EMBnet.news

Volume 13 Nr. 2
June 2007

- p53FamTaG database
- ANN-Spec
- UTOPIA
- Linux for bioinformatics

and more ...

Bioinformatics 2007
Joint meeting of EMBnet and RIB
Torremolinos, Spain

Editorial

The present issue sees daylight in synchrony with Bioinformatics 2007, the joint meeting of EMBnet with RIB (Red Iberoamericana de Bioinformatica), in Torremolinos, Spain. Aside from reporting some of our activities, such as courses and workshops, this issue also conveys information on new software and databases, and also presents a survey of Linux based collections of self contained Bioinformatics installations. EMBnet will be entering its 20th anniversary in 2008. The editorial board of EMBnet News wishes to express its gratitude to the contributors - a long list -who have generously offered articles for publication in our newsletter and to the large, and growing community of readers. This community strongly overlaps with the large community of Bioinformatics professionals and users. This community, as such, and the fact that it remains in contact and seeks continuity in time, is broadly the greatest achievement of EMBnet in its now two decades of existence.

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

proteinspotlight

> ONE MONTH, ONE PROTEIN <

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

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

We provide the EMBnet community with a printed version of issue 82. Please let us know if you like this inclusion.

Cover picture: *Araneae* sp., Mombassa, Kenya 2007
[© Erik Bongcam-Rudloff]

Contents

Editorial Board:

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

Domenica D'Elia, Institute for Biomedical Technologies - CNR, Bari, IT
Email: domenica.delia@ba.itb.cnr.it
Tel: +39-80-5929674
Fax: +39-80-5929690

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

Lubos Klucar, Institute of Molecular Biology, SAS Bratislava, SK
Email: klucar@embnet.sk
Tel: +421-2-59307413
Fax: +421-2-59307416

Kimmo Mattila, CSC, Espoo, FI
Email: kimmo.mattila@csc.fi
Tel: +358-9-4572708
Fax: +358-9-4572302

Bioinformatics 2007

Torremolinos, Málaga (Spain)
11-14 June 2007

Workshop on Collaborative Bioinformatics

EMBnet Annual General Meeting

RIB Annual General Meeting

An international workshop on collaborative Bioinformatics will be held in Torremolinos, Malaga (Spain) during 11 th -13 th of June 2007 in coordination with the Annual General Meetings of the European Molecular Biology Network (EMBnet) and the Iberoamerican Bioinformatics Network (RIBIO).

The workshop will present major scientific endeavours in four main Bioinformatics areas, with specific emphasis on their collaborative aspects and on building up prospective ties and collaborations between the participants.

WORKSHOP PROGRAM

Starting on June, 11 th in the afternoon, specific sessions will take place addressing current advances and opportunities for collaboration in major areas, with a final, informal closing session being devoted to general analysis and discussion of the topics presented taking place in the afternoon of June 13th.

Monday, 11th June 2007

14:30 - 15:00
INAUGURATION
Adelaida de la Calle (Chancellor University of Malaga, Spain)
Oscar Grau (Red Iberoamericana de Bioinformática, RIBIO-CYTED)
Erik Bongcam-Rudloff (European Molecular Biology Network, EMBnet)

COMPUTATIONAL GENOMICS AND EVOLUTION (1):

Chairperson: **Julio Collado-Vides**, Universidad Nacional Autónoma de México, Cuernavaca, México

15:00 - 15:40
David Holmes, Life Science Foundation, Santiago de Chile, Chile
Collaborative Computational Genomics in the eScience Era

15:40 - 16:0
Gabriel Valiente, Universitat Politècnica de Catalunya, Barcelona, España
Tree-Child Phylogenetic Networks

16:00 - 16:20
Francisco Melo, Pontificia Universidad Católica de Chile, Santiago de Chile, Chile
Improving tag mapping in serial analysis of gene expression experiments

16:20 - 16:40
Richard Kamuzinzi, Université Libre de Bruxelles, Belgium
IXodus, a knowledge discovery process based on the SIMDAT-Pharma GRID technologies

16:40 - 17:00 Coffee break

COMPUTATIONAL GENOMICS AND EVOLUTION (2):

Chairperson: **Winston Hide**, South African National Bioinformatics Institute, South Africa

17:00 - 17:20
Julio Collado-Vides, Universidad Nacional Autónoma de México, Cuernavaca, México
Escherichia coli: From annotation to modeling the largest electronically-encoded regulatory network of a cell

17:20 - 17:40
Andrés Pinzón, Universidad Los Andes, Colombia
Survey and analysis of microsatellites from transcript sequences in *Phytophthora species*.

17:40 - 18:00
Doménica D´Elía, Institute for Biomedical Technologies, Bari, Italy
New insights about the role of translational control on mitochondrial biogenesis from a genomewide in silico UTRs analysis

18:00 - 20:00
POSTER SESSION: "Computational genomics and Evolution"

Tuesday, 12th June 2007

STRUCTURAL BIOINFORMATICS (1)

Chairperson: **Jose Mª Carazo**, Centro Nacional de Biotecnología, Spain

08:30 - 09:10
Alfonso Valencia, Centro Nacional de Biotecnología, Madrid, Spain
Protein structure and function prediction at the light of the CASP and Biocreative community
challenges

09:10 - 09:30
Jorge Hernández-Fernández, Labinfo, LNCC, Petropolis, Brazil
Structural model for integrin-ligand interaction. Dynamics of alphavbeta3 and alpha6beta1
inhibition

09:30 - 09:50
José Martínez-Oyanedel, Universidad de Concepción, Concepción, Chile
VisualDEP: a tool to visualize the electrostatic differences between two states of a protein

09:50 - 10:10
Danilo González-Nilo, Universidad de Talca, Talca, Chile
Biomolecular simulations and structural bioinformatics of transmembrane proteins: K+ channels

10:10 - 10:30 Coffee break

STRUCTURAL BIOINFORMATICS (2)

Chairperson: **Laurent Falquet**, Swiss Institute of Bioinformatics, Lausanne, Switzerland

10:30 - 11:10
Goran Neshich, Embrapa Informatica Agropecuária, Campinas, Brasil
3D Secondary Structure Dossier

11:10 - 11:30
J. Cristian Salgado, Universidad de Chile, Santiago de Chile, Chile
Prediction of the behaviour of proteins in hydrophobic interaction chromatography

11:30 - 11:50
Andrés N. McCarthy, Inst. de Física de Líquidos y Sistemas Biológicos, La Plata, Argentina
Structural and dynamical study of the biologically active pentapeptide contained within the islet
neogenesis associated protein (INGAP) sequence

11:50 - 12:10
Alfonso Benítez-Páez, Centro de Investigación y Desarrollo en Biotecnología, Bogotá, Colombia
Sequence and structural analysis of receptor activity-modifying proteins

12:10 - 12:30
Rosana Chehín, Universidad Nacional de Tucumán, Argentina
Insights into the mechanism of membrane fusion induced by citoplasmatic
dehydrogenases: role of the positively charged crevices

13:00 - 14:30 Lunch

SