights for individual sequences in a multiple sequence alignment to compensate for the skew in the number of different family members in the original sequence alignment.
- psa2msa** convert a PSA file into a FASTA multiple sequence alignment file.
It reformats a PSA profile-sequence alignment as produced by **pfscan** or **pfsearch** into a Pearson/FASTA-formatted multiple sequence alignment (also called MSA). The output can then be reformatted through third party tools like **readseq** into other formats, or directly used by **pfmake** to create a PROSITE profile.
- ptof** convert a protein profile into a frame-search profile.
The frame-search profile created by **ptof** can be used to search for occurrences of a protein motif in DNA sequences (*i.e.* genomic or EST) previously translated with **2ft**.
- ptoh** convert a PROSITE profile into an approximately equivalent HMM.
This tool uses the reverse technique of **htop** to convert a PROSITE profile into an HMM compatible either with the SAM or HMMER1 package.

The concurrent use of these tools allows the easy creation and search for sequence motifs. They help the biologist to identify biological family members or discover functional elements in sequences. The conversion between different kinds of motif descriptors opens a whole range of possibilities to explore the structure of the biological sequence under investigation.

In the following paragraphs we will focus on the changes introduced between releases 2.2 and 2.3 of the **pftools** package.

Release 2.3 generalities

Some improvements, not necessarily visible to the user, affect the handling of input files. All programs included in the **pftools** package are now able to handle larger sequences or profiles compared to the previous release. Indeed, with the advent of genomic sequencing, the size of sequences submitted to these tools has increased during recent years.

An effort was made to catch and report to the user parse errors occurring on input files or problems that may arise during computation. Many of these errors would terminate the programs silently in release 2.2, leaving the user in doubt about the correct completion of the task. These error or warning messages are output on standard error and can be easily redirected or discarded.

The command line syntax has been rendered more Unix friendly. All options are now specified via a hyphen (-) followed by a letter and possibly a parameter value, as is common in the Unix world. The syntax used in release 2.2 is still supported but deprecated; it may not be so for future releases of the **pftools** package.

Programs having as output alignments, profiles or sequences now include an option to change the width of the output to adapt it to the users needs, *i.e.* for sequences the number of residues to display on one line can be set through the command line option **-W**.

Due to specifics of the formatted output in FORTRAN, numerical overflow could result in the display of a series of '*' characters in place of the expected value. This was corrected wherever possible in release 2.3 of the **pftools**.

In addition we also corrected the known bugs of the different programs, which nevertheless did not lead to major changes in the algorithms used by the **pftools**.

We also introduced a controlled vocabulary, implemented as *keyword=value* pairs, to describe alignments while remaining compatible with the standard FASTA sequence header. This header syntax is called **xpsa** and will be detailed below (see also the man page included in the **pftools** distribution). The need for this kind of syntax arose when transferring information between and through different sequence analysis tools, most of them relying on sequences in FASTA format. Further developments around the **pftools** will rely more heavily on this kind of syntax. In the current release, **xpsa** is essentially used for output, but a tool like **pfmake** is able to read the weight of input sequences through the use of the *'weight=<value>'* keyword.

In the next paragraphs we describe some of the more commonly used tools of the package and the major changes between the release 2.3 and 2.2.

Creating a profile

Several tools are useful when generating profiles from multiple sequence alignments. Here we briefly describe some changes found in **psa2msa**, **pfw**, **pfmake** and **pfscale**.

psa2msa is used to convert alignments produced

```

>PEX7_HUMAN|O00628/63-323 motif=PS50294|WD_REPEATS_REGION
norm_score=38.968 raw_score=1376 match_type=region seq_end=-1
SFDWNDGLFDVTWSENnEHVLITCSGDGSLQLWD TAKAAGPLQvYKEHAQEVYSDWSQT
rgEQLVVSGSWDQTVKLDWPTVGKSLCTFRGHESIIYSTIWSPHiPGCFASASGDQTLRI
WDVKAAGVRIVIPAHQAEILSCDWCKYnENLLVTGAVDCSLRGWDLRNVRQPVFeLLGHT
YAIRRVKFSPPfHASVLASCSYDFTVRFWNFSKPDSSLEtVEHHTeFTCGLDFSLQsPTQV
ADCSWDETIKIYDPACLTIPA-
>PEX7_HUMAN|O00628/63-105 motif=PS50294|WD_REPEATS_REGION
norm_score=7.437 raw_score=180 repeat_nb=1 match_type=repeat
seq_end=-219
SFDWNDGLFDVTWSENnEHVLITCSGDGSLQLWD TAKAAGPLQ
[...]
>PEX7_HUMAN|O00628/282-323 motif=PS50294|WD_REPEATS_REGION
norm_score=4.353 raw_score=63 repeat_nb=6 match_type=repeat
seq_end=-1
VEHHTeFTCGLDFSLQsPTQVADCSWDETIKIYDPACLTIPA-

```

Figure 2. This example shows the alignment between a circular profile and a protein sequence. Elements of the xpsa header are highlighted in blue. The first alignment is the total match of the circular profile on the sequence (as identified by the 'match_type=region' element). This alignment can be decomposed into 6 individual matches listed below the total match (only two have been included in this figure). They are characterised by the 'match_type=repeat' and the 'repeat_nb' elements.

by either **pfscan** or **pfsearch** (launched with option **-x**) into a Pearson/FASTA formatted multiple sequence alignment (MSA). It also allows to control the length of the resulting alignment by truncating insertions to a certain size. The behaviour of this option was changed in the new release: a value of '0' now removes all insert positions and the default value of '-1' imposes no limit on the length of insert regions.

The programs **pfw** and **pfmake** can now directly use these MSA files as input without requiring another conversion step to obtain an alignment in MSF format. The sequence weighting tool **pfw** will output the weighted multiple alignment in the same format as the input file. When the MSA format is used, the weight of each individual sequence is specified in the FASTA header using the 'weight=<value>' keyword. This information can then be recovered by **pfmake** for profile generation.

When scaling the normalisation function of a profile by using **pfscale**, it is now possible to specify which normalisation mode is to be scaled (and implicitly all the cut-off level scores using this normalisation mode will be updated as well). Note that the current version of **pfscale** does not support input score lists in **xpsa** format.

Searching with profiles

The two search tools **pfsearch** and **pfscan** have very similar behaviours, the first searches with a profile against a sequence library whereas the second searches a sequence against a profile library. Therefore the new features have been

similarly implemented in both programs.

Alignments are now printed correctly when option **-a** (report optimal alignment for all sequences regardless of the cut-off value) or **-u** (forces DISJOINT=UNIQUE) is combined with any of the **-s**, **-x** or **-y** output modifier options.

If a profile contains several normalisation modes it is possible to specify via the command line option **-M** which mode should be used for alignment score computation. This also overrides the profile's PRIORITY parameter which can be used to give a precedence for one normalisation mode over another.

The option **-k** enables the output of alignments with a **xpsa** header. As already mentioned, this is a more structured variation of the FASTA sequence header using *keyword=value* pairs to store information. It's syntax is detailed in the man page distributed with the pftools package. Briefly, the header must start with a '>' immediately followed by a sequence identifier. Alignment begin and end positions can be appended to this field using the '/' character. Then follows a list of *keyword=value* pairs which describe the raw score of the alignment, the normalised score, the motif identifier, etc. For examples refer to figure 2. This also allows to have a similar header structure when using **pfsearch** or **pfscan**, which is not the case when using the standard output format. Indeed **pfsearch** will report first the sequence identifier whereas **pfscan** uses the profile's identifier instead.

If a profile is circular, each match between a sequence and this profile can be composed

of repeats of individual matches. Release 2.3 of **pfsearch** and **pfscan**, can now report the individual matches for such circular profiles (see figure 2). Note that the scoring system of a circular profile was optimised to find total matches, therefore the normalised scores of the individual matches should be considered with caution.

Conclusion and future developments

The release 2.3 of the **pftools** package corrects many known bugs of the previous release. The tools should be less error prone due to an improved parsing of input files and buffer size tests. No new algorithm was developed for this release but many options have been added to the different programs in order to increase their flexibility and interoperability. This includes the creation of the **xpsa** header format, designed to allow the flow of information between different sequence analysis tools.

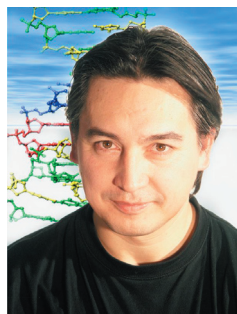
In order to allow easy implementation and testing of new algorithms, the main focus for future developments of the **pftools** will be to port the FORTRAN sources to C/C++. Emphasis will be set on developing a modular, library based design, thus enabling profile parsing and alignment functions to be easily used in other programs written in different languages. Several new sequence weighting, profile construction and scaling algorithms are currently tested and will eventually be included in future releases.

Further information can be obtained by mail at: pftools@isb-sib.ch

References

1. Bucher, P. and Bairoch, A. (1994) A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. *proc Int Conf Intell Syst Mol Biol*, 2, 53-61.
2. Bucher, P., Karplus, K., Moeri, N. and Hofmann, K. (1996) A flexible motif search technique based on generalized profiles. *Comput Chem*, 20, 3-23.
3. Luthy, R., Xenarios, I. and Bucher, P. (1994) Improving the sensitivity of the sequence profile method. *Protein Sci*, 3, 139-46.

A guide on how to install Bioinformatics tools on MacOSX (Part II)



Erik Bongcam-Rudloff

Assistant Professor
Swedish University of Agricultural Sciences
The Linnaeus Centre for Bioinformatics



Anders Nister

Project Student
Swedish University of Agricultural Sciences
The Linnaeus Centre for Bioinformatics

This is the second part in our tutorial series aimed at building a "Bioinformatics Workbench" on the MacOSX operating system. In this part we will learn how to install a web interface to the blast programs. You can also use the command line blast programs installed during our first part, but for the inexperienced Unix-users it is much easier to use a cut and paste system than using the terminal.

We will also learn how to install a graphical interface to the most popular bioinformatics package namely EMBOSS.

Installing your own databases to use with BLAST

To find the BLAST databases the blast binaries look for a file called ".ncbirc". The command line binaries look at the environment variable NCBI to be set to the directory where the .ncbirc file is located.

If you use our distribution of the ncbi-tools and ncbi-blast this file is located in the /usr/local/blast/ directory, and the NCBI environment variable should already be set for you (/usr/local/blast). Edit this file with a text editor to reflect where you store your databases.

The .ncbirc file should always start with the line (square brackets included):

```
[NCBI]
```

And then three more lines:

```
Data=/your/path/here
```

```
[BLAST]
```

```
BLASTDB=/your/database/path/here
```

where "/your/path/here" is the path to the directory where you store the BLAST matrix and data files, and "/your/database/path/here" is

the path to the directory containing the BLAST databases.

Example:

You have the data directory with matrix files in /usr/local/blast/ and want to put all your BLAST databases in the directory "databases" located in: /usr/local/, then the .ncbirc file should contain these four lines:

```
[NCBI]
Data=/usr/local/blast/data
[BLAST]
BLASTDB=/usr/local/databases
```

Preformatted BLAST databases ready to use can be downloaded from the ncbi ftp server at:

ftp://ftp.ncbi.nih.gov/blast/db/

These databases need to be unzipped, untarred and placed in the directory you have specified in the .ncbirc file.

Fasta files for some of the most common databases that can be "BLAST-formatted" are also located at the ncbi ftp site:

ftp://ftp.ncbi.nih.gov/blast/db/FASTA

Any Fasta formatted sequence file can be turned into a BLAST database. In the next paragraph of this tutorial we will explain how you can create your own custom databases.

Formatdb

To create your own BLAST databases you must use a program called formatdb.

There are other tools available, but for this tutorial we will use formatdb. Formatdb can index Fasta formatted "flatfiles" containing protein or nucleotide sequences, the program was preinstalled with the ncbi-tools package.

formatdb is run from the command line (terminal) using the following important arguments:

```
-t Title for the database file
-i Input file for formatting
-p Type of file
  T - protein
  F - nucleotide
```

Example:

To create a BLAST database called "mydatabase" using nucleotide sequences from a file called drosoph.nt (please download the file from the ncbi ftp site or use one of your own, you can also download it from the tutorials at www.ebioinformatics.org) write the following on the terminal window (you must be in the folder

containing the file drosoph.nt!):

```
$sudo /usr/local/blast/bin/formatdb -p F -t
mydatabase -i drosoph.nt
```

You should then obtain in that folder a number of files called drosoph.nt having different file suffixes like .nhr .nin .nnd .nni .nsd .nsi and .nsq. These files should be moved to your blast database directory described above. If you format a file containing protein sequences the file suffixes should be .phr .pin .pnd .pni .psi and .psq (the starting n(ucleotide) exchanged for p(rotein)).

To find out more about formatdb look in the README's located in /usr/local/blast (if you have installed the ncbi-tools of course).

Creating a web interface to BLAST: WWWBLAST

This package is based on the standalone WWW BLAST server suite of programs created by the NCBI team.

First you must install our package "wwwblast.pkg" as usual (see Figure 1 on next page).

After this automatic installation you should have a folder containing the web interface to BLAST in /Library/WebServer/Documents/blast.

In the next step you must change some parameters in the Apache web-server installation to allow the execution of the installed blast scripts. This is done by editing some configuration files.

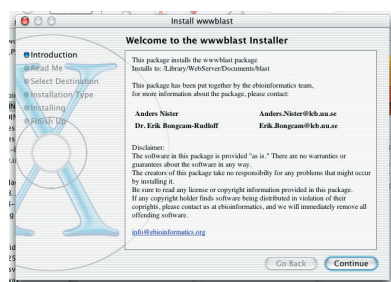
A -First you have to change your web-server parameters. You do this by editing a file called httpd.conf. We will use a text-editor called "pico", do like this (\$ is only to symbolize the prompt):

```
$sudo pico /etc/httpd/httpd.conf
```

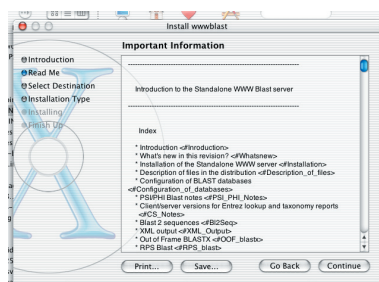
First modify the Options line adding ExecCGI where the file looks like this:

```
# This may also be "None", "All", or any
combination of "Indexes",
# "Includes", "FollowSymLinks", "ExecCGI", or
"MultiViews".
#
# Note that "MultiViews" must be named
*explicitly* --- "Options All"
# doesn't give it to you.
#
Options Indexes FollowSymLinks MultiViews
```

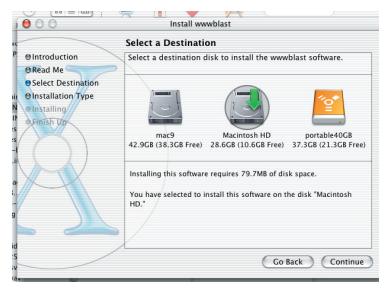
Change it to this: (the modification is shown in red)



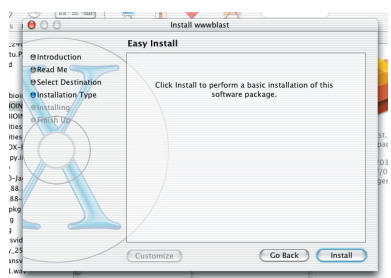
Step 1 Please start the install procedure as usual. Double-click on the unpacked installer named `wwwblast.pkg`. You will see here that the files will be installed on the Webserver that comes with MacOSX. MacOSX uses Apache as webserver.



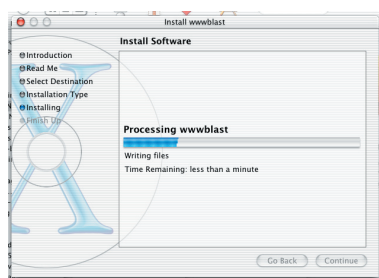
Step 2 Please read carefully this information from the NCBI team. You will find the same information later on as a web page on the installed `wwwblast`.



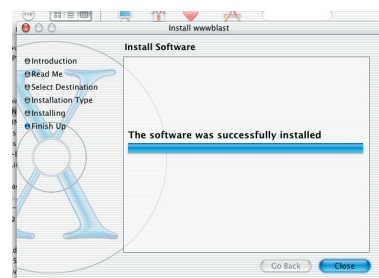
Step 3 Now you must select the harddisk where you have your WebServer, usually the same as the disk containing the System Folder.



Step 4 This is the last step before the install starts. Click on install.



Step 5 Now you can relax and the process will start, this takes a few minutes



Step 6 Now the automatic part is finished and we start with some manual operations to adjust your system.

Figure 1 Installation of the BLAST package

```
# This may also be "None", "All", or any
# combination of "Indexes",
# "Includes", "FollowSymLinks", "ExecCGI", or
# "MultiViews".
```

```
#
# Note that "MultiViews" must be named
# *explicitly* --- "Options All"
# doesn't give it to you.
```

```
# Options Indexes FollowSymLinks MultiViews ExecCGI
```

B - then find a line that looks like this:

```
# To use CGI scripts:
#
#AddHandler cgi-script .cgi
```

Uncomment (erase) the # before `AddHandler` and add `.pl`, like this:

```
# To use CGI scripts:
#
AddHandler cgi-script .cgi .pl
```

After editing press control-X , say yes, return and quit pico. (fig 2.)

C - Now you must add a file called `.htaccess` (please pay attention that the name starts with a dot, dots at the beginning of a name makes

the files "invisible" on Unix systems) in the same folder that our package installed the `wwwblast` (`/Library/WebServer/Documents/blast`). We use the program `pico` again.

A session in `pico` looks like this:

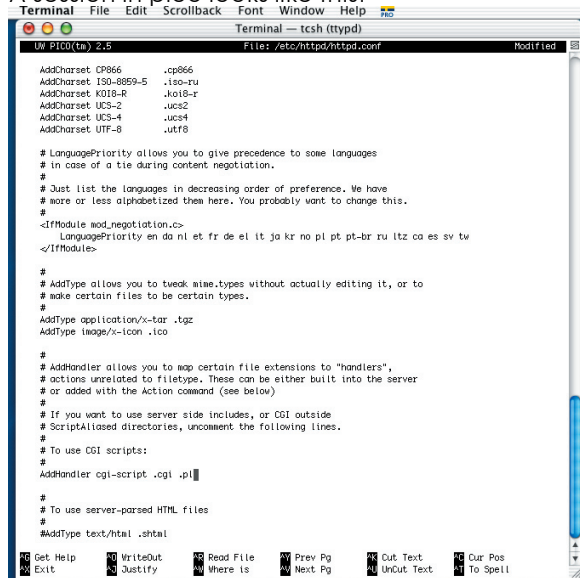


Figure 2. Shows `pico` when we edited the `httpd.conf` file. Please remember that this text editor is not like word and you can not use the mouse to navigate, But you avoid having format problems later on!!


```
$sudo pico /Library/Webserver/Documents/blast/.htaccess
```

When pico starts add the following line at the end of the file:

```
Options +ExecCGI
```

Then control-X, answer yes, return and it is done. Restart your web server in System preferences (go to Sharing, select Personal Web Sharing and click Start or Stop button once or twice depending of the current state of your web server):

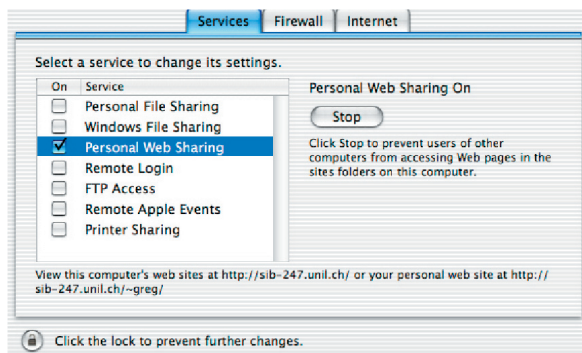


Figure 3. How to restart your web server

Now you are the lucky owner of a workstation/server running your very own blast-server.

To test your installation open your favourite web-browser and type `http://localhost/blast` (from another computer change localhost with the URL of that machine e.g., `http://mymachine.slu.se`) you should see a page like this:

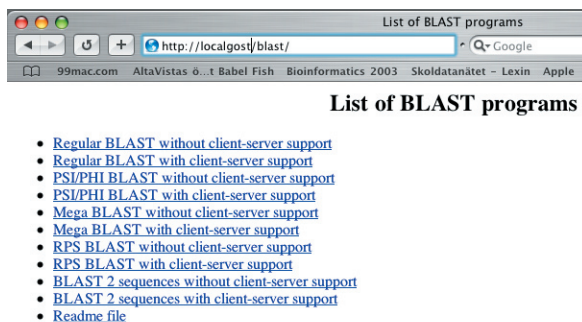


Figure 4. The picture illustrates just an example of layout you can configure in the way that suites you. This are html for template purposes only.

How to connect your databases to the web-interface?

First install your blast-formatted databases in your selected folder as described above. Secondly you must tell the web-interface where to find your databases, this is done by editing some files called: `blast.rc`, `psiblast.rc` and the

corresponding html files. Use the perl script that is supplied by the NCBI team (`config_setup.pl`).

This perl script is located in your `www-blast` folder (`/Library/WebServer/Documents/blast/`).

First we move the test databases to a new directory and subsequently we create a link "db" to the real directory:

```
$sudo mv db dbbak
```

```
$sudo ln -s /usr/local/databases db
```

Then you should run the perl script, in this case we decided to install our databases in `/usr/local/databases` therefore the command is:

```
$sudo perl config_setup.pl /usr/local/databases out
```

where `/usr/local/databases` is the place where we keep the blast databases and "out" is a temporary directory where we can find the files created use them to replace the original ones.

```
$sudo cp out/blast.rc .
```

```
$sudo cp out/psiblast.rc .
```

Please pay attention to the "."! (it means the current directory)

Now you have to edit the html pages to show your new databases.

```
$sudo pico blast.html
```

It should look like this:

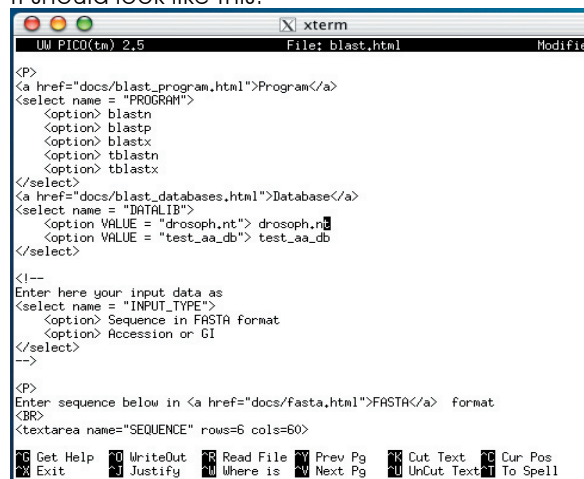


Figure 5. Edition of the blast.html file with Pico

If you have done no mistakes you will be ready to try your first search. Start your web browser and go to your `wwwblast` pages (Figure 6). Paste a query sequence, select the parameters and click "search" button. After a while you should get the result page (Figure 7).

Choose program to use and database to search:

Program Database

Enter sequence below in FASTA format

```
>test drosophila
TACCCACCA GTTAACAGT GCTATTAGT TATTAGTA ATTATGTC CTTTTTTTAAATGTTA
TAATACCAA TAAGATAGT GGACCAAT ATGTTAATG GCTCTTAATACGGGATG
GTTTGCGT GGCAATGAT GCAATGCTG CTGATGCTG TGTCAACAA TTTTATTA
ATGAATGCG CTGTACCA CCGTGA AAAAACAAC ACAGCGCCGCGCAATGCG
TAATGCGT CCAGACAAC GCTGCAATG AATGGGATG GGCAACATGGGAATGTT
```

Or load it from disk no file selected

The query sequence is [filtered](#) for low complexity regions by default.
Filter ☒ Low complexity ☐ Mask for lookup table only

Expect Matrix ☐ Perform ungapped alignment

Query Genetic Codes (blastx only)

Frame shift penalty for blastx

Other advanced options:

☒ Graphical Overview ☐ Alignment view
Descriptions Alignments Color schema

Figure 6. In this example we have the test database we formatted "drosoph.nt" (you can download this file from the tutorial pages at www.ebioinformatics.org).

There are also two test blast databases on this installation test_na_db (nucleic) and test_aa_db (protein).

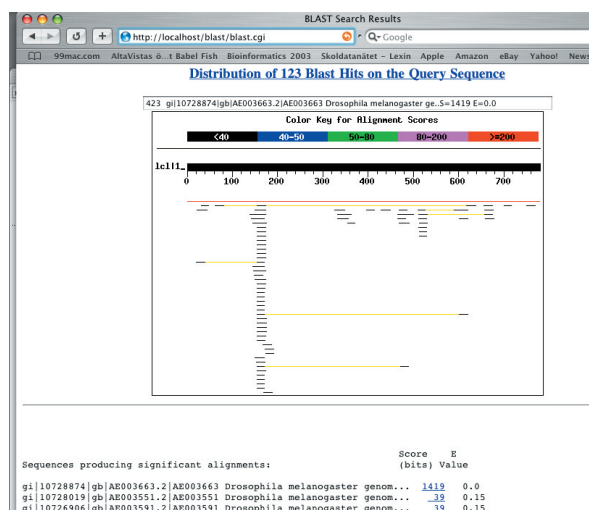


Figure 7. The result of a simple search using a nucleic acid sequence against the Drosophila genomic database.

Now you have a working web interface to the blast programs. Install your databases and connect them as explained. Lycka till! as we say in Sweden.

Installing a GUI interface to EMBOSS

There are several graphical user interfaces (GUI) and Web interfaces for the EMBOSS package. Please visit the EMBOSS web site to read more about this. In this second tutorial we have chosen to install Thomas Siegmund's creation. He uses "Kaptain" created by Terék Zsolt to wrap GUIs to the EMBOSS applications. Kaptain is a universally applicable graphical front-end based on context-

free grammars. You need EMBOSS (installed in our first part), Kaptain and QT-libraries installed. Our packages contain:

- QT installation package (install it first)
- Kaptain installation package (install it next)
- Emboss _ kaptain package (install it last)

If all the steps are made correctly you can start all EMBOSS programs from an X-Windows terminal with a module called "embosslauncher.kaptn" you start the GUI like this:

```
$ embosslauncher kaptn
```

You should get a window like this one:

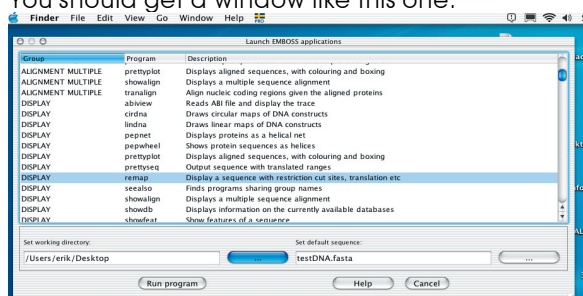


Figure 8. Shows the central command in Kaptain, you can start all EMBOSS programs from here.

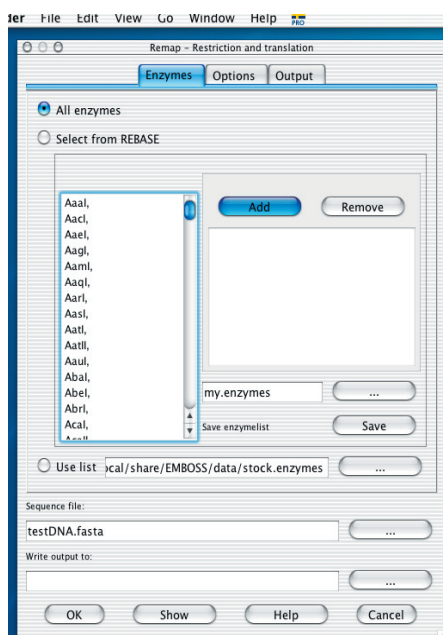


Figure 9. Here you see Kaptain running remap. OBS: You can also start all programs individually.

Most configurations are done automatically, but if you encounter any problem, read the Problems note (see Addendum). For up to date changes please visit <http://www.ebioinformatics.org>

Use "showdb" to check that it worked
\$showdb.kpt

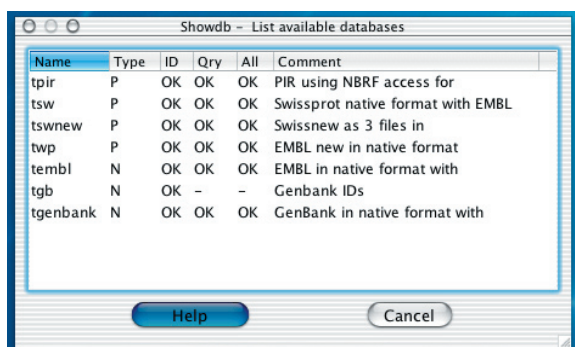


Figure 10. You should now get a list of the installed databases. You can later adapt the emboss.default to reflect your database installation

Nedit

The most interesting feature with this EMBOSS GUI is the incorporation of a sequence editor. For the preview of text and sequence data and for the display of error messages Thomas Siegmund uses Nedit. Nedit is a fast, powerful and lightweight open source text editor. We have created a new Fink package including Nedit in the installation, please install it. You can of course install the binaries downloading directly from Nedit's home page (See web resources).

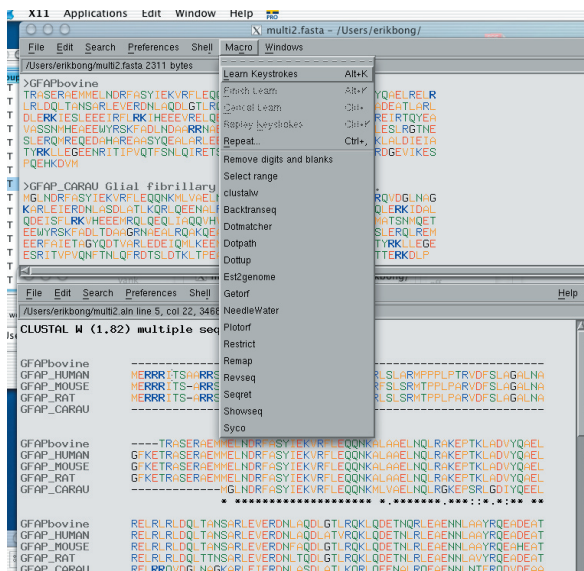


Figure 11. In this figure you see Nedit with a multi fasta protein list. From Nedit we launched ClustalW and we obtained the result on another fully editable window.

Thomas created also a .nedit settings file. Placing this .nedit file in your home directory allows Nedit to show some common sequence file formats with a beautiful highlighted syntax. Please download

his file from

<http://userpage.fu-berlin.de/~sgmd/.nedit>

Installing the parameters mentioned above enables to start a nice ClustalW alignment from Nedit.

Start nedit from your X-Windows terminal (fig. 11).
\$nedit

Extra packages:

ClustalW and T-Coffee

To work with EMBOSS.kaptain and other future package installations, we also recommend to install the following packages in the section "New-packages": ClustalW and T_Coffee, visit: <http://www.ebioinformatics.org> to download them.

This two packages work on the command line without any special configuration and ClustalW works also through other installed GUI interfaces, eg. Emma and Kaptain. Do exactly in the same way as we explained in tutorial part I (see EMBnetNews vol, 9 nr 1: <http://www.embnet.org/download/embnetnews/index.html>).

ClustalW is a general purpose multiple sequence alignment program for DNA or proteins.

T_Coffee is also a multiple sequence alignment package with some extra features: e.g., it allows you to combine results obtained with several alignment methods. For instance if you have an alignment coming from ClustalW, and another alignment coming from Dialign, T-Coffee will process the information and produce a new consensus multiple sequence alignment using the incoming data.

Next Part?

In part 3 we are going to explain how to install Jembooss and a web-interface to EMBOSS. We will also install a GUI for T_Coffee.

For all of you wondering why MacOSX and bioinformatics: we will also review the new 64-bit G5 Macintosh machines and do tests to compare Altivec-enhanced programs and the same programs running on other platforms. Do not miss it!

Web addresses

<http://www.embnet.org>

<http://www.ebioinformatics.org>

<http://www.ncbi.nlm.nih.gov/>

<http://www.emboss.org>

<http://userpage.fu-berlin.de/~sgmd/>

<http://www.trolltech.com/>

<http://kaptain.sourceforge.net/>

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

<http://www.ebi.ac.uk/clustalw/>

http://igs-server.cnrs-mrs.fr/~cnotred/Projects_home_page/t_coffee_home_page.html

References

- 1 Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389-3402.
- 2 Rice P., Longden I., Bleasby A. (2000) EMBOSS: The European Molecular Biology Open Software Suite *Trends in Genetics*, 16, 276-277.
- 3 Thompson J.D., Higgins D.G., Gibson T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673-4680.
- 4 Notredame C., Higgins D., Heringa J. (2000) T-Coffee: A novel method for multiple sequence alignments. *Journal of Molecular Biology* 302, 205-217.
- 5 Bongcam-Rudloff E. (2003) A guide on how to install Bioinformatics tools on MacOSX (Part1). *Embnetnews*, vol.9 nr.1, 13-16.

Glossary

sudo This command means "superuser do". It enables allowed users to execute a single command with the root user privileges. You must have administrator status to invoke the sudo command. This command is used to access permissions and ownerships or perform tasks reserved for root. To be used with caution!

pico Pico is a simple, display-oriented, text editor based on the Pine message system composer. As with Pine, commands are displayed at the bottom of the screen, and context-sensitive help is provided.

As characters are typed they are immediately inserted into the text.

QT Qt is a multiplatform, C++ application development framework created by Trolltech, Norway. Qt lets developers write a single application source that will run natively on Linux/Unix, Windows, Mac OS X, and embedded Linux with a simple recompile.

unzipped From unzip. Archives compressed in zip format can be unpacked using the command unzip. Normal command: `$unzip myfile.zip`

untarred From tar. The tar command creates, adds files to, or extracts files from an archive file in tar format.

Normal command: `$tar -xvf yourfile.tar`

You can also use the free Stuffit Expander, Stuffit unzip and untar files to a folder automatically.

Addendum

Most configurations are done automatically, but if you encounter any problem try this:

- 1 To instruct some of the Kaptain-EMBOSS programs where you keep your data files, like restriction enzymes, you must change your environment.

In our first tutorial (part I) you created a file in your home directory called «.cshrc».

```
source /sw/bin/init.csh
setenv PLPLOT_LIB /usr/local/share/emboss
set path=( /usr/local/bin /usr/local/blast/bin
/usr/local/staden/macosx-bin ${path} )
setenv DISPLAY :0.0
```

Add the following line to that file (use pico!):

```
setenv EMBOSS_DATA /usr/local/share/EMBOSS/data
```

so that it looks like this:

```
source /sw/bin/init.csh
setenv PLPLOT_LIB /usr/local/share/emboss
setenv EMBOSS_DATA /usr/local/share/EMBOSS/data
set path=( /usr/local/bin /usr/local/blast/bin
/usr/local/staden/macosx-bin ${path} )
setenv DISPLAY :0.0
```

- 2 To get access to sequence databases connected to EMBOSS you have to edit a file called `emboss.default`. The normal installation installs a template. Now we are going to change it to reflect our installation. You should adapt this file when you install your own databases, for the moment we are going to use some preinstalled test databases.

```
$cd /usr/local/share/EMBOSS
$sudo cp emboss.default.template emboss.default
$sudo pico emboss.default
```

Change a line that looks like this:

```
#SET emboss_tempdata path_to_directory_
$EMBOSS/test
remove the # and change the path to:
SET emboss_tempdata /usr/local/share/EMBOSS/test
```

Use "showdb" to check that it worked

```
$showdb.kpt
```


THE PHYSIOLOGICAL HOLY GRAIL?

By Vivienne Baillie Gerritsen

Though we seem to be quite solid, we are in fact quite liquid. The best part of us is water. In fact, all living creatures – from bacteria to plants and man – are made up of roughly 70% water. And that water needs to flow into us, out of us and inside us. We sweat water, we cry water, we digest with water, we think thanks to water, we pee water – to name but a few aqueous physiological activities. Imagine that, on a daily basis, hundreds of litres of water go through a human kidney! It has been known for decades that water molecules can quite happily cross cell membranes unassisted. However, transit in this way could not account for the huge amounts a kidney would need for example. There needs to be another system. In the 1990s such a system was discovered: aquaporin. Aquaporins are proteins which are embedded within cellular or intracellular membranes and function as a rather high-tech channel specific to water molecules. What is more, aquaporins have been found in plants, animals and yes, bacteria.

In the 1950s, scientists were already wondering how on earth huge amounts of water could possibly diffuse through membrane lipid bilayers. Diffusion in this way is indeed effective but rather slow and simply could not account for it. Surely there must exist some kind of water pore. It took a further thirty years before the very first aquaporin was actually discovered...and quite by chance. Indeed, scientists had been rummaging around red blood cells on the lookout for a specific Rhesus factor molecule when they stumbled upon an odd protein lodged in the plasma membrane. After close inspection, it turned out that this protein was in fact a pore which let water in and out of the cell. Aquaporin has been called the Holy Grail of fluid-transport physiology....

What could a physiological Holy Grail possibly look like? Well...a dumbbell. Or to be more precise an assembly of four dumbbells. Indeed, one aquaporin channel is made up of four aquaporin monomers, each of which acts as a specific water pore. An aquaporin monomer has a diameter of about 30Å and a height of about 60Å. The four monomers are quite tightly bound and form a stable complex in the plasma membrane. Were you to unravel one, you would

find six longish alpha helices and two shortish ones. Imagine a cylinder. Take six tubes and place them vertically around the cylinder. That is the way the six helices wrap themselves around the centre. Tie a ribbon around the centre of the pore. It forms a constriction in the middle, of a diameter of about 8Å. This is what gives the dumbbell shape to the aquaporin monomers – though there is no ribbon, simply molecular forces working away. The two short helices tuck themselves into the very middle of the pore and form in effect a barrier to molecules other than water. These two short helices are what make an aquaporin what it is, i.e. a pore for water molecules only.

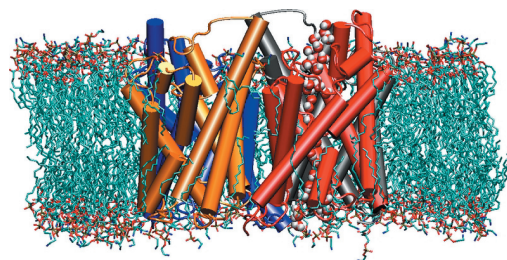


Fig. 1 3D model of aquaporin

Source: http://www.psc.edu/science/2002/schulten/precious_bodily_fluids.

How is water driven through an aquaporin then? The passage of water molecules through the pores is rather charming and has been described as a molecular ballet. Nothing complicated, no back-breaking arabesque or hurried pas de deux, no just a smooth glide and then a graceful pirouette. What happens is that these aqueous ballerinas drift into an aquaporin one by one, face down so to speak, i.e. their oxygen facing the inside of the pore. Though perhaps 'drift' is not quite the word since one billion molecules cross a membrane...per second! However, drift or no drift, once they reach the middle of the channel, the water molecules are grabbed by side chains which line the interior of the pore and swung around so that they exit the pore, bottom down so to speak, i.e. hydrogens first.

What on earth would they do this for? Well the system is ingenious. It all has to do with

protons. Indeed, not only do aquaporins let water through but they also inhibit the transit of protons. And this is a very wise move. Why? Protons are needed to charge cells; if cells lose their protons, they lose their energy. The thing is, protons usually hitch a ride on the backs of water molecules. A line of water molecules forms what is known as a proton wire, a path on which protons can move from one place to another. Now if this path is disrupted, the protons have only one choice but to go back from where they came. This is precisely what happens in an aquaporin. And thanks to the two short helices mentioned earlier. It is precisely at this point that the incoming water molecules perform a pirouette and disrupt the proton wire. The protons go back home while the water molecules continue their way to the other side. Besides, the two short helices, there is also a second major mechanism that stops other molecules from passing through: the constriction. The size of the constriction is such that it only lets molecules the size of water molecules through...

Evidently, aquaporins must be at the heart of a number of diseases if their role is, as is the role of water, so important. That is why it is crucial to get to know aquaporins on a very intimate

level, with a view to develop new therapeutics thanks to the world of drug design. As an example, corneal transparency, i.e. vision, requires a precise regulation of water content. Indeed, it could be that water has a role in keeping the diameter and spacing of collagen regular, which in turn confers transparency to the lens. Cataracts are a direct consequence of aquaporin malfunction. Nephrogenic diabetes insipidus is a disorder in which patients' kidneys cannot reabsorb water correctly; patients have to go to the bathroom frequently and end up dehydrated. Defective aquaporins are no doubt at the heart of this form of disease. Aquaporin malfunctions in salivary and lacrimal glands also result in a disorder known as Sjogrens syndrome, or dry mouth. Besides the malfunction of aquaporins, their overexpression can be indicative of physiological states such as congestive heart failure or pregnancy. Though the latter should not be a cause to worry. No doubt, within the next decade, a growing number of clinical disorders will point their finger at a deregulation of aquaporin function or expression. One interesting biotechnological development could be to incorporate aquaporins to certain materials which could then filter ions, such as salt, from seawater and produce fresh water in countries where it is so scarce.

Cross-references to Swiss-Prot

P29972: human aquaporin 1
P43285: *Arabidopsis thaliana* (mouse ear-cress) aquaporin PIP1.1
Q9C4Z5: *Methanobacterium thermoautotrophicum* aquaporin aqpM
Q23808: *Cicadella viridis* (Green leafhopper) aquaporin AQPcic

References

1. Tajkhorshid E., Nollert P., Jensen M.Ø., Miercke L.J.W., O'Connell J., Stroud R.M., Schulten K. Control of the selectivity of the aquaporin water channel family by global orientational tuning Science 296:525-30(2002). PMID: 11964478
2. Kozono D., Yasui M., King L.S., Agre P. Aquaporin water channels: atomic structure and molecular dynamics meet clinical medicine J. Clin. Invest. 109:1395-9(2002). PMID: 12045251
3. Structure, Dynamics, and function of aquaporins
NIH resource for macromolecular modeling and bioinformatics
<http://www.ks.uiuc.edu/Research/aquaporins/>

National Nodes

Argentina

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

Australia

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

Austria

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

Belgium

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

Brasil

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

Canada

Laura Brown
Canadian Bioinformatics Resource, National Research
Council Canada, Institute for Marine Biosciences,
Email: manager@cbr.nrc.ca
Tel: +1-902-426 7310 Fax: +1-902-426 9413
<http://www.ca.embnet.org>

Chile

Dr. Ricardo Baeza-Yates
Dept. of Computer Science, Santiago,
Email: rbaeza@dcc.uchile.cl
<http://www.embnet.cl>

China

Jingchu Luo
Room 303, Exchange Centre, Peking University
Email: luojc@cbi.pku.edu.cn
Tel: +86-10-6275 9001
<http://www.cbi.pku.edu.cn>

Colombia

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

Cuba

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

Denmark

Hans Ullitz-Moeller
BioBase, University of Aarhus
Email: hum@biobase.dk
Tel: +45-86-13 9788
<http://www.dk.embnet.org>

Finland

Kimmo Mattila
CSC, Espoo
Email: Kimmo.Mattila@csc.fi
Tel: +358 9 457 2708
<http://www.fi.embnet.org>

France

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

Germany

Sandor Suhai
EMBnet node at the German Cancer Research Center
Department of Molecular Biophysics (H0200)
Email: genome@dkfz.de
Tel: +49-6221-422 342 Fax: +49-6221-422 333
<http://www.de.embnet.org>

Greece

Babis Savakis
Institute of Molecular Biology and Biotechnology
Heraklion, Crete
Email: savakis@nefeli.imbb.forth.gr
Tel: +30-81-391 114 Fax: +30-81-391 104
<http://www.imbb.forth.gr>

Hungary

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

India

H.A.Nagarajaram
Laboratory of Computational Biology & Bioinformatics
facility, Centre for DNA Fingerprinting and Diagnostics
(CDFD), Hyderabad
Email: han@www.cdfd.org.in
Tel: +91 40 7155607 / 7151344 ext:1206
Fax : +9140 7155479
<http://www.in.embnet.org>

Israel

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

Italy

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

Mexico

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

The Netherlands

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

Norway

Rune Groven
The Norwegian EMBnet Node
The Biotechnology Centre of Oslo
Email: admin@embnet.uio.no
Tel: +47 22 84 0535
<http://www.no.embnet.org>

Poland

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

Portugal

Pedro Fernandes
Instituto Gulbenkian de Ciencia
Unidade de Bioinformatica
2781-901 OEIRAS
Email: pfern@igc.gulbenkian.pt
Tel: +351 214407912 Fax: +351 2144079070
<http://www.pt.embnet.org>

Russia

Sergei Spirin
Biocomputing Group, Belozersky Institute Moscow
Email: sas@genebee.msu.su
Tel: +7-095-9328825
<http://www.genebee.msu.su>

Slovakia

Lubos Klucar
Institute of Molecular Biology SAS Bratislava
Email: klucar@embnet.sk
Tel: +421 7 5941 2284
<http://www.sk.embnet.org>

South Africa

Ruediger Braeuning
SANBI, University of the Western Cape, Bellville
Email: ruediger@sanbi.ac.za
Tel: +27 (0)21 9593645
<http://www.za.embnet.org>

Spain

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

Sweden

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

Switzerland

Laurent Falquet
Swiss Institute of Bioinformatics, CH-1066 Epalinges/
Lausanne
Email: Laurent.Falquet@isb-sib.ch
Tel: +41 (21) 692 5954 Fax: +41 (21) 692 5945
<http://www.ch.embnet.org>

United Kingdom

Alan Bleasby
UK MRC HGMP Resource Centre, Hinxton, Cambridge
Email: ableasby@embnet.org
Tel: +44 (0) 1223 494535
<http://www.uk.embnet.org>

Specialist Nodes

EBI

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

ETI

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

EU

Dr Bernard Mulligan
DG Research - European Commission
Brussels BELGIUM
Email: bernard.mulligan@cec.eu.int

ICGEB

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

LION Bioscience

Peter Rice
LION Bioscience AG, Heidelberg, Germany
Email: Peter.Rice@uk.lionbioscience.com
Phone: +44 1223 224700
<http://www.lionbioscience.com>

MIPS

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

Pharmacia - Biovitrum

Timothy Wood
Biovitrum, Stockholm, Sweden
Email: timothy.wood@eu.pnu.com
Phone: +46 (8) 695 9134
<http://www.pharmacia.se>

Sanger Institute

Michelle Clamp
Wellcome Trust Sanger Institute, Wellcome Trust
Genome
Campus, Hinxton Hall, Hinxton CB10 1SD, Cambridge,
United Kingdom
Email: michele@sanger.ac.uk

Phone: +44-1223-494 967
Fax: +44-1223-494 919
<http://www.sanger.ac.uk>

UMBER

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

TECH-MGR

Email: tech-mgr@embnet.org
The team gives support to EMBnet nodes and helps
them with maintenance and troubleshooting.
The team is formed of experienced system administrators
and programmers who ensure the availability of local
services for all EMBnet users.



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 Internet E-mail.

Past issues of embnet.news are available as
PostScript or PDF files (ISSN 1023-4144). You can
get them by anonymous ftp from:
the EMBnet organisation Web site
<http://www.embnet.org/download/embnetnews>
the Belgian EMBnet node
<ftp://ftp.be.embnet.org/pub/embnet.news>
the UK EMBnet node
<ftp://ftp.uk.embnet.org/pub/embnet.news>
the EBI EMBnet node
<ftp://ftp.ebi.ac.uk/pub/embnet.news>

Submission deadline for next issues:

October 31, 2003
February 29, 2004

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

EMBnet.news

Volume 9 Nr. 3
December 2003

- 
- Training and education with EMBnet
 - Australian & New Zealand undergraduate degrees in Bioinformatics
 - Bioinformatics in Latin America
 - MacOSX and Bioinformatics (part 3)
 - EMBnet E-meeting portal

Editorial

Education is one of the key component of the EMBnet community. Not only we have thousands of users around the world that are participants in our EMBnet courses and training sessions, but most of us organise and participate in the teaching of many academic degrees in bioinformatics.

Two articles in this issue exemplify this. One describes the current situation of the bioinformatics degrees in Australia and New Zealand, whereas the second explains the role of the E&T PC (Education and Training Program Committee) and details our new web submission tool for course organisers. Latin America is also very active in the Bioinformatics fields as demonstrated by a review of the non-EMBnet groups.

Don't miss also the last part of our MacOSX and Bioinformatics series. Finally a little report with some pictures of our last AGM in Poland.

The editorial board: Erik Bongcam-Rudloff, Laurent Falquet, Pedro Fernandes, Oscar Grau, Gonalo Guimaraes Pereira

Protein Spotlight

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

Publisher:

EMBnet Administration Office
c/o Jack Leunissen
Dept. of Genome Informatics
Wageningen UR, Dreijenlaan 3
6703 HA Wageningen
The Netherlands
Email: Jack.Leunissen@wur.nl
Tel: +31 317 484074

Contents

Editorial	2
Australian & New Zealand undergraduate degrees in Bioinformatics	3
Bioinformatics in Latin America	6
Training with EMBnet	10
AGM03 in Poland	12
MacOSX and Bioinformatics (Part III)	14
EMBnet E-meeting portal	18
Protein spotlight 40	20
National nodes	22
Specialist nodes	24

Editorial Board:

Erik Bongcam-Rudloff, The Linnaeus Centre for Bioinformatics, SLU/UU. SE
Email: erik.bongcam@bmc.uu.se
Tel: +46-18-4716696
Fax: +46-18-4714525

Laurent Falquet, SIB, Lausanne. CH
Email: laurent.falquet@isb-sib.ch
Tel: +41-216925954
Fax: +41-216925945

Pedro Fernandes, Instituto Gulbenkian. PT
Email: pfern@igc.gulbenkian.pt
Tel: +315-214407912
Fax: +315-214407970

Oscar Grau, IBBM, AR
Email: grau@biol.unlp.edu.ar
Tel: +54-221-4259223
Fax: +54-221-4259223

Gonalo Guimaraes Pereira, UNICAMP. BR
Email: goncalo@unicamp.br
Tel: +55-19-37886237/6238
Fax: +55-19-37886235

Cover picture: a zinc cast of the sculpture by Konstanty Hegel, made in 1855. Syrena, Old Town (Stare Miasto) Warsaw, September 2003 [© Laurent Falquet]

Australian & New Zealand undergraduate degrees in Bioinformatics



S.K. Cattley

Australian National Genomic Information Service
University of Sydney
Camperdown 2006,
NSW, Australia

In Australia and New Zealand there has been a movement away from general science to more specialized degrees. Bioinformatics has been a hot topic for the past 5 years and as such, a number of bioinformatics degrees have been instigated. For students in next years intake there are eight specialized undergraduate bioinformatic degrees to choose from and another seven bioinformatic streams/majors within existing Bachelor of Science degrees.

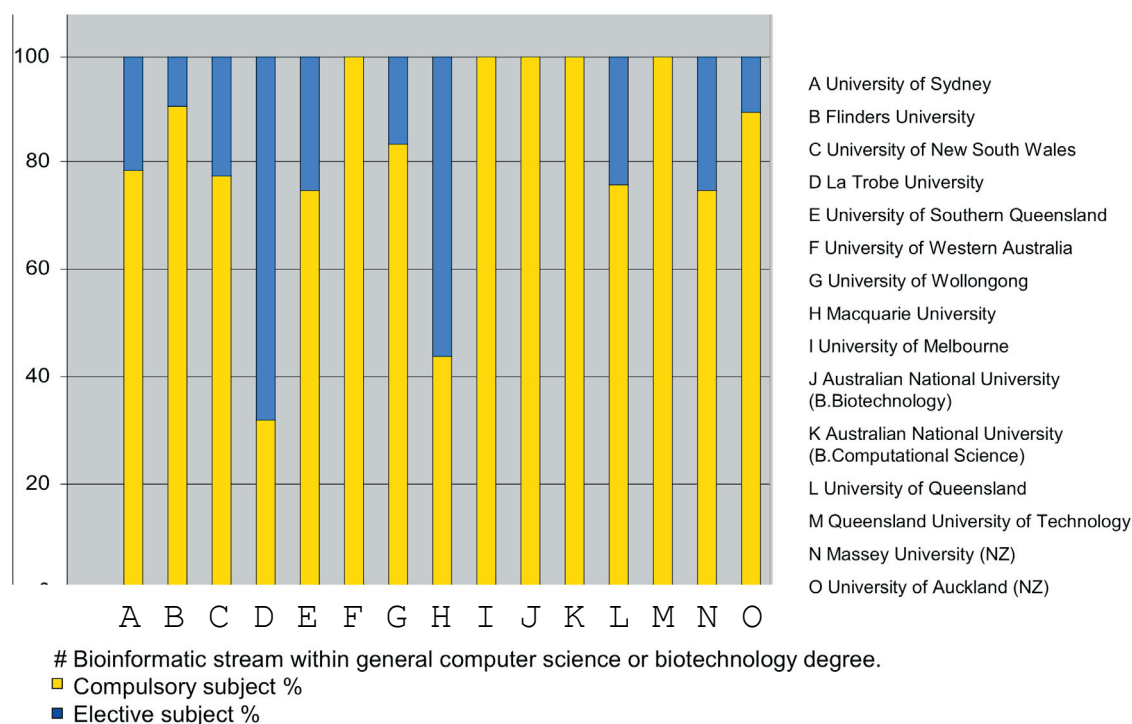
The subject content between and within these degrees is not uniform. In this analysis, the degrees were examined to enable a comparison of the relative strengths and weaknesses in the curricula. Most of the degrees included a number of electives. In some cases only one or two electives were permitted but in other instances electives consisted of more than 50% of the entire degree. A graph of the proportion of elective subjects in each of the degree is given in Figure 1. Five of the degrees allowed for no variation. The majority of the dedicated

Abbreviations:

UWA (University of Western Australia), UNSW (University of New South Wales), USQ (University of Southern Queensland), UofW (University of Wollongong), USyd (University of Sydney), Flinders (Flinders University of South Australia), La Trobe (La Trobe University), Macquarie (Macquarie University), QUT (Queensland University of Technology), UQ (University of Queensland), ANU (Australian National University), Melbourne (University of Melbourne), Massey (Massey University, NZ), Auckland (University of Auckland, NZ).

Figure 1: Ratio of compulsory to elective subjects in undergraduate bioinformatic degrees in Australia and New Zealand

Elective component indicates the portion of subjects that can be taken from different subject area (biology, computer, statistics, mathematics, bioinformatics, other).



degrees allow around 10-20% of the total course content to vary depending on the students own interest. Two degrees allow 50% or more of the subjects to be chosen by the student.

The fifteen degrees were examined from the websites listed in the appendix and six subject areas were chosen as markers for course content. Due to the differences in the number and choice of electives between the degrees the compulsory curricula only were used in the comparison.

- Biological Sciences includes all biology and chemistry subjects.
- Computer Science includes programming, computer studies and other subjects generally taught by the Department of Computer Science.
- Statistics includes all statistic and probability subjects.
- Mathematics includes mathematical subjects not covered in the area of statistics.
- Subjects that included a combination of biology and computer science (e.g. phylogenetics, micro array analysis or Bioinformatic project) and/or were titled bioinformatics were grouped into the area Bioinformatics.
- The area Other, included subjects that ranged from ethics, project management, experimental design and language comprehension through general electives outside the Faculty of Science.

While each of the courses surveyed contains a number of biological science and computer science subjects, the number of mathematics, statistics and bioinformatic subjects vary greatly and are even absent in some degrees. Figure 2. Statistics as a compulsory was absent from three degrees (Sydney, Flinders, Melbourne) while mathematics was absent from two degrees (Sydney and Macquarie). Dedicated Bioinformatic courses were not present in four degrees (Maquarie, La Trobe, USQ and Melbourne).

This last omission is surprising. In all cases the computer science and biological science subjects are taught by different departments. Only in the bridging bioinformatics subjects are the students exposed to a field of study where all the information they have learnt from biology, computer science, mathematics and statistics can be integrated together into one unit. Without this subject incorporated into the curricula the student is required to make these connections themselves.

The greatest variation observed occurs with the proportion of each subject area. Invariably the largest portions are dedicated to biological and computer science subjects. The actual proportions vary from between 25% (UNSW) through 60% (ANU) for biological science and 15% (ANU) through 50% (Melbourne) for computer science. The other subject areas also show variation. With selection of elective subjects, the proportions can change slightly. This can result in two students who undertake the same degree but graduate with vastly different levels of knowledge.

The explosion of bioinformatics degrees in Australia has already resulted in a casualty. In 2003, University of Western Sydney instigated a Bachelor of Science (Bioinformatics) but this was cancelled after just one year. It is important that the bioinformatics degrees we already have fill a specific niche or produce a particular kind of graduate in order to guarantee themselves a long life. For instance, UNSW is a computer engineering degree designed towards producing computer programmers, USQ almost fills the niche of a distance education degree, Massey and Auckland fits in the geographic need for bioinformatics education New Zealand (and surrounding South Pacific nations).

Appendix:

URL references bioinformatics degrees and streams:

Australian National University

<http://www.maths.anu.edu.au/bio.html>

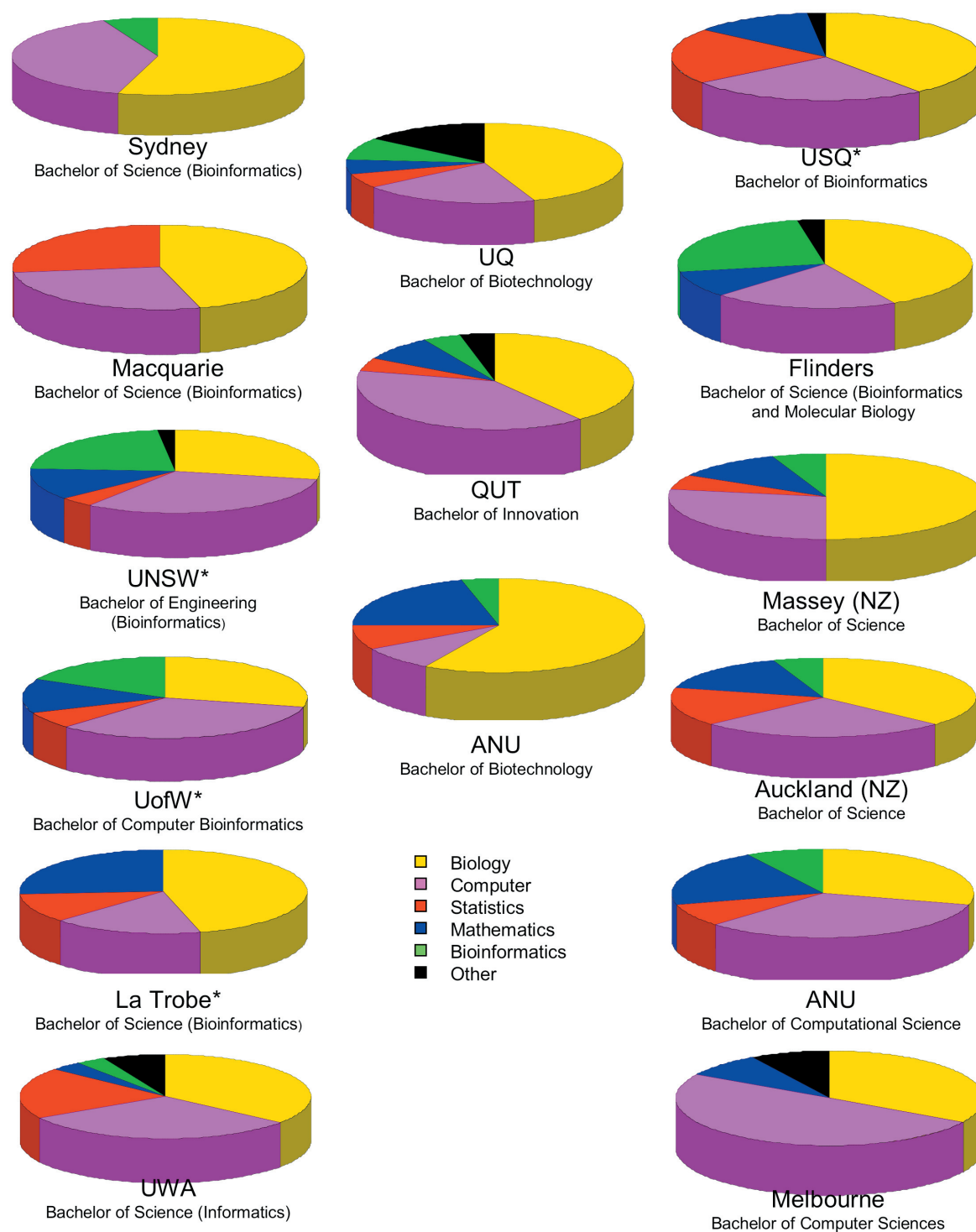


Figure 2: Distribution of compulsory subjects in bioinformatic streams and degrees available at Australian and New Zealand universities in 2004. All are 3 year degrees except those marked * which are 4 year

Curtin University

<http://handbook.curtin.edu.au/courses/30/301135.html>

Flinders University

<http://www.scieng.flinders.edu.au/courses/bioinformatics/>

La Trobe University

<http://www.latrobe.edu.au/handbook/scitech/stbsbb.htm>

Macquarie University

<http://www.hscandbeyond.com.au/bio.htm>

Massey University

http://study.massey.ac.nz/major.asp?major_code=2511&prog_code=92411

Queensland University of Technology

<http://www.studyfinder.qut.edu.au/cgi-bin/WebObjects/StudyFinder>

Use search term bioinformatics

Royal Melbourne Institute of Technology

<http://www.cs.rmit.edu.au/courses/postgraduate/escience/>

University of Melbourne

<http://www.cs.mu.oz.au/courses/ugrad/bcsbioinf.html>

University of New South Wales

<http://www.eng.unsw.edu.au/prospect/ug/be/bio.htm>

University of Queensland

http://www.uq.edu.au/study/plan.html?acad_plan=BIOINX2055

University of Southern Queensland

<http://www.usq.edu.au/handbook/2003/BBIN.html>

University of Sydney

<http://www.science.usyd.edu.au/future/ug/bsc-bio.html>

http://www.scifac.usyd.edu.au/future/pg/pgc_bioinform.html

University of Western Australia

<http://www.publishing.uwa.edu.au/handbooks/2003/sciences/BScInformatics.html>

University of Wollongong

<http://www.uow.edu.au/discover/courses/yr2002/bcompbioinf.html>

University of Auckland, New Zealand

<http://www2.auckland.ac.nz/science/subjects/bioinform.ptml>

Bioinformatics in Latin America



Oscar Grau

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

Introduction

Bioinformatics, as other emergent areas of science, has had a lag period to be established in the Region; however, in some countries it has sprung vigorously accompanying genetic engineering and genomic projects.

The contacts with members of EMBnet have been instrumental for the consolidation of the Cuban (<http://www.cu.embnet.org/>) and Argentinean (<http://www.ar.embnet.org>) nodes which had been providing services since the early 90s and later on for the Mexican (<http://embnet.cifn.unam.mx>), Colombian (<http://bioinf.ibun.unal.edu.co>), Chilean (<http://www.embnet.cl>) and Brazilian (<http://www.br.embnet.org>) nodes.

Inspired in the network organization of EMBnet and following the mandate of RELAB (Latin American Network of Biological Sciences) and with the sponsorship of FAO and BIOLAC a meeting was conveyed in Rio de Janeiro in June 2001. Then, a Latin American Bioinformatics Network project was elaborated and proposed to UNU-BIOLAC (United Nations University-Biotechnology for Latin America and the Caribbean, figure 1) for support. It includes members from Argentina, Bolivia, Brazil, Chile, Colombia, Cuba, Mexico and Venezuela, and provides an efficient clearhouse for information.

<http://www.lacbionet.org>

In 2002 an Ibero-american network was proposed to CYTED (Ciencia y Tecnología para el Desarrollo) and, as a consequence, a preparatory meeting was held in Buenos Aires in March 2003. A symposium, in which 24 members of 10 countries of Latin America

and Spain presented their interests and results in Bioinformatics was held at the same time.

A project was elaborated and submitted to CYTED for support. All participants agreed to join efforts with the Latin American Network and coordinate activities.

These initiatives are fully operative at the moment and more details can be observed in the site <http://portal-bio.ula.ve>

Further information about EMBnet nodes from Argentina, Brazil, Colombia, Cuba, Chile and Mexico can be found on their respective web site and in the links provided by EMBnet at: www.embnet.org

A short description of other groups follows:

Argentina:

In Rosario, the Center of Computational Biology is a recent initiative of the Molecular Biology Institute and the Faculty of Exact

Sciences and Engineering of the National University of Rosario. They are interested in the design of motif detection algorithms for T. cruzi y G. lamblia proteins having capacity of interaction with the DNA or participation in the initiation complex. In the area of analysis of microarray data, they work in the design of classification algorithms based on error correcting codes and in the design of biclustering algorithms suitable for capturing coregulation mechanisms in circadian genes expression. In the present year the first course on Introduction to Bioinformatics will be offered.

The Bioinformatics Group at the Institute of Biotechnology, University of General San Martín, current interest is to develop an automated sequence analysis pipeline, customized to the unique features of single-pass sequences like ESTs and GSSs. This currently includes: clustering of sequence

LacBioNet The Bioinformatics Network for Latin-America and the Caribbean

Goals & Objectives | Training & Human Resources | Research & Development | Planning Meeting: report | Information & Communication | Bioinformatics Institutes | How to Join | Nodes & Services | News - Courses - Events

HOME

Welcome to the LacBioNet homepage

- Register in our database of people and Institutes active in Bioinformatics and Computational Biology
- Map of Bioinformatics Centers in Latin-America and the Caribbean
- Access our Helpdesk with utilities, information, courses, events and on-line training modules

Physical networks in the Region

- Internet 2 through Americas Path Network (AMPATH)
- Project CLARA - Cooperación Latino Americana de Redes Avansadas

Other Organizations

- ABionet - The African Bioinformatics Network
- APBioNet - The Asia Pacific Bioinformatics Network
- CYTED - Red Ibero Americana Bioinformatica
- ISCB - International Society for Computational Biology

Sponsored By

United Nations University
"advancing knowledge for human security and development"

UNU - BIOLAC
Biotechnology for Latin America and The Caribbean
Programa de Biotecnología para América Latina y el Caribe

Also supported by

Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, Brazil

UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases - TDR

To participate, please register in the database. Registration will automatically add you to the LacBioNet discussion list and place your Institute on the map.

To contact us, with questions, suggestions, links and utilities, send mail to lacionet@fiocruz.br

The LacBioNet website is hosted at the Oswaldo Cruz Foundation, Fiocruz, Rio de Janeiro, Brazil.
Coordination: Wim Degraeve - DBBM/IOC FIOCRUZ Webdesign: Leonardo Leite Database: Rafael Glória, André Raw

Figure 1. LacBioNet home page

data and analysis of the resultant non-redundant datasets to discover new genes. The focus is on annotation and prediction of function. Their laboratory is also interested in regulation of gene expression in trypanosomatids characterizing motifs and regulatory signals present in the UTRs of mRNAs. They provide access to two standalone sequence analysis suites through a web server: Paul Stothard's 'Sequence Manipulation Suite (SMS)' and Malay Basu's 'Sequence analysis using Web Resources (SeWer)'.

At the computing department of the University of Buenos Aires there is a database of repeat elements for *Trypanosoma Cruzi*, *T. brucei* and *Leishmanias* and they offer undergraduate and postgraduate courses on Bioinformatics.

At the University of Quilmes, a course for graduate students is offered and an enterprise "Digital genes" offers access to EMBOSS and a limited group of services.

Bolivia

Universidad Católica Boliviana, Cochabamba IIIA - Instituto de Investigación en Informática Aplicada.

They have developed the PC version MELANIE (Medical Electrophoresis ANALysis Interactive Expert) project in collaboration with several Swiss institutions.

Also, they have migrated the protein database EXPASY to a Linux platform and installed a mirror at <http://bo.expasy.org>. They have also developed new Bioinformatics tools for sequencing and identifying proteins by mass spec.

Brazil

Besides the four laboratories that constitute the EMBnet node there are very many sites developing Bioinformatics.

For brevity only a list is included, but their coordinates can be found at <http://www.lacbionet.org>

- **Belo Horizonte:** DCC - Departamento de Ciências da Computação da Universidade Federal de Minas Gerais, Laboratório de Biologia Computacional do Departamento

de Bioquímica e Imunologia da Universidade Federal de Minas Gerais, Laboratório de Computação Heurística, Núcleo de Bioinformática da Universidade Federal de Minas Gerais.

- **Brasília:** Embrapa - Genetic Resources and Biotechnology - Bioinformatics Laboratory, Bioinformatics Laboratory at Universidade Católica de Brasília, Projeto de Bioinformática do Laboratório de Química Computacional da Universidade Federal de Brasília.

- **Campinas:** Departamento de Genética e Evolução - IB - Unicamp. Instituto de Computação Universidade Estadual de Campinas, Laboratório de Bioinformática da Universidade de Campinas, Laboratório Nacional de Luz Síncrotron, Sting Millennium Suite - Structural Bioinformatics Group - EMBRAPA/CNPq

- **Curitiba:** Laboratório de Bioinformática do CEFET - PR, Tecpar - Instituto de Tecnologia do Paraná.

- **Florianópolis:** Laboratório de Bioinformática da Universidade Federal de Santa Catarina.

- **Foz do Iguaçu:** LABI - Laboratório de Bioinformática.

- **Recife:** CIn - Centro de Informática da Universidade Federal de Pernambuco GENOMA/PE - cana-de-açúcar.

- **Rio de Janeiro:** Biologia Matemática e Computacional, LNCC - Laboratório Nacional de Computação Científica, Federal University of Rio de Janeiro, Instituto Virtual de Bioinformática e Modelagem de Biosistemas, Pontifícia Universidade Católica do Rio de Janeiro - Departamento de Informática, FIOCRUZ - Fundação Oswaldo Cruz.

- **Porto Alegre:** LabBioInf - Laboratório de Bioinformática da Universidade Federal do Rio Grande do Sul, Ana Lúcia C. Bazzan - Laboratório de Informática da Universidade Federal do Rio Grande do Sul, Federal University Rio Grande do Sul.

- **São Carlos:** Laboratório de Inteligência Computacional.

- **São Leopoldo:** Laboratório de Bioinformática e Biologia Computacional.

- **São Paulo:** ICMC - Instituto de Ciências Matemáticas e da Computação

da Universidade de São Paulo, IME - Instituto de Matemática da Universidade de São Paulo, Laboratório de Bioinformática do Instituto de Química da Universidade de São Paulo, Laboratório de Bioinformática - Hospital do Câncer - AC. Camargo, Laboratório de Biologia Computacional do Instituto Ludwig, LEMB - Laboratório de Evolução e Bioinformática, NIB - Núcleo de Informática Biomédica da Universidade de Campinas, BIOINFO - Núcleo de Pesquisas em Bioinformática da USP.

- **Viçosa:** Biologia Molecular Computacional da Universidade Federal de Viçosa

Cuba

The Centro Nacional de Bioinformática (BIOINFO, <http://www.bioinfo.cu>) was created in October 2001 in the Ministerio de Ciencia Tecnología y Medio Ambiente (CITMA). Their interests are focused in developing artificial intelligence tools for data mining, interaction databases and annotation. They have a server (<http://www.bioinfo.cu/iPPI.html>) for the prediction and modelling of proteins and their particular interest is the prediction of the affinity of peptides to MHC class II.

Chile

The "Laboratorio de Bioinformática y Matemática del Genoma" (LBMG) is located in the Centro de Modelamiento Matemático (CMM) from Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile. Its main interests are: automatic genome annotation, microarrays analysis, metabolic pathway modelling and regulation, in particular in bacteria involved in biolixiviation of minerals.

<http://www.dim.uchile.cl/~genoma/proyectos.shtml>

At the Pontificia Universidad Católica de Chile, Facultad de Ciencias Biológicas, Depto. Genética Molecular y Microbiología, Molecular Bioinformatics Laboratory, the main interest is focused in Computational Biology, Protein Structure and Function, Prediction and interaction of them with other

elements. They provide access to ANOLEA, a program to assess the quality of a three-dimensional protein structure that uses a statistical potential at the atomic level and gives an energy profile as output.

<http://protein.bio.puc.cl/index.html>

The University of Talca has implemented an undergraduate career in Bioinformatics based on the Faculty of Engineering and the Institute of Plant Biology and Biotechnology. The group is interested in data mining and visualization of macromolecules.

Uruguay

Very recently, a group of experimental and theoretical biologists as well as informaticians created a consortium named CUBIC (Consortio Universitario de Bioinformática y Biología Computacional) to develop infrastructure, an undergraduate career in Bioinformatics and a postgraduate programme in computational biology.

<http://bioinformatics.fq.edu.uy/cubic/index.shtml>

Venezuela

"Servidor de Bioinformática" a Venezuelan Bioinformatics Server is a strategic alliance involving an interdisciplinary partnership of four Venezuelan institutions. It is a shared effort between the Centro Nacional de Cálculo Científico Universidad de Los Andes (CeCaCULA), The Instituto Venezolano de Investigaciones Científicas (IVIC), Corporación Parque Tecnológico de Mérida (CPTM), and Universidad de Los Andes (ULA).

The server offers support for analyzing DNA data, protein sequences, prediction of three-dimensional structures (protein and epitopes), development of web interfaces for biological software, data-mining of genomic data, and informatics tools for educational and research exchange among collaborators. This server is physically located at CeCaCULA in Mérida, Venezuela

<http://www.cecalc.ula.ve/bioinformatica>

E&T PC Training with EMBnet

The Education and Training Committee currently comprises four members based at three different institutes. Martin Bishop and Lisa Mullan are resident at the HGMP-RC (Human Genome Mapping Project Resource Centre), UK, where Martin is head of the Bioinformatics Division, and Lisa is involved in coordinating and teaching the various courses in the HGMP-RC training schedule. Janet Kelso is involved in the Stack Database at SANBI (South African National Bioinformatics Institute) and Isabel Marques coordinates administration for the Bioinformatics courses at the Gulbenkian Institute in Oeiras, Portugal alongside assisting researchers across the country with their bioinformatics queries.

Our mission is to hold the funds that are reserved for training from the EMBCORE grant and offer them to EMBnet members wishing to run bioinformatics training courses for students in their own, and EMBnet communities. Many successful courses have been run over the years, emphasis is slowly shifting away from introductory courses and more towards specialist courses. This is indicative of a community becoming more familiar with the area of bioinformatics tools.

Further money (Stichting funding) is available for those nodes outside the EU member countries, although it is not in such plentiful supply and subject to greater revision.

Courses tend to offer training to at least 20 students (although in the case of China, 120 students were taught by six trainers) and the training is generally conducted by two or three colleagues. Specialist courses may involve a lower ratio of trainers to students.

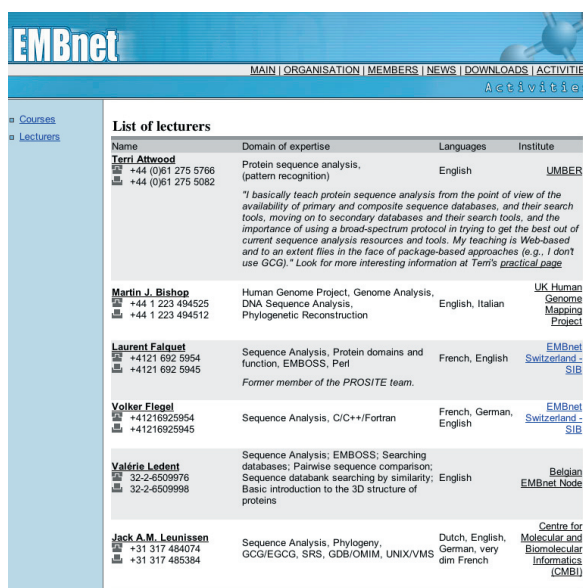
I have personally been involved in several courses across the world due to invitations to teach from the EMBnet community, and I am grateful to them all. It is a pleasure to be hosted by those interested in both enhancing the range of training offered in their country and wanting to explain their culture and customs to a foreigner. Also the students

must be applauded for their willingness to be taught in English. As a member of the ET committee, it is also important for me to see that the money has been well spent!

Last year (2002) the ETPC supported seven courses at a grand total cost of €40,800 - this represented courses supported by both EMBCORE and Stichting funding. Only six institutes applied for this support.

We now have an absolute fortune of more than €100K to spend and we must do so before December 2004. We have already funded several courses but we need applications for more! Specialist courses would be nice, but we understand that introductory courses still play a very important role in the training for students of all countries. We offer to cover 50% of the course costs and there is no restriction on the funding, however each application is reviewed by the training committee who will exercise judgment on the necessity for the award. Come and join the regular band who use EMBnet money to support their training!

Fortunately the process for proposing a course is now made easy by a web interface. First of all we set up a list of available lecturers with their topic of interest. This can be very



Name	Domain of expertise	Languages	Institute
Terri Atwood +44 (0)61 275 5766 +44 (0)61 275 5082	Protein sequence analysis, (pattern recognition)	English	UMBER
<i>"I basically teach protein sequence analysis from the point of view of the availability of primary and composite sequence databases, and their search tools, moving on to secondary databases and their search tools, and the importance of using a broad-spectrum protocol in trying to get the best out of current sequence analysis resources and tools. My teaching is Web-based and to an extent flies in the face of package-based approaches (e.g., I don't use GCG)." Look for more interesting information at Terri's practical page</i>			
Martin J. Bishop +44 1 223 484525 +44 1 223 484512	Human Genome Project, Genome Analysis, DNA Sequence Analysis, Phylogenetic Reconstruction	English, Italian	UK Human Genome Mapping Project
Laurent Falquet +4121 692 5954 +4121 692 5945	Sequence Analysis, Protein domains and function, EMBOS, Perl <i>Former member of the PROSITE team.</i>	French, English	EMBL Switzerland - SIB
Volker Flegel +41216925954 +41216925945	Sequence Analysis, C/C++/Fortran	French, German, English	EMBL Switzerland - SIB
Valérie Ledent 32-2-6509978 32-2-6509998	Sequence Analysis; EMBOS; Searching databases; Pairwise sequence comparison; Sequence databank searching by similarity; Basic introduction to the 3D structure of proteins	English	Belgian EMBnet Node
Jack A.M. Leunissen +31 317 484074 +31 317 485384	Sequence Analysis, Phylogeny, GCG/EGCG, SRS, GDB/OMIM, UNIX/VMS	Dutch, English, German, very dim French	Centre for Molecular and Biomolecular Informatics (CBMI)

Figure 1. Lecturers list

Figure 2. Index page

useful when a course organiser is looking for a teacher in a particular domain (figure 1):
<http://www.embnet.org/php/lecturers.php>

Every node manager can propose a course using its authentication (username/password) to access our internal pages:
<http://www.embnet.org/php/index.php>

Please follow this rule:

- File a course proposal
- Wait for approval by ETPC
- Fill out the contract
- Run the course
- File the report for the course
- Send a formal reimbursement request

By clicking on the button "Create course proposal" (figure 2), you will be asked to fill a form (figure 3) and click the "Send course proposal to ETPC" button. After a while you will receive an email from the ETPC to inform

Figure 3. Course proposal form

Figure 4. Courses status page

you about the status of your proposal. Should you not receive this email, you can check the current status of your course on the courses page (figure 4):

<http://www.embnet.org/ET/courses.php>

Once your course is approved you must fill in the contract with either EMBCORE (EU countries) or Stichting (non-EU countries).
<http://www.embnet.org/EMBCORE/GrantManagementstructure.php>

You can now run your course, ask your participants to fill in the evaluation form
<http://www.embnet.org/ET/questionnaire.php>

Finally you must file a report (figure 5) using the web form accessible from the index page (figure 2) and send a reimbursement request to EMBCORE or Stichting manager.

Lisa Mullan & Laurent Falquet

Figure 5. Course report form

AGM03: EMBnet annual general meeting 2003 in Poland

The EMBnet community met in Warsaw beginning of last september for its 17th Annual General Meeting (AGM03).

The meeting was hosted by the Polish EMBnet node team and its node manager, Piotr Zielenkiewicz, did a great job. The meeting was held at the Institute of Biochemistry and Biophysics of the Polish Academy of Sciences.

The session was conducted by the chairman Jack Leunissen. The usual administrative procedure took place with a very dense discussion on the future of EMBnet organisation. Several proposals were discussed and the nodes agreed to organise workshops on different topics in 2004 (see list next page).

Elections:

Executive Board

Erik Bongcam-Rudloff (SE) chairman (new)
Martin Grabner (AT) secretary
Sandor Pongor (ICGEB) treasurer
Terri Attwood (UMBER) member



Education & Training PC

Martin J. Bishop (UK) chairman
Isabel Marques (PT) secretary
Lisa Mullan (UK) treasurer
Rob Harper (FI) member (new)

Technical Management PC

José R. Valverde (ES) chairman
Nils-Einar Eriksson (SE) secretary
Kristian Vlahovick (ICGEB) treasurer
Endre Barta (HU) member (new)

Publications & Public Relation PC

Gonçalo Guimarães Pereira (BR) chairman
Pedro Fernandes (PT) secretary
Laurent Falquet (CH) treasurer
Oscar Grau (AR) member (new)



Planned Activities for 2004

Workshops

January 29: EMBER - a European Multimedia Bioinformatics Educational Resource, Amsterdam NL.
attwood@bioinf.man.ac.uk

February 2-5: EMBOSSE, ICGEB Trieste IT.
kristian@icgeb.org

April 22-23: Videoconferencing systems, CSC Espoo FI.
eija.korpelainen@csc.fi

May 5-7: Regulatory sequence motif discovery, LCB Uppsala SE.
Erik.Bongcam@lcb.uu.se

May 27-28: Web-services, ETI NL.
marc@eti.uva.nl

August 30-31: BioMinT - Biological Text Mining Summer School, Geneva CH.
attwood@bioinf.man.ac.uk

September 16: Collaborative EMBnet workshop day. In connection with the AGM. Brussels, BE.
rherzog@dbm.ulb.ac.be

date to be announced

Federating the SRS servers within the EMBnet infrastructure, BE or UK.
rherzog@dbm.ulb.ac.be

Meetings

February 25-27: EMBnet Interim meeting, Volendam NL.
asfra@asfra.nl

September 16-19: Annual General Meeting (AGM), Brussels, BE.
rherzog@dbm.ulb.ac.be

For more detailed information, consult our web site: <http://www.embnet.org>

Announcements

Visitor at BEN

BEN (Belgium EMBnet Node) welcomes Viorica Ghita who will act as bioinformatics teacher within the EMBER project. She will be with us between November 1 2003 and January 31 2004.

CINEMA 5 beta release

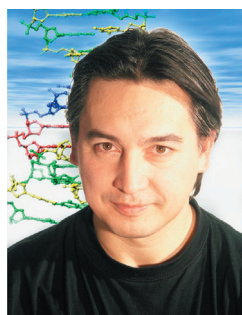
Cinema 5 (Colour INteractive Editor for Multiple Alignments) is an interactive visual tool for the interpretation and manipulation of protein and DNA sequences. Cinema 5 allows the user to interactively construct and edit and visualise multiple sequence alignments. It provides a highly intuitive, flexible and user-friendly interface with many features:

- Intuitive mouse-controlled interface for navigating and editing an alignment
- Optimised rendering for large sequences
- Multiple different alignments may be opened for editing simultaneously
- Many views can be created simultaneously onto a single alignment
- Support for many popular protein file formats
- Breakpoints allow aligned sections to be protected from edits elsewhere in the sequence
- Flexible zoom controls allow an alignment to be visualised at varying levels of detail
- Hierarchical grouping to simplify editing tasks
- Flexible sub-sequence searching allowing wildcard and fuzzy searching using customisable substitution matrices
- Advanced editing tools, such as hierarchical sequence groups & search-guided alignment

Cinema 5 is available for Linux, Windows and Mac OS X, and can be downloaded for free from:

<http://aig.cs.man.ac.uk/utopia/download>

A guide on how to install Bioinformatics tools on MacOSX (part III)



Erik Bongcam-Rudloff

Assistant Professor
Swedish University of Agricultural Sciences
The Linnaeus Centre for Bioinformatics



Anders Nister

Project Student

Introduction

Today, the desktop of a typical researcher in Life Sciences has two or more computers: a UNIX workstation for research and analysis applications and a PC or a Macintosh for mainstream applications such as writing, Web-surfing or creating Power-point presentations. With the arrival of MacOSX and the new 64 bit G5 computers, a researcher can now run both productivity applications and high-performance UNIX applications on a single system.

Apple is not known for server or super computer production but these past months have changed all that. As an example: Virginia Tech has built one of the Top Supercomputers in the world from approximately 1,100 of Apple's new Dual 2.0GHz PowerMac G5s.

This cluster achieved the #3 spot according to: <http://www.top500.org/>.

The Virginia Tech built their 10.2 Tflop/s cluster in a few weeks for \$5.2 million, while Japan's 35.9 Tflop/s Earth Simulator, currently the fastest system in the world, costs around \$350 million.

Introduction to Panther

During the development of our bioinformatics tutorials and packages Apple released a major update of its OS and presented a new computer architecture. The new Mac OS

version number is 10.3 and its code name is Panther.



Figure 1. A schematic picture of MacOSX Panther system architecture.

Panther comes on four CDs: three discs for installing the OS and one disc containing the developer tools. For our purposes be sure to install the developer tools.

Panther is now synchronized with FreeBSD 5, and has adopted gcc 3.3 as its default compiler. The new version of gcc gives better optimized code for the PowerPC architecture. More importantly, it includes performance profiles for the G5 (PowerPC 970), including native 64-bit integer arithmetic as well as features like the G5's hardware square-root function. Worth mentioning is the fact that Panther was compiled using gcc 3.3 and and this produced more optimized code for the G5.

Panther improves the open-standard, UNIX-based foundation of MacOSX and now includes X11 for Mac OS X. Along comes an enhanced Terminal and updated open source projects such as Python, FreeBSD 5 commands and libraries. New is also that 10.3 has Perl 5.8.1RC3 and Java 1.4.1 as default. This makes our previous packages of X11 and Perl 5.8 obsolete for Panther.

G5

The Power Mac G5 is the first to bring a 64-bit architecture to the personal computer market. This new G5 is a new implementation of the PowerPC architecture, the G5 processor is based on IBM's 64-bit POWER4 processor. More practical and useful for bioinformatics applications, the Power Mac G5 can be configured with 8GB of RAM,

four times more than what is supported by a typical PC. The G5 also has a 1GHz frontside bus, an advanced system controller, fast 400MHz 128-bit memory and Serial ATA disks.

New package system

In previous issues of EMBnet.news we provided two releases of bioinformatics tools, namely Packages Version 0.9 and Packages Version 2.0.

Now we have created a single, uniform installation package installing the following bioinformatics programs/packages: EMBOSS, NCBI-blast, T-coffee, Staden, ClustalW1.8.2, wwwtools, wise, primer3, HMM 2.2g (Altivec optimized) and Phylip.

The installation also installs nedit, the QT libraries, kaptain., and other necessary libraries.

New in this release is that this installation does not need Fink (an Open Source project that adapts Unix software so that it compiles and runs on Mac OS X ("port" it)). The X11 (for Panther, optional for Jaguar) and Perl 5.8, are already installed in Panther. New also is that we adopted a new destination for all packages: `/usr/ebiotools`. This will make it easier for the user to remove the packages or to install our new releases. Another change is the inclusion of a Staden license, no need to apply for one (please read this information: <http://www.mrc-lmb.cam.ac.uk/pubseq/news.2003.html>).

The requirements for the package are MacOS X 10.2 (Jaguar) or 10.3 (Panther) and Apple's X11.app (beta 3 for Jaguar or version 1 for Panther). It also runs under XDarwin with XFree86 4.3.

How to install our new release:

- 1- Go to our site
<http://www.ebioinformatics.org/downloads/>
- 2- Download a single package, choose your MacOSX: Jaguar or Panther.
- 3- Unpack
- 4- Click twice on the icon
- 5- This action will start a visual install procedure, all Unix commands will run in the background.

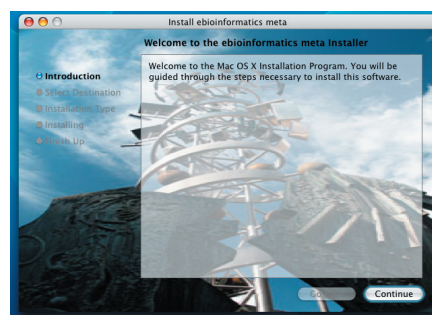


Figure 2. This is the window that starts the easy install.

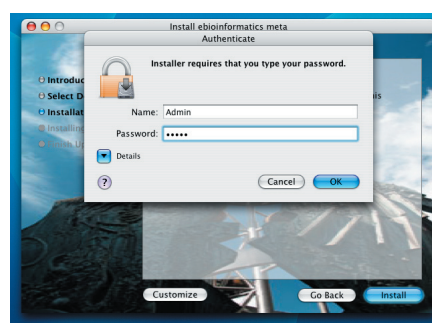


Figure 3. You must have administrator rights on your machine. Please fill in and press "install".

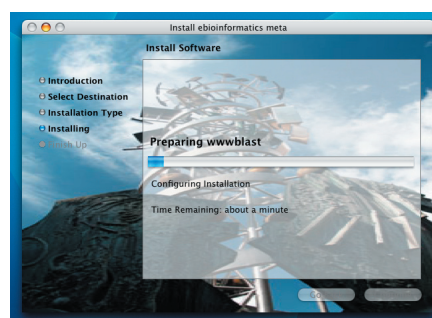


Figure 4. The installation starts

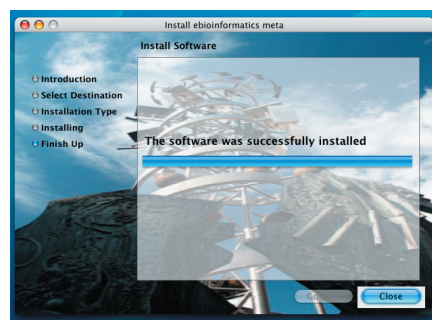


Figure 5. The installation will take around 15 minutes on a single 400 Ghz G4 or less than one minute on a dual 2Ghz G5.

Notes

You will need to read our previous tutorial (EMBnet.news 9.2) to make the wwwtools to work. First of all you must start your Macs web-sharing and then modify the Apache configuration file `httpd.conf` to be able to execute the web-blast scripts. Later on you should also configure your blast databases following the instruction in EMBnet.news 9.2. For addendum and updates (always on development): www.ebioinformatics.org. If you want to adapt your environment follow the instructions on EMBnet.news 9.2, (pp 16-22) and create a ".cshrc" file in your home directory.

Most packages work directly from the command line and the GUI through the X-Windows system (X11).

Blast comparison

To test the power of our new systems we performed a series of Basic Local Alignment Search Tool (blast) tests. The blast experiments were conducted on a G4 Xserve (Dual 1Ghz G4, 2GB RAM, MacOSX 10.2.8 (Jaguar) and a G5, dual 2Ghz G5, 4 GB RAM and 10.3.1 (Panther). The executable used was ncbi-blast 2.2.6 compiled on the individual machines. The databases were downloaded from EBI in blast format.

Blasts were run in a fashion similar to the Apple-genentech-blast described at:

<http://developer.apple.com/hardware/ve/acgresearch.html>

<http://developer.apple.com/hardware/ve/pdf/PerfData18Apr02.pdf>

With an increasing word size as the only variable parameter, the BLOSUM62 matrix and an e-value of 0.000001 was used. Three randomly selected human ests were used as query strings for all searches.

Plots depict the time taken for a blast-run with the word count W on the x-axis and time in seconds on the Y-axis. Databases plotted are nt, est, other_genomic and sts.

Our tests are only comparisons between different MacOSX versions and computers based on different PowerPC architectures. The reader will also wonder how this architecture compares with other systems.

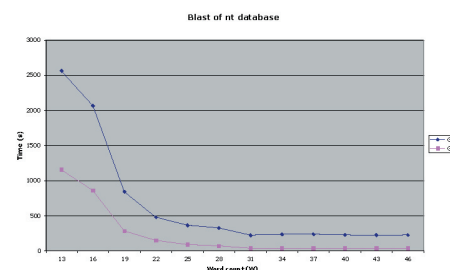


Figure 6. Comparison of blast runs on a Jaguar Xserve (G4) and a Panther G5. Parameter Word (x-axis) count plotted against run time (y-axis). The G5 has run times that range from more than twice as fast as the G4 for low Word count values, and up to 6,5 times as fast for higher word counts.

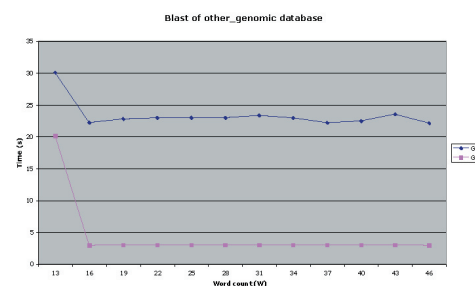


Figure 7. Comparison of blast runs on a Jaguar Xserve (G4) and a Panther G5. Parameter Word (x-axis) count plotted against run time (y-axis). The G5 is 4,5 times faster for all values of W .

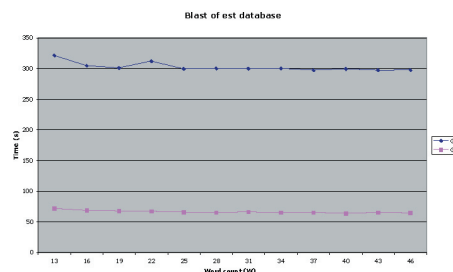


Figure 8. Comparison of blast runs on a Jaguar Xserve (G4) and a Panther G5. Parameter Word (x-axis) count plotted against run time (y-axis). The G5 is consistently about 8 times faster than the G4.

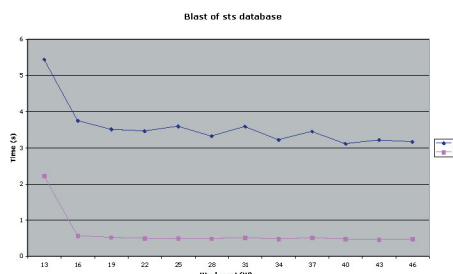


Figure 9. Comparison of blast runs on a Jaguar Xserve (G4) and a Panther G5. Parameter Word (x-axis) count plotted against run time (y-axis). The G5 is about 7,5 times as fast for the small sts database.

We did not have the time to do our own tests and we can only present a figure showing an experiment done by Apple. This is shown for your information only, we are planning to run our own comparisons in the future. To demonstrate the performance advantages of the Power Mac G5 for processor-intensive scientific analysis Apple used blast. All Power Mac G5 systems were tested using A/G BLAST. 2 The Dell Dimension XPS, Alienware Aurora and Dell Precision 650 ran NCBI BLAST on Red Hat Linux.

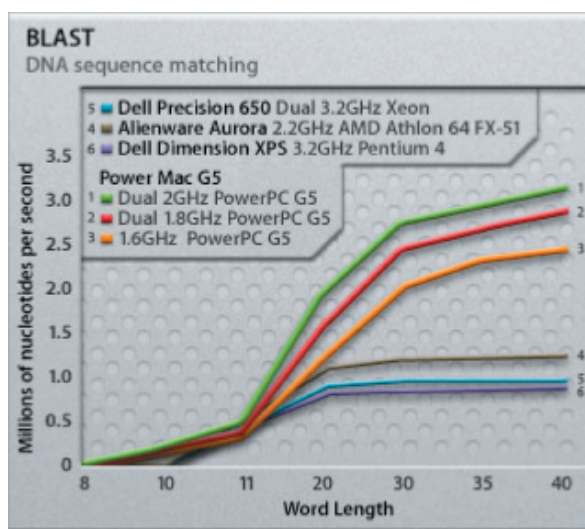


Figure 10. In common searches using a word size of more than 11 the Power Mac G5 far outperformed the Pentium 4-, dual Xeon- and Athlon 64 FX-51-based systems

Conclusion

The improvements in the last MacOSX version make an upgrade to Panther a significant performance boost. Together with our new installation package you will easily build a "Bioinformatics Workbench" on your desktop machine. Our new approach will also make it easy to enhance and upgrade the whole system. We are at the moment compiling a list of Open Source bioinformatics tools to add as modules to our system.

References

- 1 <http://www.top500.org/>
- 2 http://www.apple.com/downloads/macosx/math_science/applegenentechblast.html
- 3 <http://www.ebioinformatics.org>

Addendum

Free/Open Source Packages for the MacOSX (a preliminary list):

- EMBOSS, ClustalW, Staden, T-coffee, Fasta, HMMer, NCBI Toolkit, Phylip, Primer3, Wise2 (Protein comparison):

www.ebioinformatics.org

- Amber:

amber.scripps.edu

- Apollo Genome Browser:

www.ensembl.org/apollo

- AutoDock Toolkit:

www.scripps.edu/pub/olson-web/doc/autodock

- BASE:

base.thep.lu.se

- BioPerl, BioJava, BioXML, BioCorba, BioRuby, BioPython

open-bio.org

- BLAT:

contact_kent@soe.ucsc.edu

- CINEMA 5:

aig.cs.man.ac.uk/utopia/download

- Cn3D:

www.ncbi.nlm.nih.gov/Structure/CN3D/cn3dinstall.shtml

- Glimmer2, MUMmer

www.tigr.org/software

- GROMACS:

www.gromacs.org/

- MacGDE 1.3:

www.msu.edu/~lintone/macgde

- MolMol, O:

www.msg.ucsf.edu/local/programs/MacOSX/MACOSX_Downloads.htm

- OmniGene/OmniView

omnigene.sourceforge.net

- Phred-Phrap-Consed:

www.phrap.org

- PyMOL:

fink.sourceforge.net/pdb/package.php/pymol

- Python Molecule Viewer:

www.scripps.edu/~sanner/python/pmv/webpmv.html

- RasMOL:

proclus.tripod.com/darwin/

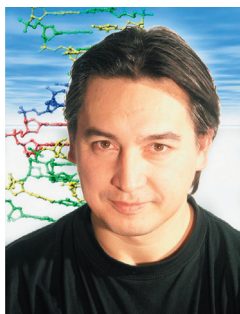
- Scilab:

scilabsoft.inria.fr

- VMD:

www.ks.uiuc.edu/Research/vmd

EMBnet E-meeting portal



Erik Bongcam-Rudloff

Assistant Professor
Swedish University of
Agricultural Sciences
The Linnaeus Centre for
Bioinformatics

The EMBnet organization is more global than ever. We have an increasing need for fast and personal contacts as a complement to our usual meetings (AGM, PC-meetings, courses and workshops). The telephone and E-mail are useful, but we need ways to meet face to face with node managers and PC-members (to see the facial reactions)—without spending all our time and money on airlines and hotels!

The solution we were looking for is videoconferencing. Using the Internet for videoconferencing isn't a new concept; different projects have tried for years to create

such applications. But videoconferencing applications from five years ago, like CUseeMe and early versions of Microsoft NetMeeting, proved too difficult to set up for mainstream use, and they tended to work badly. Many people at EMBnet have had bad experiences with these media.

We did a quick research of newly developed software and we found one that fulfilled our pre-requisites. Easy of use, clients for all major OS (Linux, Windows, MacOSX and Solaris), encrypted communication and good sound and video quality. The software we found is Marratech. (www.marratech.se)

The next-most-important consideration was whether the application works from behind a firewall, after testing we can state that Marratech does without problems.

EMBnet has now a Video-conferencing portal server (Marratech E-meeting Portal server) and the only thing an EMBnet member need to connect to an E-meeting is a Web-camera, a headset and client software.

The Marratech Work Environment consists of the Marratech E-meeting Portal server and the free Marratech Pro client software.

The steps to connect to an E-meeting are:

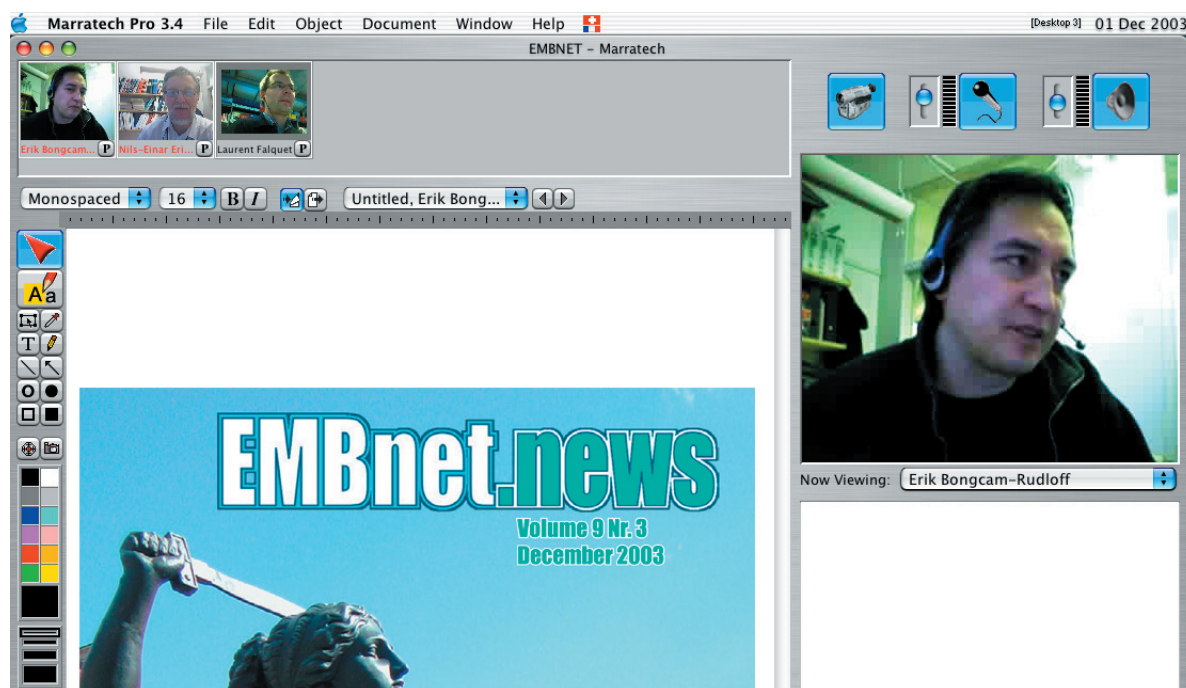


Figure 1 Shows a typical session with 3 participants and one shared document page. This meeting place can easily accommodate 20 simultaneous participants.

1- Download the Marratech client at:
<http://www.marratech.se/download/>
 2- Install Marratech Pro on your computer (Windows, Mac, Sun or Linux), connect a headset and a web camera (it is possible to only communicate with sound!). Start Marratech or point your web browser to the EMBnet E-meeting Portal link. Marratech Pro will then start and let you securely interact with other EMBnet node managers or PC-members. You will need a username and a password for the meeting room. That's all!

Embnet's Executive Board (EB) has made use of the system and we can report that three virtual meetings have taken place during the past 3 months. The system works rather well in our hands.

The E-meeting room allows you to share documents as: Power-point presentations, PDF-files, pictures, text files etc. All the participants can move and control the meeting-room, this of course requires some meeting discipline among users.

I have personally used the Marratech system

on The Linnaeus Centre for Bioinformatics (LCB) own server installation. I distributed my classroom over the Internet by having students join via LCB's meeting portal. Before the class started, I imported my PDF-files and other material to the meeting room's electronic whiteboard. When the class begun, all the students had the material pre-loaded into their Marratech Pro clients; changing slides took then a fraction of a second.

This environment enables truly interactive e-learning, and students could ask questions (Marratech uses proprietary algorithms for high quality sound) and use the whiteboard to show me problems.

The system can also be used to give remote seminars, the teacher and the audience connects then to the portal (see Figure 2). The seminar room needs a computer beam and good speakers for the audience to see and hear the teachers lecture. To make the experience truly interactive is a good idea to have a couple of microphones to handle to the audience. I have used this system several

times in Sweden and from Sweden to Norway. The first time it is rather strange to look into the camera and not to the screen showing the audience, but after a while everyone can learn to act as a pro. All sessions can be recorded to the hard disk making it possible for the students to review the lecture afterwards.

For more specific notes about recommended hardware please visit:

<http://www.marratech.se/support.shtml>

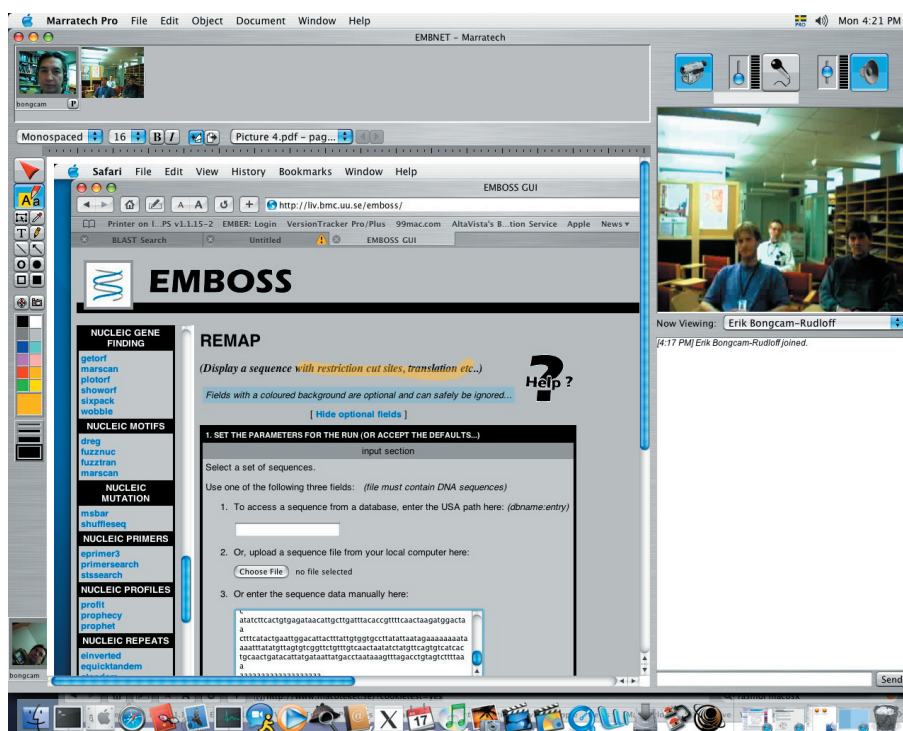


Figure 2 Depicts a typical lecture situation in this case between Uppsala, Sweden and Bergen, Norway.

IN VINO VITA?

By Vivienne Baillie Gerritsen

For many, a prolonged – but of course healthy – life is a tempting prospect. Especially, in a day and age where there seems so little time to fulfill our ever-growing aspirations and the gift of an extra few hours a day is just not on. Researchers have been giving a lot of their time to the problem of longevity since the 1930s when it was first discovered that calorie restriction actually lengthened life expectation in mammals. It took many more years of research before scientists got a glimpse of the molecular pathway underlying such a process. And, for the time being, it really is just a glimpse but an encouraging one. A family of proteins, known as sirtuins (sir-too-ins) or SIRS, an abbreviation of Silent Information Regulators, known to exist from bacteria to humans, seem to have a definite role in the ageing process of cells, and hence the ageing process of organisms.



Fig.1 Bacchus, Hermitage, St. Petersburg
Peter Paul Rubens (1610-1695)

It is now clear that the ageing pathway in different organisms is different. In yeast, for example, an accumulation of what is known as toxic extrachromosomal rDNA circles (ERCs) is a cause of cell senescence. Whilst in worms, interference in the insulin signaling pathway have an effect in the organism's longevity. The exciting discovery is that sirtuins have a direct effect in these different ageing processes. They may act in different molecular ageing pathways but their mode of action is the same. Sirtuins are NAD-dependent protein deacetylases. What they do is remove acetyl groups from specific target

proteins. Their function however depends on the intracellular concentration of nicotinamide adenine dinucleotide (NAD), which in the oxidized form greatly enhances sirtuin activity. In yeast for example, genomic instability and hence the accumulation of ERCs, stakes its fate. Sirtuins can brake this fate by acting on certain histones. Through deacetylation of the said histones, chromatin structure is modified, the lecture of rDNA disrupted and the formation of ERCs hindered.

What kind of event would delay cell senescence? Calorie restriction for one. It is thought that calorie restriction either shifts the ratio of oxidized/reduced NAD or alters its level, thereby having an effect on sirtuin activity. How? Well, a number of proteins which are also NAD-dependent may become less active and release their NAD which is then free to activate the sirtuins. Another hypothesis for an increase in NAD is a metabolic shift to respiration. Indeed, when glucose is limiting, respiration is preferred. An increase in respiration yields a higher rate of electron transport thereby increasing the overall cellular NAD/NADH ratio. Either way, an increase in NAD activates sirtuin activity and as a result, in the event of environmental stress, such as less food but also fungal infection for instance, organisms would be given a better chance to survive.

So, sirtuins do seem to have an overall role as regulators of programmed cell death and cell differentiation. As we have seen, they are more heavily summoned during environmental stress. In the quest for eternal youth and health, scientists set about to find out if they could pin down a molecule or two which could trigger off sirtuin activity, instead of just counting on calorie restriction. A number of chemical compounds were examined amongst which a few polyphenols, which are metabolites found in plants. The most potent activator turned out to be resveratrol, a plant metabolite found in the skins of red and white grapes for instance.... and in the eyes of the convivial 'grapes' spells 'wine'.... Another plant metabolite, quercetin, is also an activator of sirtuins, but to a lesser degree and is found in olive oil for example. Both compounds are synthesized in response to environmental stress and are already known for their health

benefits, namely their antioxidant properties, their capacity to lower cholesterol levels and decrease blood clot formation.

There is a theory that, when there is a lack of nutrients, plants may have the power to 'say' so to other organisms via resveratrol for example. Various animals feeding on plants would assimilate the increased ratio of resveratrol and thereby transfer the warning message to their own organism; i.e. that reproduction has to be halted and survival means undertaken. However, there is no reason not to believe that organisms other than plants do not have their own sirtuin activity enhancers which have simply not been identified yet. Time will tell.

So should we all take to drinking a glass of red wine or two per day? Or adding olive oil to our diet? Certainly, statistics have shown that the French, who are generous drinkers of red wine, suffer less from the side effects of overweight than do the Americans for example. A Mediterranean diet, which includes not only red wine but also ample amounts of olive oil as well as fresh fruit and vegetables, is also statistically far kinder to humans than a more Western diet of coke, hamburgers and chips. Let us make a finer comparison for the gourmet readers: wines produced in cooler regions or areas where grapes are more prone to disease often have more resveratrol. Hence Pinot Noir, for instance, has rather high doses of the chemical while a

Cabernet-Sauvignon is not so rich in the compound. What is more, resveratrol is unstable on exposure to air and goes off within 24 hours. Would it be wiser then to drink a bottle a day?

Well no. In fact, it has been shown that high doses of resveratrol stunts its effect on sirtuins – so no point in overdrinking. Nothing has been proved yet. No one can say whether even one glass of wine a day keeps the doctor away or not. Furthermore, in human cells, resveratrol – via sirtuin activity – ultimately deactivates a tumor suppressor, thus fueling concerns that increasing resveratrol levels may predispose humans to cancer. Though a number of scientists disagree since calorie restriction *per se* has never proved to do so. Much more has to be done, sirtuin's 3D structure for one, before any conclusions can be made. Naturally, a lot of money could be involved in this kind of discovery. Yet another miracle diet: eat your cake and swill it down with a glass of wine enriched with resveratrol. Just imagine. However scientists insist that the findings are not so much to prolong life than to avoid the bothersome effects of ageing such as neurodegeneration, carcinogenesis and atherosclerosis. Well, the message is always the same isn't it? Just eat a healthy varied diet: some wine, some olive oil, fresh fruit and vegetables, cereals and even chocolate. Just don't eat too much of it and you should be all right.

Cross-references to Swiss-Prot

P06700: *Saccharomyces cerevisiae* (Baker's yeast) NAD-dependent histone deacetylase SIR2

References

1. Finkel T.
Ageing: a toast to long life
Nature 425:132-133(2003).
PMID: 12968159
2. Howitz K.T., Bitterman K.J., Cohen H.Y., Lamming D.W., Lavu S., Wood J.G., Zipkin R.E., Chung P., Kisieleski A., Zhang L.-L., Scherer B., Sinclair D.A.
Small molecule activators of sirtuins extend *Saccharomyces cerevisiae* lifespan
Nature 425:191-196(2003).
PMID: 12939617
3. Hall S.S.
In vino vitalis? Compounds activate life-extending genes
Science 301:1165-1165(2003).
PMID: 12947168

National Nodes

Argentina

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

Australia

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

Austria

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

Belgium

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

Brasil

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

Canada

Canadian Bioinformatics Resource, National Research
Council Canada, Institute for Marine Biosciences,
Email: manager@cbr.nrc.ca
Tel: +1-902-426 7310 Fax: +1-902-426 9413
<http://www.ca.embnet.org>

Chile

Dr. Ricardo Baeza-Yates
Dept. of Computer Science, Santiago,
Email: rbaeza@dcc.uchile.cl
Tel: N/A
<http://www.embnet.cl>

China

Jingchu Luo
Room 303, Exchange Centre, Peking University
Email: luojc@cbi.pku.edu.cn
Tel: +86-10-6275 9001
<http://www.cbi.pku.edu.cn>

Colombia

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

Cuba

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

Denmark

Hans Ullitz-Moeller
BioBase, University of Aarhus
Email: hum@biobase.dk
Tel: +45-86-13 9788
<http://www.dk.embnet.org>

Finland

Kimmo Mattila
CSC, Espoo
Email: Kimmo.Mattila@csc.fi
Tel: +358 9 457 2708
<http://www.fi.embnet.org>

France

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

Germany

Sandor Suhai
EMBnet node at the German Cancer Research Center
Department of Molecular Biophysics (H0200)
Email: genome@dkfz.de
Tel: +49-6221-422 342 Fax: +49-6221-422 333
<http://www.de.embnet.org>

Greece

Babis Savakis
Institute of Molecular Biology and Biotechnology
Heraklion, Crete
Email: savakis@nefeli.imbb.forth.gr
Tel: +30-81-391 114 Fax: +30-81-391 104
<http://www.imbb.forth.gr>

Hungary

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

India

H.A.Nagarajaram
Laboratory of Computational Biology & Bioinformatics
facility, Centre for DNA Fingerprinting and Diagnostics
(CDFD), Hyderabad
Email: han@www.cdfd.org.in
Tel: +91 40 7155607 / 7151344 ext:1206
Fax : +9140 7155479
<http://www.in.embnet.org>

Israel

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

Italy

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

Mexico

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

The Netherlands

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

Norway

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

Poland

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

Portugal

Pedro Fernandes
Instituto Gulbenkian de Ciencia
Unidade de Bioinformatica
2781-901 OEIRAS
Email: pfern@igc.gulbenkian.pt
Tel: +351 214407912 Fax: +351 2144079070
<http://www.pt.embnet.org>

Russia

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

Slovakia

Lubos Klucar
Institute of Molecular Biology SAS Bratislava
Email: klucar@embnet.sk
Tel: +421 7 5941 2284
<http://www.sk.embnet.org>

South Africa

Ruediger Braeuning
SANBI, University of the Western Cape, Bellville
Email: ruediger@sanbi.ac.za
Tel: +27 (0)21 9593645
<http://www.za.embnet.org>

Spain

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

Sweden

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

Switzerland

Laurent Falquet
Swiss Institute of Bioinformatics, CH-1066 Epalinges/
Lausanne
Email: Laurent.Falquet@isb-sib.ch
Tel: +41 (21) 692 5954 Fax: +41 (21) 692 5945
<http://www.ch.embnet.org>

United Kingdom

Alan Bleasby
UK MRC HGMP Resource Centre, Hinxton, Cambridge
Email: ableasby@embnet.org
Tel: +44 (0) 1223 494535
<http://www.uk.embnet.org>

Specialist Nodes

EBI

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

ETI

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

EU

Dr Bernard Mulligan
DG Research - European Commission
Brussels BELGIUM
Email: bernard.mulligan@cec.eu.int

ICGEB

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

LION Bioscience

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

MIPS

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

UMBER

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

TECH-MGR

Email: tech-mgr@embnet.org
The team gives support to EMBnet nodes and helps
them with maintenance and troubleshooting.
The team is formed of experienced system administrators
and programmers who ensure the availability of local
services for all EMBnet users.



ISSN 1023-4144

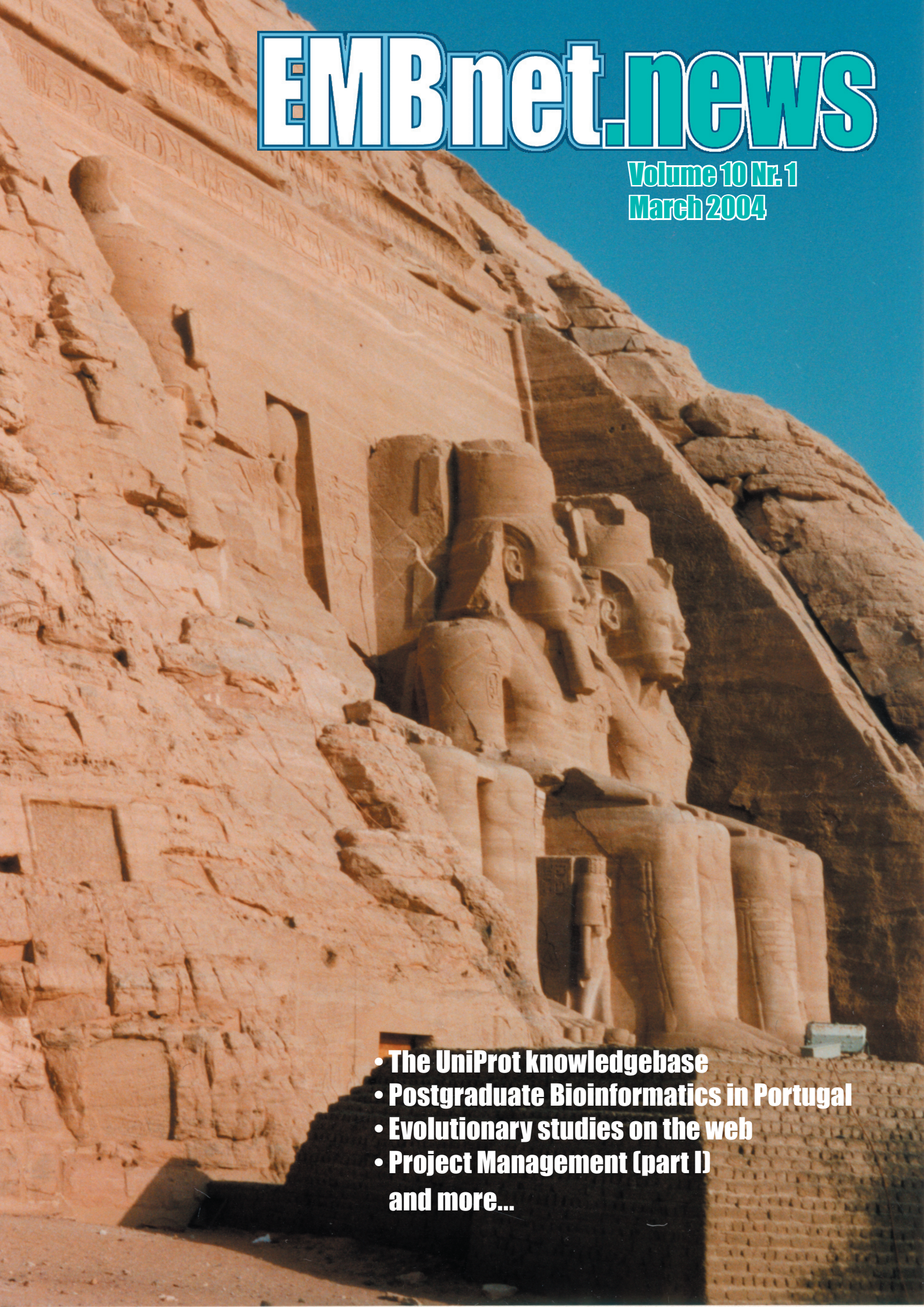
Dear reader,

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 (ISSN 1023-4144). You
can get them by anonymous ftp from:
the EMBnet organisation Web site
<http://www.embnet.org/download/embnetnews>
the Belgian EMBnet node
<ftp://ftp.be.embnet.org/pub/embnet.news>
the UK EMBnet node
<ftp://ftp.uk.embnet.org/pub/embnet.news>
the EBI EMBnet node
<ftp://ftp.ebi.ac.uk/pub/embnet.news>

Submission deadline for next
issues:

February 29, 2004
Mai 31, 2004
August 15, 2004
October 31, 2004

The background of the cover is a photograph of the Great Sphinx of Giza in Egypt. The Sphinx is carved into the sandstone cliff face, with its head turned towards the right. The rock surface is covered in hieroglyphs. The sky is a clear, bright blue.

EMBnet.news

Volume 10 Nr. 1

March 2004

- The UniProt knowledgebase
 - Postgraduate Bioinformatics in Portugal
 - Evolutionary studies on the web
 - Project Management (part I)
- and more...

Editorial

The creation of the UniProt knowledgebase in December 2003 illustrates a unification of forces. Just as in ancient Egypt the strength and skills of many workers were combined to build one of the most impressive human creations, researchers in today's proteomics field are working together to achieve a common goal.

While UniProt may not last a millennium (we must stay humble), other joint projects already proved to be successful (e.g., InterPro, and of course, EMBnet!) the impact of these collaborative projects might incite other to follow their example and join forces. What about a UniRNA, UniMicroarray, Uni*?

In this issue you will also discover how to use the PHYLIP package on the web, how to manage a project, and a report of the Portuguese Post graduate programme in Bioinformatics.

The editorial board: Erik Bongcam-Rudloff, Laurent Falquet, Pedro Fernandes, Oscar Grau, Gonçalo Guimaraes Pereira

Protein Spotlight

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

Publisher:

EMBnet Administration Office
c/o Jack Leunissen
Dept. of Genome Informatics
Wageningen UR, Dreijenlaan 3
6703 HA Wageningen
The Netherlands
Email: Jack.Leunissen@wur.nl
Tel: +31 317 484074

Contents

Editorial	2
UniProt the universal protein knowledgebase	3
Evolutionary studies on the web	6
A guide book for Microarray	10
Postgraduate in Portugal	12
Project management (Part I)	17
MacOSX and Bioinformatics (short report)	21
Protein spotlight	44
National nodes	24
Specialist nodes	26

Editorial Board:

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

Email: erik.bongcam@bmc.uu.se

Tel: +46-18-4716696

Fax: +46-18-4714525

Laurent Falquet, SIB, Lausanne. CH

Email: laurent.falquet@isb-sib.ch

Tel: +41-216925954

Fax: +41-216925945

Pedro Fernandes, Instituto Gulbenkian. PT

Email: pfern@igc.gulbenkian.pt

Tel: +315-214407912

Fax: +315-214407970

Oscar Grau, IBBM, AR

Email: grau@biol.unlp.edu.ar

Tel: +54-221-4259223

Fax: +54-221-4259223

Gonçalo Guimaraes Pereira, UNICAMP. BR

Email: goncalo@unicamp.br

Tel: +55-19-37886237/6238

Fax: +55-19-37886235

Cover picture: Two statues of Ramsès II of 20 m high at the entrance of Abu Simbel temple, September 1997
[© Monique Zahn & Marta Puente]

A great achievement within the realm of Proteomics: UniProt, the universal protein knowledgebase



Vivienne Baillie-Gerritsen
Science
Communication
Swiss Institute of
Bioinformatics
Swiss-Prot Group
CMU - 1, rue M.-Servet
CH-1211 Geneva 4
Switzerland

In December 2003 a press release, which informed the world at large of the birth of yet another database in the Life Sciences, was enthusiastically launched. The news may have sounded tedious to some but to a great number of life scientists, whose research is based on proteins and the fascinating world these macromolecules revolve in, the news was paramount. Here is a database – no better still a knowledgebase – which is a central hub not only for all existing public information on protein sequences but also a number of tools for their analysis and a colossal amount of cross-references for all kinds of additional knowledge. And to which researchers can refer for free.

VSMGLDAVDE SSMTGSFSGS NAQTSTEEVS QDSTDIMALL DNNMLGSMGD
TAAASLTFF TKRANWVEEELRDFLQI ANNVPGAGPL PAGPFAQMNL
KLLIHDDFVLELSSGNARFYPSILNVTGY SVEIQDIFL
NAPFQA FMAPYTKNELLSEFJHFFETVGHAI AGSKFAPNPN
QSGERSDVT FSEGTGSSN FFPSTHGL MELRREEEA ENDEAQKQWM

UniProt

the Universal protein resource

PIR-PSD, the Protein International Resource protein sequence database, was the very first protein sequence database to exist. It was developed by the late Margaret O. Dayhoff (1925-1983), a pioneer in bioinformatics, and existed in book form between 1965 and 1978 as the 'Atlas of Protein Sequence and Structure' before continuing its existence within the entrails of a computer. Dayhoff

has actually been credited as one of the founders of the field of Bioinformatics. The Atlas' first edition held information on barely 65 different protein sequences.

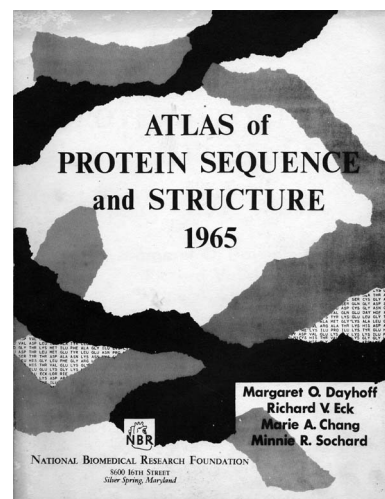


Figure 1 Cover of the Atlas of Protein Sequence and Structure 1965

Then came Swiss-Prot in 1986, a protein sequence database developed by Dr Amos Bairoch in Geneva. Why start a second one? Whilst working on his thesis, Bairoch felt the need not only to gather protein sequences but also to gather information related to each protein sequence – something he felt lacked in Dayhoff's database. Annotation was born... The first Swiss-Prot release boasted a little less than 4000 annotated protein sequence entries. Naturally, Swiss-Prot – like all databases – bounced along with the evolution of computer facilities; first stored on tapes, then disks and CD-ROMs, Swiss-Prot became available over the internet with the World Wide Web facilities. Today the Swiss-Prot knowledgebase prides itself with almost 150'000 entries and is developed by the Swiss-Prot group of the Swiss Institute of Bioinformatics in Geneva and the Swiss-Prot group of the European Bioinformatics Institute in Hinxton, England.

Swiss-Prot is renowned for its manual annotation. However, where something manual is involved, so are humans and humans do things at a human speed. In the past decade, the amount of biological information at hand has simply soared with

growing improvements in biotechnology. Gene sequences, and hence protein sequences, are churned out faster than you can swear. Manual annotation simply cannot keep up with the pace. That's why an additional database which complements Swiss-Prot, TrEMBL, saw the light of day. TrEMBL is a database of protein sequences which are automatically annotated and await further manual annotation by the Swiss-Prot team. It is an offshoot of the EMBL-Bank, a database of gene sequences, developed by the European Molecular Biology Laboratory whose main office is in Heidelberg, Germany. Currently, TrEMBL sports over one million entries.

So, here we have three large and distinct protein sequence databases: PIR-PDS, Swiss-Prot and TrEMBL, in addition to the multitude of more specialised protein sequence databases. Besides TrEMBL, which is the anteroom for protein sequences awaiting manual annotation, PIR-PDS and Swiss-Prot fulfilled the same kind of role. And it soon became apparent that it would be a wise step to merge the two in order to offer an even more comprehensive and non-redundant database of protein sequences to the scientific community. In 2002, the National Institutes of Health donated \$15 million, over a period of three years, to do just that. In December 2003, the new merged database, UniProt – the Universal Protein knowledgebase – was completed and went on-line. What happened is that all the suitable PIR-PDS entries which were missing in Swiss-Prot + TrEMBL were incorporated into the latter, and cross-references are made to

the original PIR entries for easy tracking back to PIR-PDS. Hence, UniProt is a bloated version of Swiss-Prot + TrEMBL. It sports two databases: one which is manually annotated and one which is only automatically annotated. The two continue to be called Swiss-Prot and TrEMBL, respectively.

Besides the UniProt knowledgebase, two other resources are an integral part of the Universal Protein Resource: UniProt Archive (or UniParc) and UniProt NREF (or UniRef). UniParc is the most comprehensive, publicly accessible pool of protein sequences. Which would make Swiss-Prot + TrEMBL quite redundant then you are thinking. Well no. The same protein sequence can be found in many different databases, not to mention more than once in one database. What UniParc does is assign a single identifier to an identical protein sequence found in a number of different sources and provides cross-references to the source databases, so that a UniProt user is not only aware of



Figure 2 The UniProt Knowledgebase home page

the Swiss-Prot + TrEMBL protein sequence but also its existence within other databases. So, in effect, a UniParc entry is in no way similar to a UniProt entry: there is no annotation. It is a catalogue which informs you of all the various places to go for a given protein sequence. The third resource, UniRef, is itself made up of three sublayers: UniRef100, UniRef90 and UniRef50. The UniRef databases use recently developed procedures to combine closely related sequences into a single record. UniRef100 is a non-redundant version of all the sequences in the knowledgebase; UniRef90 collapses all the sequences that are at least 90% identical, or more, into a single record and UniRef50 collapses sequences that are at least 50% identical. UniRef50 not only speeds up searching significantly but also does not reduce the effectiveness of homology searching. Hence, the three UniRef sections give the user the opportunity to choose between either a fast search or a truly comprehensive one.

What are the aims of UniProt? UniProt provides the scientific community with a comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and query interfaces. What is more, the data is publicly available. As mentioned before, a UniProt entry is identical to the familiar Swiss-Prot entry. Besides the protein sequence, there are a number of comments made on a protein's function, post-translation modifications, subcellular location etc. as well as information on specific amino acids or domains within the sequence itself. Numerous cross-references are made to the literature from where information was extracted as well as to other databases and indeed to analysis tools which contribute to an ever-widening knowledge of a given protein sequence. So, we have certainly come a very far way from Margaret Dayhoff's 'Atlas of Protein Sequence and Structure' but without it protein sequence knowledgebases would not be where they are today.

Links

<http://www.expasy.uniprot.org/>
<http://www.ebi.ac.uk/uniprot/>

Bibliography

Apweiler R., Bairoch A., Wu C.H., Barker W.C., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M., Martin M.J., Natale D.A., O'Donovan C., Redaschi N., Yeh L.S.
UniProt: the Universal Protein knowledgebase
Nucleic Acids Res. 32:D115-D119(2004)
PMID : 14681372

Planned Activities for 2004

Workshops

April 19-20: Videoconferencing systems, CSC Espoo FI.
eija.korpelainen@csc.fi

May 6-7: Regulatory sequence motif discovery, LCB Uppsala SE.
Erik.Bongcam@lcb.uu.se

August 30-31: BioMinT - Biological Text Mining Summer School, Geneva CH.
attwood@bioinf.man.ac.uk

September 16: Collaborative EMBnet workshop day. In connection with the AGM. Brussels, BE.
rherzog@dbm.ulb.ac.be

date to be announced

Federating the SRS servers within the EMBnet infrastructure, BE.
rherzog@dbm.ulb.ac.be

Meetings

September 17-19: Annual General Meeting (AGM), Brussels, BE.
rherzog@dbm.ulb.ac.be

For more detailed information, consult our web site: <http://www.embnet.org/TM/workshops.php>

Evolutionary studies on the Web



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

The analysis of evolution is more than just an interesting subject or a curious oddity. Certainly, knowing when dinosaurs died is appealing to the public at large, though it has little repercussion on everyday life. However, the study of evolution encompasses much more than that and actually has an impact on everyday life when used appropriately: from allowing us to identify relevant genes to tracing the origin of an epidemics and how it changes, or just tracing a donor for a transplant.

It is therefore of the utmost importance, for those who must deal with evolutionary techniques that applying them and understanding their results becomes as easy and efficient as possible. Furthermore, as its applications extend, new collectives must deal with technologies that seem the most alien to them since they were developed for a different collective with different ideas, needs and culture in mind. Ease of use is a must if we want the society (and ourselves) to benefit from the advantages of evolutionary studies.

In this article we have a look at Web interfaces for PHYLIP, a package for comprehensive phylogenetic analysis developed by Joe Felsenstein (see <http://evolution.genetics.washington.edu/phylip.html>).

PHYLIP

PHYLIP is a free package of programs for evolutionary studies that has become a standard used in many institutions for a number of reasons:

First, **PHYLIP** is free. Compared to other classical phylogenetic analysis packages

(like e.g. PAUP) which are commercially distributed at very expensive costs, PHYLIP cost is certainly unbeatable; specially at academic institutions where funds are always at a stake, this supposes a serious advantage.

PHYLIP is open source: anybody may see how it works, amend and correct it if necessary and, what is more important, use it to develop new tools, methods and approaches. And so there are additional programs or enhanced versions of PHYLIP programs (like e.g. FastDNAML or PROTML) developed at other sites.

It is a comprehensive package including methods for most of the common phylogenetic analysis, and it certainly covers successfully most needs of the average and advanced user. While it is true that some methods are better represented than others, or that the philosophical tendencies of the authors in the interpretation of phylogeny are reflected in the package, one would need to go for a very expensive commercial package to get a more comprehensive integrated environment.

Programs in **PHYLIP** are well proven and tried. Most of them have been around for many years (well over one or two decades) and being open source have been scrutinized and corrected at large. Surely, new programs may have hidden bugs, but the wide professional user community and its very openness reduce them to a minimum.

On the minus side, **PHYLIP** lacks the niceties of a modern, graphical user interface, which draws aback many potential users, and might be somewhat slower than some of its commercial counterparts which -with current computers- is seldom significant.

WebPHYLIP

This is a very useful web interface developed exclusively for PHYLIP. Although taste and preferences widely vary, it still has some niceties missing in other interfaces, providing a really practical user environment.

WebPHYLIP (see <http://sdmc.lit.org.sg:8080/~lxzhang/phylip/>) was developed

in 1999 by A. Lim and L. Zhang. It is a simple yet easy to use and powerful interface to most PHYLIP programs. The only ones missing are the interactive programs (which are unnecessary on a Web environment), a few programs of very narrow application (DNAINVAR, CONTRAST and FACTOR) and the unsupported programs MAKEINF and PROTML. Shortly put, you get access to many more programs than you do with EMBOSS and hence get extended phylogenetic productivity.

To use **WebPHYLP** all you need is to install PHYLIP (from Joe's site) and the web interface (from WebPHYLP site). Installation of WebPHYLP is very easy and is described in the accompanying README file: it involves a configuration step using Perl and a building step with make and a C compiler. All in all, it is a lot simpler than the many-step installation required by EMBOSS and thus it is better if all you are interested in are PHYLIP phylogeny applications.

Once installed, you may proceed to use it. WebPHYLP divides the screen in three parts: a hierarchically organized menu on the left, and two sections on the right: one above with the help needed to use any selected program, and one below with a form to fill in

your experimental data (Figure 1).

The menu on the left has been organized so it is easy to locate the appropriate programs for your data and analysis type. It starts with a section on conversion tools so you may import data from other programs (like Clustal, Paup or GCG), and goes on to provide options for distance computations, data sampling and phylogeny methods for DNA, Proteins, Restriction sites, Gene frequencies, 0-1 data and distance matrix, as well as tools for analysing and plotting trees.

As you navigate options towards programs you may request at any time additional help (to be shown on the right, upside) or run the program (with the form on the bottom right). This allows for a very comfortable work method, where you may have both the program and its description and exhaustive help at the same time in your screen. You may even resize the different portions to suit your needs.

This kind of layout, when you are dealing with complex analysis, is a really useful life-saver. Sure enough you may get help with other web interfaces, on a separate pop-up window, but for the average user with not-so-big screens this is often more of an annoyance than a help and results in people avoiding documentation (and making mistakes). WebPHYLP layout makes it easy to

Figure 1. WebPhylip example

The screenshot shows the WebPhylip web interface. On the left is a menu with the following items: 1. Parsimony (with links for Help and Run), 2. Parsimony + Branch&Bound (with links for Help and Run), 3. Compatibility (with links for Help and Run), 4. Max. Likeli. (with links for Help and Run), 5. Max. Likeli. with mol. clock (with links for Help and Run), Do consensus, tree editor, Draw trees, and Back to Menu. The main content area is titled 'DNAPENNY - Branch and bound to find all most parsimonious trees for nucleic acid sequence parsimony criteria'. It contains a paragraph of text about the program, followed by a section for 'Use ordinary parsimony' or 'threshold parsimony?' with an input field for 'Input threshold value (larger than 1):' set to 2.0. Below this is a 'Simple branch and bound?' section with a 'Yes' dropdown and a text input for 'How many groups of 100 trees?' set to 1, with a note 'After 10 trees, report progress of run.' At the bottom, there is an 'Input sequence type?' dropdown set to 'Interleaved', a 'Number of data sets:' input set to 1, a 'Use previous data set?' dropdown set to 'No', and a link for 'Formatted Input Sequences: an example'.

keep everything at sight, which most people find more convenient and practical.

PHYLIP and EMBOSS

EMBOSS, the European Molecular Biology Open Software Suite (see <http://www.emboss.org>), developed within EMBnet to provide an integrated environment for molecular biological studies, has facilities for a wide number of sequence analysis tasks, and being free and open source as well, has become widely popular too. EMBOSS does not cover everything, most notably it lacks in molecular evolution tools.

Not to worry though: EMBOSS provides a mechanism for other application to be integrated within, and since there is already a good, well proven, free and comprehensive package for evolution, it made more sense to integrate it than to build a new one. In this way, PHYLIP programs have been integrated within EMBOSS as an EMBASSY application (an external application integrated with EMBOSS).

Thus, to be able to use PHYLIP within EMBOSS you must get and install all, PHYLIP (from Joe Felsenstein's site), EMBOSS and the "EMBASSY" interface between both (from EMBOSS site). In addition, you need to install the WWW user interface of your choice (e.g. wEMBOSS, W2H, SRS...).

Once integrated, any application (and PHYLIP programs are not an exception) is dealt with as another EMBOSS application, and therefore, all EMBOSS user interfaces (Jemboss, wEMBOSS, W2H, SRS) will provide a seamlessly integrated access to PHYLIP applications (Figure 2).

The advantages of using PHYLIP within EMBOSS are many, but mainly they consist of taking advantage of an integrated environment where sharing sequences, analysis and results between applications is fully transparent.

PISE

We can't ignore the "interface of interfaces", **PISE**. Developed by Catherine Lethondal

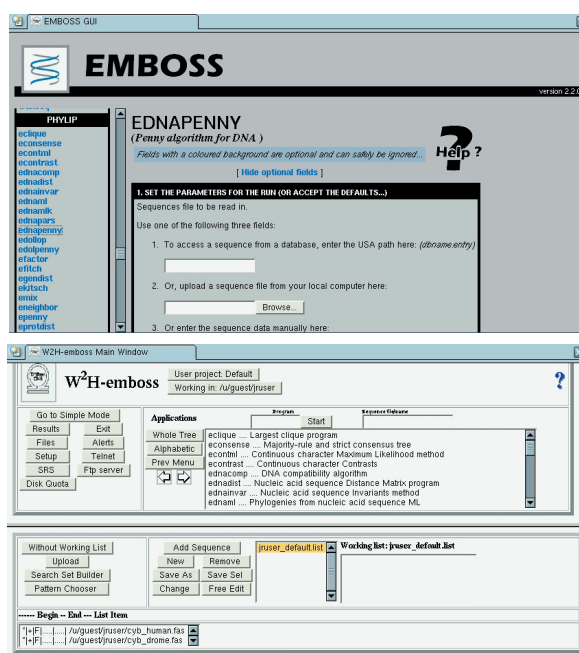


Figure 2 Examples of EMBOSS interfaces

at the Pasteur Institute (see <http://www.pasteur.fr/recherche/unites/sis/Pise/>) this is a generic framework for building web interfaces to command line applications

Almost any application may be used with PISE after it has been appropriately described to the system using an XML schema. There are already schemas provided for many applications (EMBOSS and PHYLIP among them) and therefore normally you do not need to take care of these details yourself, just use the provided definitions. In a sense, you may include PISE among the above mentioned EMBOSS interfaces, although you don't need EMBOSS to run PHYLIP from PISE. Using PISE you get a uniform user interface for all the supported programs. Actually this is not totally true: you get two interfaces for each: a basic interface for simple, quick use of the programs and an advanced form for more elaborate use where you can tweak options to your heart's content.

Installing PISE is not such an easy task however: although all you need are the programs and the interface generator (PISE) which is written in Perl, it is in itself a highly complex system that relies on many additional Perl libraries,

some of which in turn depend on additional libraries to various depth levels, but once you are done with all of them, you get a really comprehensive web environment.

When you access your PISE server you get a listing of supported applications, with links to both, the basic and advanced forms and to the program documentation, which open as independent web pages. This, as we have mentioned, often becomes a hindrance for the average user with a medium-sized screen, who is forced to switch among windows constantly to navigate.

Being automatically generated, PISE forms tend to be terse and less appealing than other, more elaborate or hand-crafted interfaces, but everything needed is there, and they usually come with basic descriptions of relevant fields (Figure 3).

Summary

When dealing with phylogenetic analysis, one would like to have the easiest interface possible while keeping access to advanced features. Several Web interfaces are available for PHYLIP programs, among which we have reviewed EMBOSS, WebPHYLIP and PISE.

WebPHYLIP provides a pleasant, helpful interface to almost all of PHYLIP programs, and is probably to be preferred when you are mainly interested in PHYLIP programs. It has been hand-crafted specifically for PHYLIP and that shows off. It is easier to install, easier to use, well thought out and its ability to integrate help in the main window allows for an easier introduction of the novice user

and for an easier way to explore advanced features as well.

If you (or your users) need to deal with other applications, integrating PHYLIP with them may prove a definitive advantage and you should consider a more generic interface for your work.

EMBOSS provides access to most common molecular biological applications, and as long as it covers all your needs, your best bet is to use it and integrate PHYLIP as an «EMBASSY» external application. Then any EMBOSS interface will do (Jembooss, wEMBOSS, W2H, SRS, PISE...). Just have a look and see which of them suits better your needs (a review of EMBOSS interfaces is beyond the scope of this article).

However, if EMBOSS is not enough for all your needs, and you still want to use one single interface for all your work, then PISE is certainly the way to go, despite the extra effort needed to install it. It will surely be terse and less appealing, but it allows you to integrate much more applications and process data with them.

Generally speaking, the best option is often starting with the more specific interfaces (WebPHYLIP for all of PHYLIP, an EMBOSS-specific one for PHYLIP+EMBOSS or PISE for a wider range of applications). As you go from generic to more specific, interfaces usually go from terseness to more tailored and full featured (albeit exceptions always exist). In case of doubt, install and try various options and decide which is the one that best suits you.

In any case, when speaking of phylogenies, you will always be interested in having **a good tree drawing program** that goes beyond the basic diagrams drawn by PHYLIP. Don Gilbert's Phylodendron TreePrint is one of the nicest web based tools and you should consider adding it to your toolbox (see <http://iubio.bio.indiana.edu/soft/molbio/java/apps/trees/>).

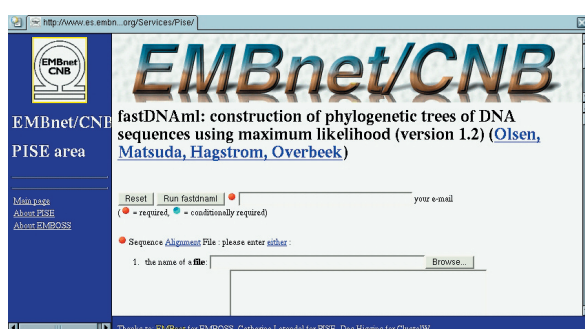


Figure 3 Example of PISE interface

A guidebook for DNA Microarray Data Analysis



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

CSC has produced a guidebook on DNA Microarray data analysis. It also has a series of PowerPoint presentations dealing with the subject, and stored in the media archive at Funet TV there are a set of videos.

This guidebook on DNA Microarray data Analysis was a collaboration between several Finnish researchers from different universities and research institutions.

The purpose of this book is to serve as course and teaching material to introduce the basic concepts of microarray data analysis. We hope that researchers who are starting microarray data analysis can benefit from the book.

The first edition of the guidebook was written by M. Minna Laine (chapters 4, 8 and 14), Tomi Pasanen (chapter 11), Janna Saarela (chapters 2 and 3), Ilana Saarikko (chapter 8), Teemu Toivanen (chapter 14), Martti Tolvanen (chapter 12), Jarno Tuimala (chapters 4, 6, 7, 8, 9, 10, 13 and 15), Mauno Vihinen (chapters 10, 11 and 12), and Garry Wong (chapters 1 and 5).

Each chapter has a section on suggested reading, which introduces some of the relevant literature. Several chapters also include data analysis examples using GeneSpring software.

We are very interested in receiving feedback about this publication. Please email your comments about this book to Jarno Tuimala: jtuimala@csc.fi



Copyright © All rights reserved by CSC - Scientific Computing Ltd., Finland.

The PDF version of this book or parts of it can be used for academic non-profit purposes, provided that this copyright notice is included. This publication may not be sold or included as part of other publications without permission of CSC. For users outside Finland, the PDF version is available. Unfortunately, we cannot send hard copies of the book abroad.

How to read the pdf-version of the book?

First, if you don't have it already, download Adobe Acrobat Reader (5.0 or later) software from Adobe (<http://www.adobe.com/>).

The whole document (13.2 MB) can be downloaded from

<http://www.csc.fi/oppaat/siru/siruwww.pdf>

Or four parts of the book separated into individual pdf files:

Part I Introduction (3.9 MB)

<http://www.csc.fi/oppaat/siru/sirupartI.pdf>

Part II Analysis (7.7 MB)

<http://www.csc.fi/oppaat/siru/sirupartII.pdf>

Part III Data mining (0.6 MB)

<http://www.csc.fi/oppaat/siru/sirupartIII.pdf>

Part IV Tools & data management (0.7 MB)

<http://www.csc.fi/oppaat/siru/sirupartIV.pdf>

PowerPoint presentations concerning Microarray analysis

A whole series of PowerPoint presentations on the subject of microarray analysis can be found at

<http://www.csc.fi/molbio/opetus/sirukokous/sirupresentations.html>

The presentations from the symposium called "Microarray data analysis developers' days" include the following topics:

1. The functionality of BASE, and how to extend it. **Carl Troein**, Complex Systems Division, Department of Theoretical Physics, Lund University, Sweden

2. J-Express - an integrated system for microarray data analysis **Inge Jonassen**, Department of Informatics, University of Bergen, Norway

3. Expression Profiler **Jaak Vilo**, EGeen Inc., Tartu, Estonia

4. What do Biologists want from Microarrays. What do computer scientists need to know **Garry Wong**, A.I. Virtanen Institute, University of Kuopio

5. Data mining of microarray data for inference of biological insight **Mauno Vihinen**, Institute of Medical Technology, University of Tampere

6. Microarray bioinformatics at TUCS/CBT Bioinformatics Laboratory **Tapio Salakoski**, University of Turku and Turku Centre for Computer Science

7. Analysis of microarray data: current research and future challenges

Jaakko Hollmén, Laboratory of Computer and Information Science, Helsinki University of Technology

8. ArrayExpress and more **Misha Kapushesky**, European Bioinformatics Institute, Cambridge, UK

Microarray video files on the FUNET media archive

A symposium called "Microarrays and data mining" dealing with microarrays was recorded and stored to the videosever of CSC in Mpeg1 and RealVideo-format. To be able to watch the files, your need to have an appropriate video player installed. The videos are available from

<http://www.csc.fi/molbio/workshop.html>

Listed below are the names of the lecturers and the topics under discussion:

Victor Jongeneel

Using EST data to produce a catalog of genes: the challenges of clustering, assembly and probe choice for cDNA arrays

Heikki Mannila

Global analysis of expression data sets

Esko Ukkonen

Computational Problems in Structural Biology

Garry Wong

Analysis of gene expression data using self organizing maps

Yike Guo

Kensington Discovery Edition: Towards a Web Platform for e-Science

Jaak Vilo

Extracting information from microarray data

Mauno Vihinen

Microarray analysis: development of B and T cells

Benedict Arnold

Supporting scientific discovery through the integration of heterogeneous data sources and applications

Michael Zhang

Super-Paramagnetic Clustering (SPC): A New Tool for Large-Scale Gene Expression Data Analysis

PGBIOINF – a Portuguese Post Graduate Programme in Bioinformatics



Pedro Fernandes

Head, Bioinformatics Unit
of the IGC

Member of the PGBIOINF
Executive Committee

Teaching Bioinformatics

Bioinformatics' usefulness comes from the fact that it addresses biological problems and finds solutions for them in the information world when they exist.

The Life Scientist of the early 21st century is continuously referencing the objects in his research to the existing knowledge through the vast amounts of information that are made available in the public and private domains. Bioinformatics is many times inevitable and the need to train people for using it in their daily life has become evident.

The Instituto Gulbenkian de Ciência (IGC) hosts the Portuguese node of the EMBnet since 1991. In an effort to address the needs of the local community, in particular – but not only – the ones that use the facilities provided by the national node, the IGC has stepped-in as a provider for such training skills and, in the past five years, has organized training courses in which more than 600 scientists have had first level or specialist education in Bioinformatics. A continuous operation was created, named the Gulbenkian Training Programme in Bioinformatics (GTPB). Training courses are held for 20 attendees, lasting for 4 to 5 days. Numerous teachers, many from the EMBnet community, were invited to run the courses. The training courses provide the training needed to use the tools, not the knowledge to build or modify them. The entry-level ones are, naturally, too broad in scope to allow for any in-depth knowledge.

The specialized ones are focused on specific subjects such as Molecular Evolution or Population Genetics but their design is such that acquiring the skill to use Bioinformatics tools is still the main focus.

User level education is insufficient for the professional practice of Bioinformatics, in other words, for the research and the development that is needed to create and continuously improve the methodology and the tools. The PGBIOINF is an educational step that aims at the creation of young, productive professional Bioinformaticians, therefore it focuses on the methods themselves and the research around them, keeping in mind the broad knowledge that a professional Bioinformatician must have in order to be able to borrow experience across the various subjects that he encounters in the profession itself. Bridging between these areas is therefore an essential ingredient in the design of PGBIOINF.

Professional Bioinformaticians need to acquire considerable knowledge both in Biology and in Computer Science in order to understand biological problems from the information side, conceive new solutions, build programs to query large databases and screen the results with reliability (quality) criteria.

Origins of the Programme

The PGBIOINF was created in 2002, as a partnership, by the Instituto Gulbenkian de Ciência (IGC) and the Faculdade de Ciências da Universidade de Lisboa (FCUL). In the IGC, providing user level training extensively (120 course attendees per year for the last 5 years) had revealed that a deepening of the level of such courses was possible, extending them into the research side. Many of the teachers involved in providing the training courses in Oeiras (GTPB) are active researchers in Bioinformatics. Only a few of the course attendees would be able and willing to follow Bioinformatics professionally as a research activity, which was normal, but receiving applications from very capable undergraduates was becoming

more frequent. Such undergraduates were finishing their first degree in Biology, Biochemistry, Medicine, Engineering, Mathematics, etc. and had discovered that Biology and its problems could be addressed with Computer Science methods. Naturally the source of such students was mainly the Portuguese Universities but, as the GTPB receives foreign students quite frequently, we were aware that some demand from foreign countries could be expected.

Several Portuguese Universities were contacted to establish a partnership that could conciliate the necessary competences and the ability to provide degrees. The idea was to start with Post-Graduate education and reach the possibility of offering a PhD within a year or two. The FCUL was building similar plans and looking for complementary teaching resources, namely the means to provide advanced thematic seminars, so a program was set-up in only a few months and was approved at the Universidade de Lisboa. The proposal within FCUL was initially headed by the departments of "Informatics" and "Chemistry and Biochemistry" but, within a month or two the "Statistics" and the two departments of Biology were in the team.

Scope and objectives of PGBIOINF

PGBIOINF aims clearly at creating new competences in the field, starting mainly from freshly graduated students. Bioinformatics is to be taken by them as a profession, not an additional skill. To do this, PGBIOINF is providing them with a coherently designed framework of knowledge acquisition opportunities, in close contact with the community of users and a selected set of active Bioinformaticians that teach the advanced topics with a strong emphasis on the current research in the area.

"Biology and medicine are moving from bench-based to computer-based sciences as models replace some experiments and complement others" as stated by charter of the Biomedical Information Science and

Technology Initiative (BISTI) created at the NIH. Biologists can hardly do without, but their focus is and must be on the Biology itself not on the ways that need to be professionally created to handle these problems and to tackle them from the information side. And as the information and knowledge repositories grow, the need for smarter ways to use them grow even more, and the need for people that can be professionally dedicated to this activity becomes increasingly evident.

Creating such professionals is a positive move towards this set of objectives. It is also an affirmative acknowledgement of the need to provide well-equipped intellects into a field where traditionally only self-made skills have been used for years. It is our firm conviction that a Bioinformatician that works in the prediction of protein function has nothing to loose if he also masters Phylogenetic analysis, for example, and that all Bioinformaticians can only work much better if their knowledge in Biostatistics is fresh and articulated.

Future perspectives for PGBIOINF students

PGBIOINF students are able to pursue their studies to a Master's (in Research) by writing a Master's Thesis and to step directly into a PhD project. This can be done at FCUL, in which these degrees have been created in early 2003, or elsewhere. This is also easy because PGBIOINF students are graded according to the methods of the European Credit Transfer System (ECTS) that allows for the easy interchange of students between schools that have implemented the EU directives, following the decisions of the Bologna convention and the "Lisbon strategy" (see reference URLs at the end of this paper).

PGBIOINF students are ready to step into a Master or PhD project once they finish, but they also have the aptitudes that allow them to take a professional job in several existing and foreseeable areas of employment such as:

- Biotechnological industries
- Pharmaceutical companies
- Food industry
- Research Laboratories
- Nature conservation
- Non-Governmental organizations

Moreover, many of these students will be able to respond to new opportunities in business that naturally arise from making their aptitudes available in a market that extends easily beyond national borders. They can also decide to create and manage Bioinformatics support units and actively deliver services to the research community in a completely autonomous way.

Programme structure

An executive committee with members of FCUL and IGC manages the PGBIOINF. The committee defines major orientations for the courses, chooses the teaching staff, selects the student candidates and monitors the progress of the Programme.

The PGBIOINF is organized in two main parts:

Part A follows a more classic university model. It has a workload 30 ECTS credits, consisting

of classes and laboratory sessions at the FCUL, complemented with weekly trips to IGC to ensure a direct contact with the local research labs, the local scientists (many of them Bioinformatics users) and the Library is taken. Students obtain competencies in the various subjects that provide the basic science supporting Bioinformatics: on one hand, Chemistry, Biology and Biochemistry; on the other hand, Data Analysis, and Systems Design and Programming.

The choice of courses allows each student to acquire the skills that are needed to complement his/her graduate profile. For example, a Software Engineer picks-up Biochemistry and Genetics, while a Biologist picks-up Programming and Database Design.

The executive committee, according to the student's profile, initially takes this choice, and each student receives individual follow-up to ensure that not only the choice but also the learning rates are appropriate throughout the Programme.

Part B is structured after a model of thematic seminars and intensive courses conducted by well-recognized experts, also totalling 30 ECTS units, to be held in the IGC. At this time, albeit

Table 1. Part A Courses

<i>Title</i>	<i>Type</i>	<i>Credits (ECTS)</i>
Introduction to Bioinformatics	3	6
Biostatistics	3	6
Biologically Inspired Algorithms	3	4
Biodiversity	3	4
Introduction to Programming	2	4
Algorithms and Data Structures	2	4
Introduction to Databases	2	6
Molecular Genetics	1	4
Molecular Structures, Metabolism and Diversity	1	4
Genetic Engineering	1	4
Chemistry and Biochemistry Laboratory	1	2

Type 1: for Engineering, Maths and Computer Science students (Info Profile)

Type 2: for Biology, Biochemistry and Medicine Students (Bio Profile)

Type 3: for all students

Table 2. Part B Courses

<i>Title</i>	<i>Type</i>	<i>Credits (ECTS)</i>
Protein Structure and Function	3	2
Genetic Analysis	3	2
Phylogenetics and Molecular Evolution	3	4
Population Genetics	3	2
Functional and Comparative Genomics	3	4
Genetic Expression and Microarrays	3	2
Data Warehousing and Data Mining	3	2
Limits and Expectations in Bioinformatics	3	2
Gene Prediction and Identification	3	4
Proteomics, Transcriptomics and Metabolomics	3	2
Gene Ontology	3	2
Population Dynamics and Epidemiology	3	2

Type 3: for all students

their different origin and interests, students are sufficiently synchronized in aptitudes to be able to attend these seminars, in which major specialized Bioinformatics topics are covered in considerable depth. The main methodologies are explained in detail, not only in their theoretical roots but also in terms of the details of the several implementations that have been made available so far. Bringing recent publications to classroom discussions, in the Journal Club style, reveals the current status of research in each of these topics and allows each student to better judge his/her own interests. It also gives them exposure in the sense of preparing them to address audiences, a practice that is not sufficiently stimulated in undergraduate courses and is an essential ingredient the life of any scientist.

The first edition

The first edition of the PGBIOINF started late, relative to the normal academic year. The announcement of the Programme could only be out in the beginning of the summer,

thus conditioning the rest of the calendar. There were 24 candidates, 6 from abroad, of which 9 were admitted. The first lecture was held on Jan 4th 2003. Part A ended 3 weeks beyond schedule in consequence of adjustments to avoid superimpositions and to the need to ensure a deeper training in fundamental subjects such as Biostatistics (mandatory for all students). Part B started in Oeiras in the planned format and students gradually adapted to the seminar form of the courses. Most of the students have chosen a subject to follow PGBIOINF with a Master's



thesis. The only exceptions are the ones that were already in PhD programmes and had interrupted them to follow the PGBIOINF.

The appraisal from the teaching staff was rather good both in the course's organizational features and in the overall quality of the end product: the student's skills.

IBM chose the student that obtained the highest grades for the "Student Achievement Prize" of 2003, that was awarded to two post-graduate students per country.

The second edition

The 2003-2004 edition was announced in January, thus allowing for a longer period for the applications. 36 applicants responded, many of them with excellent first-degree classifications. This time a considerable amount of candidates had a much better idea of where this Programme can take them in terms of their professional careers. Part A began on Sept 8th 2003, synchronized with most academic activities, which makes the planning and logistics far easier to manage. Part B is about to begin (March 22nd 2004). We now have 8 students, 6 of them are Portuguese.

Considerable amounts of experience from the first edition were used in order to make fine adjustments to this edition. For instance, more time was given to Biostatistics as we found that progress is slower than we thought, and teachers of the Computer Science courses were encouraged to use working examples from Biology in place of more abstract ones.

Present Status

At this moment, the call for applications for the third edition is open. If possible, due to the demand, we will admit more than 10 applicants. We are eager to admit foreign students, so please consider recommending it to members of your local community.

At this point I consider that the key characteristics that make the PGBIOINF an attractive option for a fresh graduate interested in practising Bioinformatics professionally are:

- Mixture of student origins and skills
- Taught and fully documented in English
- Graded in ECTS (student mobility)
- Advanced seminars in 12 subjects with research emphasis
- Individual attention to the student (small number of students, time to talk and discuss)
- Problem-driven classes, deep into the biological subjects

**The FCUL/IGC
Post-Graduate
Programme in
Bioinformatics
2004/2005 Edition
Lisboa and Oeiras, Portugal**

**INTERNATIONAL CALL
FOR APPLICATIONS**
Open from January 15th 2004
Until February 15th 2004 (early applications)
Until June 14th 2004 (final deadline)

The programme is jointly organized by the Faculdade de Ciências da Universidade de Lisboa (FCUL) and the Instituto Gulbenkian de Ciência (IGC), with participation of several research laboratories.

The 2003/2004 course will train up to 10 specialists in Bioinformatics. The classes will take place in the metropolitan area of Lisboa.

Candidates must have a degree in the life sciences, mathematics, computer science or related fields.

English is the working language of the programme and international students are welcome.

The website <http://bioinformatics.fc.ul.pt> publishes detailed information on the course and the way to apply.
For additional queries send an e-mail to: pgbioinf@igc.gulbenkian.pt

FUNDAÇÃO CALOUSTE GULBENKIAN
FACULDADE DE CIÊNCIAS UNIVERSIDADE DE LISBOA

For further reference

Course website

<http://bioinformatics.fc.ul.pt>

European Credit Transfer System (ECTS)

http://europa.eu.int/comm/education/programmes/socrates/ects_en.html

European directives towards mobility and educational strategies (Bologna, "Lisbon strategy")

http://europa.eu.int/eur-lex/en/com/rpt/2004/com2004_0021en01.pdf

Project Management (part I): a gentle introduction



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

Basic concepts

In this article we introduce some basic concepts about Project Management. Here you will find the basic concepts about what a project is, the respective roles of project members and a general overview of the Project Life Cycle (PLC).

Introduction

Project Management is mainly about *common sense*¹. However, it covers such a wide variety of tasks and abilities that it is easy to forget some if you rely on some unconscious, vague confidence on your experience. Having a structured view of the overall process and its most relevant details will help you to avoid forgetting some odd important step.

It is highly advisable to read about the discipline. There are many excellent books, and you will find that many of them concentrate on exposing the acquired experience (both good and bad) in previous projects which, when well written, makes in addition for an entertaining reading. You will find references to books and pointers in the Project Management section of the EMBnet/CNB portal:

http://portal.cnb.uam.es/tiki/tiki-directory_browse.php?parent=46

Whether you are a project manager, you plan to become one in due time or you are simply just a worker in a project, learning what a Project is all about will help you better understand the process and make it work to your advantage.

Actually, it is arguable that you are not a Project Manager yourself. As we shall see,

managing a project involves delivering a product to some interested stakeholders. Whether you are Research Policy Coordinator for the EU, a Group Leader at EMBL or an intern is moot: you still have someone (your client, a funding agency, or your boss) that is interested in the timely delivery of some results that you are responsible for, and thus you must manage your own projects (even of it is with a team of only one –**you!**).

What is a project?

A project, as defined in the “PMBok²” is *a temporary endeavor undertaken to create a unique product or service*.

This has several implications:

It is **temporary**, something with a definite start and end times, not a continuing labour like, e.g. ongoing maintenance or provision of a service. There are those witty ones who say that you know when you start a project but not when it will end. This is a typical error: you must be able to realize when a project goal cannot be achieved and stop there.

It is an **endeavour**, a complex process that requires a number of resources to be put into place in the appropriate order to succeed. Usually a project is executed by a team, whose existence is tied to the project itself, but it may as well be a one-person task. In addition it is a complex enterprise, composed of one or more tasks and associated resources that must be orchestrated to complete the goal.

It is **unique**, meaning that it purports to create something unique. Its goal may be to create thousands of the same unique product, but still it attempts to create something new or in a new way. A repetition of something that has already been done does not need the entire project apparatus, it just needs to go by the book.

It has a **definite goal**, which is the production of a **product or service**. This goal determines the creation of a new project, defines when the project ends and the whole characteristics of the project.

1 The least common of all senses

2 The Project Management Body of Knowledge, compiled, maintained and published by the PMI, the Project Management Institute, one of the main reference bodies for PM in the World.

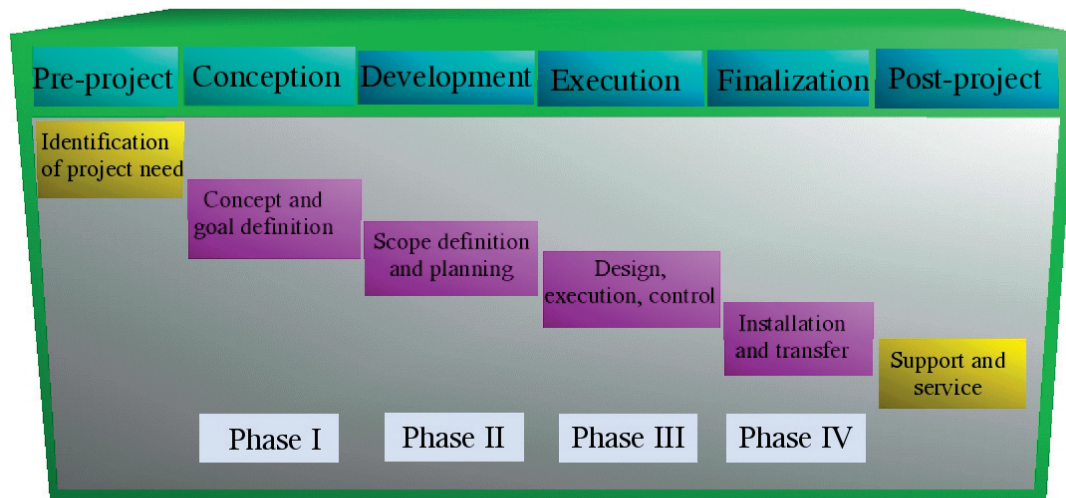


Figure 1 The Project Life Cycle

A Project Life Cycle

A project goes through several phases in what is termed a Project Life Cycle (Figure 1).

There is some general consensus on the division, although not on the names. But basically, it is obvious that before a project comes to live, something must happen: an idea must pop up about a need to be fulfilled. This idea, when accepted will result in the onset of a process whose goal will be to fulfil the need addressed.

The conception phase

Once the need has been accepted, work starts towards satisfying it. While you will probably find at most sites that corporate culture is "let's go straight for it without costly delays", common sense dictates that it is never a good idea undertaking a new enterprise without giving it a second thought and some planning. So, before rushing up, stop and take time to think.

The first step in every project should be the definition of the project goals. What do we want to achieve? The goal needs to be refined until everybody clearly understands what it is that we want to achieve. Otherwise, different people will start with incompatible conceptions and chaos will ensue, often resulting in a scandalous failure.

The development phase

Once we know what we want, it is time to make decisions. Ideally we would like to

have everything, now, perfect and at no cost. Practically we can't. The old saying "Price, Quality, Time, pick two of three" is just as true: given enough time anything can be built to the most exquisite quality standards at marginal costs. Given enough money you may always build a quality product in very little time. If quality is not a problem, you can get anything now at minimum cost. But you can't get everything except in very rare occasions (Figure 2).

Therefore, before starting, it is important to make it clear how it is exactly that we will achieve the goal proposed: we need to make compromises, establish the project scope, its reach, the resources available to accomplish the goal, deadlines, cost and deliverables. It is compulsory to make a breakdown of the project into smaller subtasks/subprojects, and these assigned



Figure 2 The trade off.

the resources needed and to the people that will be responsible for them.

This can only be accomplished successfully through in-depth communication among all the parties involved, and that in turn means involvement of everyone, the client, project managers, developers, etc.

The execution phase

Once the project has been refined, planned and approved, it is time to get the work done. The people involved need to be assembled, usually building a team and resolving all group dynamics problems, getting it to work until production level is reached, maintaining productivity until the goal is achieved and taking corrective action when something may get out of control.

The final phase

Once the product is finished, we are not done yet: it must be delivered to the client, and we need to make sure everything works as expected. The product must be deployed, installed, and possibly the client trained into using it. As the saying goes, the show isn't over until the fat lady sings, and it is always a pity to lose a lot of work because the final product hasn't been finally delivered.

Project Post-mortem

Excellent! Your project has finished and you are now able to pursue other, more exciting endeavours. But this is not the end of the story. Rather, it links to the beginning of the next. Whether your project was an astonishing success or an utter failure, there is a lot to learn from it, what you did well, what you didn't, what you may expect of yourself and the people involved... If you did your homework while the project was alive, there will be a wealth of information on how everything went, and you can milk this information to help you better plan and manage your next projects. Ignoring this information will lead you to repeat all over again the same mistakes and, don't be naïve: if your project was a success you may not be so lucky next time.

Collecting project statistics and experiences once it is finally dead is the single and only way to learn from the past for the future. You

may delay it to the next proposal, but we contend that at least some degree of self-examination should take place right after the project has ended.

The Project Manager

The role of the project manager is to be responsible for the successful completion of the project. This entangles a number of tasks during all the project life cycle.

More specifically, the project manager will have to coordinate all the labours in the process, build and cater for the team that will carry out the tasks, ensure that all resources needed are available in a timely fashion, provide a hub for communication among all parties involved (the team, upper management, the client), and monitor and control the project well-being, taking corrective action when anything deviates from the project plan.

An important aspect in Project Management is the attitude you take to it. While the approach you use may change in the different phases, there is something that is of the utmost importance for success: whether you act proactively or not.

The very inception of project management is proactive. That's why planning is the start of all the process. You should therefore strive to preview and plan in advance for all possible contingencies, trying to have a contingency plan to save the day if need comes.

Waiting and hoping for the best is easier. Sometimes you will be lucky, and if no problem occurs, you save a lot of headaches. And, if you and your team are skilled enough you may get out of difficulties fast enough to reduce impact on the project schedule, quality or costs. But whenever something happens, the only way out will be through additional, unexpected stress on the team, and this will quickly lead to burnout and a snowball of additional problems.

So, for the Goodness sake, always try to be proactive and preview risks before they happen, working out alternative solutions ahead of time.

The project team

It is a resource like any other... or is it not? Most other resources are reliable and well

known, but the team deals with people and its intrinsic unpredictability.

Most of the project life cycle is usually spent in the execution phase, and carried out by the project team. It is therefore paramount that it works as a well-greased engine. Understanding how teams coalesce and build up, how people interact within them, and providing for successful interactions is at the core of project success.

You can never consider team members as interchangeable pieces in your puzzle. You need to know each of them, their needs and expectations and their abilities. People need to be matched to the appropriate tasks, and interest and morals kept up all the way through.

While the project manager has the ultimate responsibility, team members will do the actual work, and without their backing, no project can survive. It is therefore important to get team members involved from the earliest phases, getting their experience and agreement on the plan and schedule, work subdivision and resource allocation. This actually implies making each team member a project manager for the tasks under his responsibility, and helping them succeed as managers of them.

In other words, don't just think of team members as workers, but as skilled people and fellow project managers under your wings. Care for them and shield them as much as you may from external interference: your success depends on easing their way.

The client

It is him, or her, or them, who has the need that will -hopefully- be satisfied by the successful completion of the project. It may be an actual client from outside the company, or another department, or upper management, or even yourself. Anyhow, the very existence of the project is to satisfy his/her needs.

An important part of the preparation of any project is the successful identification of all project stakeholders, that is, everybody with a direct or indirect interest in the project and influence in its development and outcome. Failure to do so results in ignoring key people and influences and usually leads to disaster

due to unforeseen influences and conflicts of interest.

Just like the team, the client must be an active part in all the process. It seldom works to follow the old method of carefully laying out the needs and then enclosing yourself in an ivory tower to make the product, which you will magically produce to the client's astonishment and awe at the delivery session. If you are a client, demand to always have a say (and a *see*) on how the project develops: it is done in your interest, and it is in your interest to supervise its evolution.

The client original expectations may change during the lifetime of the project, often as a result of the project itself, and these shifting expectations must be met to guarantee client satisfaction and hence production of a successful deliverable.

Things may go wrong, out of schedule, problems appear. Every decision that may affect the final outcome, its costs, quality, or delivery date must be negotiated with the client. Every sensible change must be acknowledged and agreed. Otherwise you will end up releasing something other than what was initially agreed, and the client will have a lawful right to consider him/herself sacked.

Want to know more?

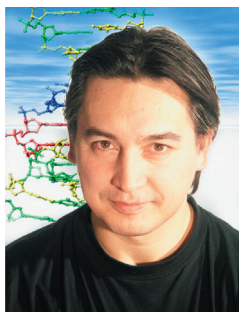
You may now take the "What is a project" quiz at the EMBnet/CNB portal and test your understanding of the topics covered.

http://portal.cnb.uam.es/tiki/tiki-take_quiz.php?quizId=3

To learn more about how projects get started, see the upcoming article on Project Conception or get yourself some good references like those included in the portal directory.

This article will be publicly available at the EMBnet/CNB portal, which we encourage you to visit for future updates, extensions, etc. If you disagree, have experiences to share or feel something is wanted or needed, please, visit the EMBnet/CNB portal and feel free to contribute and help build a significant help resource that may become a reference for the community.

MacOSX and Bioinformatics: a short report



Erik Bongcam-Rudloff

Assistant Professor
Swedish University of Agricultural Sciences
The Linnaeus Centre for Bioinformatics



Anders Nister

Project Student

New versions

We created new versions of the packages and we offer them for download at the usual place: www.ebioinformatics.org.



Two different packages are compiled for G4 or G5 processors, be sure to download the correct one.

New in the packages

All packages are now installed in `/usr/ebiotools/`, this new location makes it easier to upgrade and/or remove packages/programs. All licenses and documentations are now installed in `/usr/ebiotools/share`.

New as well:

Staden new version 1.4, a new **shlibs-package** with tcl/tk, itcl, iwidgets, f2c and tablelist, **Emboss 2.8** plus emboss-db (containing PRINTS, PROSITE, CUTG, AAINDEX, REBASE, and TRANSFAC), **Qt** updated from 3.2.0 to 3.2.3, and **EMBOSS.kaptn** updated to 0.95.

A program from TIGR for gene finding in microbial DNA, **glimmer**, has been added (see also: <http://www.tigr.org/software/glimmer/>)

Plus many bug fixes and changes to make the packages easier to upgrade or to update components.

Troubleshooting

Many users report having problems starting emboss programs, emboss-launcher.kaptn, or Staden programs. Please be sure to have the latest Apple X11 or X-Darwin version installed, most of the problems are related to old libs on previous X11 releases. Be also sure to have your `.cshrc` file with the line: `setenv DISPLAY :0.0` or if you are using bash then ensure that your `.bashrc` file has the line `DISPLAY=:0.0; export DISPLAY` as described in previous EMBnet.news issues.

Success story

We are also very proud to report having users in many different countries: the USA, Germany, Canada, New Zealand, Australia, Sweden, and Denmark among others.

Example of comments from people around the world:

From USA

«I really appreciate what you've done in putting these packages together AND writing a tutorial. I tried downloading from biotools and NCBI, and that was a waste of time.»

From Canada

«Thank you very much for your ebiotools release. It certainly makes installation much easier to have so many packages in one installation»

We appreciate comments and reports on successful Installations. We are also open to comments for improvements and if you have suggestions on programs that should be included. Please mail to: Erik.Bongcam@lcb.uu.se or Anders.Nister@lcb.uu.se.

More details and a new rewritten tutorial in next EMBnet.news issue.

WHAT MOSQUITOES SNIFF

By Vivienne Baillie Gerritsen

In these parts of the world, it is hardly the time to talk of mosquitoes when the cold winter winds are still blowing. In other parts of the globe however, mosquitoes are out and about and causing millions of deaths yearly through their ability to transmit diseases, such as encephalitis, dengue, yellow fever and of course malaria. According to the World Health Organization, malaria alone is the cause of over two million deaths in Africa, one million of which are children under the age of five. A number of different mosquito species transmit malaria; these are the anopheline mosquitoes.

The female member of *Anopheles gambiae* is the mosquito which transmits malaria to humans. The question is why does it choose humans? The answer is it may well have to do with body odor. And where odor is involved, odorant receptors are also. It has recently been discovered that female *A.gambiae* do indeed have odorant receptors on their olfactory tissues which are specific for certain chemicals found in human sweat. And this is why they make a mosquito-line, so to speak, for certain humans. A greater understanding of the molecular processes, which underlie vector-host interaction, will help develop mosquito traps or repellents in the quest to beat malaria – or indeed other types of parasitic or viral diseases.

What causes malaria? It is neither a virus nor a bacteria but a plasmodium, a single-celled animal distantly related to an amoeba. *Plasmodium falciparum* and *Plasmodium vivax* are the two main culprits for causing malaria, the former being the more wicked of the two. When a mosquito sucks in human blood from a person suffering from malaria, it also sucks in the creatures causing the disease. The latter take advantage of this by reproducing inside the mosquito, waiting for the mosquito to dig its teeth into another host and seizing the opportunity to be released into the victim's blood thus causing a new infection. So what *A.gambiae* does – quite innocently in effect – is not only act as a taxi, which ships the disease from one host to another, but also as a cosy and quiet corner for the plasmodium to

reproduce. And whilst *A.gambiae* acts as a taxi, human blood acts as a highway for the plasmodium.



Fig.1 *Anopheles gambiae* at work

One way then to eradicate malaria altogether would be to empty all humans of their blood, wouldn't it? However, that is most obviously not an option. Interestingly, research on ways to fight off the disease has come round in a circle in the past 50 years. By the second half of the 20th century, entomologists were aware that it was the female mosquito which stung. A far cry from what was believed in the times of Julius Caesar. In those days, it had been noted that people who lived close to swamps and marshes were more susceptible to be struck down by malaria. Bereft of the knowledge of biology we possess today, the people assumed that the actual stench was the perpetrator of the disease. And this is where the word 'malaria' originates from: 'mal' meaning 'bad' and 'aria', air... And while we are flirting with the science of etymology, the word 'mosquito' is derived from the Spanish 'little fly'.

So, by the 1950s, scientists all over the world were imagining all sorts of experiments to understand not only why it was the female mosquitoes that were attracted but also how they were attracted. One of the experiments involved a human steel robot which was invariably warmed to body temperature, imbibed in human sweat or even made to exhale CO₂. Research was blossoming until

DDT made its appearance in the 1960s and was so effective as an insecticide that scientists' eagerness in understanding the vector-host interaction was somewhat dulled. However today, resistance to DDT coupled with second thoughts on the uncontrolled use of insecticides has flared an interest in alternative modes of fighting off mosquitoes.

The *Anopheles gambiae* genome has been completely sequenced. On it are found around 100 odorant receptor genes. These different odorant receptors are dispersed all over the mosquito's olfactory tissues. One of them, odorant receptor Or1 is found solely on female mosquito antennae and is particularly attracted to one of the 300 chemical compounds found in human sweat: 4-methylphenol. Or1 is a transmembrane G-protein coupled receptor, typically around 400 amino acids long. Its role, like all odorant receptors, is to bind its specific odorant. As a result, a number of downstream effector enzymes induce second messengers, which in turn stimulate odorant neurons. And the whiff of human sweat is transmitted to the brain thus giving the mosquito the drive to sting. What is more, not only does Or1 seek out 4-methylphenol for the mosquito's blood-feed but it also seems to have a role in turning off the process once the mosquito has had its

fill. Indeed, once the insect is replete, the odorant receptor seems to be inhibited, perhaps following belly distension.

The existence of Or1 only in female mosquitoes – and hence its indirect role in its involvement in the transmission of malaria – as well as the discovery of its ligand should lead to interesting biotechnological applications in the quest for mosquito traps or indeed repellents. Substitute ligands could be synthesized which would trick the odorant receptors, and hence the mosquitoes, by disorienting them and leading them into a trap where they would be left to buzz aimlessly. Alternatively, ligands could be thought up which would inhibit the receptors and act as insect repellents. The discovery is promising. More research must be done to grasp in greater molecular detail the intricacies of vector-host interaction and then a cheap solution must be found to help fight off a disease which is one of the scourges of developing countries. Unfortunately, we cannot change how we smell – that is not the smell the mosquitoes are out for – but biotechnology could invent a way for mosquitoes to smell differently. And who would refuse a hot, humid summer night without unrelenting, whining mosquitoes?

Cross-references to Swiss-Prot

Q8WTE7: *Anopheles gambiae* (African malaria mosquito) odorant receptor Or1

References

1. Hallem E.A., Fox A.N., Zwiebel L.J., Carlson J.R.
Mosquito receptor for human-sweat odorant
Nature 427:212-213(2004)
PMID: 14724626
2. Kanzok S.M., Zheng L.
The mosquito genome – a turning point?
Trends Parasitol. 19:329-331(2003)
PMID: 12901929
3. Enserink M.
What mosquitoes want: secrets of host attraction
Science 298:90-92(2002)
PMID: 12364778

National Nodes

Argentina

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

Australia

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

Austria

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

Belgium

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

Brasil

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

Canada

Canadian Bioinformatics Resource, National Research
Council Canada, Institute for Marine Biosciences,
Email: manager@cbr.nrc.ca
Tel: +1-902-426 7310 Fax: +1-902-426 9413
<http://www.ca.embnet.org>

Chile

Dr. Ricardo Baeza-Yates
Dept. of Computer Science, Santiago,
Email: rbaeza@dcc.uchile.cl
Tel: N/A
<http://www.embnet.cl>

China

Jingchu Luo
Room 303, Exchange Centre, Peking University
Email: luojc@cbi.pku.edu.cn
Tel: +86-10-6275 9001
<http://www.cbi.pku.edu.cn>

Colombia

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

Cuba

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

Denmark

Hans Ullitz-Moeller
BioBase, University of Aarhus
Email: hum@biobase.dk
Tel: +45-86-13 9788
<http://www.dk.embnet.org>

Finland

Eija Korpelainen
CSC, Espoo
Email: eija.korpelainen@csc.fi
Tel: +358 9 457 2030
<http://www.fi.embnet.org>

France

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

Germany

Sandor Suhai
EMBnet node at the German Cancer Research Center
Department of Molecular Biophysics (H0200)
Email: genome@dkfz.de
Tel: +49-6221-422 342 Fax: +49-6221-422 333
<http://www.de.embnet.org>

Greece

Babis Savakis
Institute of Molecular Biology and Biotechnology
Heraklion, Crete
Email: savakis@nefeli.imbb.forth.gr
Tel: +30-81-391 114 Fax: +30-81-391 104
<http://www.imbb.forth.gr>

Hungary

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

India

H.A.Nagarajaram
Laboratory of Computational Biology & Bioinformatics
facility, Centre for DNA Fingerprinting and Diagnostics
(CDFD), Hyderabad
Email: han@www.cdfd.org.in
Tel: +91 40 7155607 / 7151344 ext:1206
Fax : +9140 7155479
<http://www.in.embnet.org>

Israel

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

Italy

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

Mexico

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

The Netherlands

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

Norway

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

Poland

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

Portugal

Pedro Fernandes
Instituto Gulbenkian de Ciencia
Unidade de Bioinformatica
2781-901 OEIRAS
Email: pfern@igc.gulbenkian.pt
Tel: +351 214407912 Fax: +351 2144079070
<http://www.pt.embnet.org>

Russia

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

Slovakia

Lubos Klucar
Institute of Molecular Biology SAS Bratislava
Email: klucar@embnet.sk
Tel: +421 7 5941 2284
<http://www.sk.embnet.org>

South Africa

Ruediger Braeuning
SANBI, University of the Western Cape, Bellville
Email: ruediger@sanbi.ac.za
Tel: +27 (0)21 9593645
<http://www.za.embnet.org>

Spain

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

Sweden

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

Switzerland

Laurent Falquet
Swiss Institute of Bioinformatics, CH-1066 Epalinges
Email: Laurent.Falquet@isb-sib.ch
Tel: +41 (21) 692 5954 Fax: +41 (21) 692 5945
<http://www.ch.embnet.org>

United Kingdom

Alan Bleasby
UK MRC HGMP Resource Centre, Hinxton, Cambridge
Email: ableasby@embnet.org
Tel: +44 (0) 1223 494535
<http://www.uk.embnet.org>

Specialist Nodes

EBI

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

ETI

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

EU

Frederick Marcus
DG Research - European Commission
Brussels BELGIUM
Email: Frederick.Marcus@cec.eu.int

ICGEB

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

LION Bioscience

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

MIPS

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

UMBER

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

TECH-MGR

Email: tech-mgr@embnet.org
The team gives support to EMBnet nodes and helps
them with maintenance and troubleshooting.
The team is formed of experienced system administrators
and programmers who ensure the availability of local
services for all EMBnet users.

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 (ISSN 1023-4144). You
can get them by anonymous ftp from:
the EMBnet organisation Web site
<http://www.embnet.org/download/embnetnews>
the Belgian EMBnet node
<ftp://ftp.be.embnet.org/pub/embnet.news>
the UK EMBnet node
<ftp://ftp.uk.embnet.org/pub/embnet.news>
the EBI EMBnet node
<ftp://ftp.ebi.ac.uk/pub/embnet.news>

Submission deadline for next
issues:

May 31, 2004
August 15, 2004
October 31, 2004
February 29, 2005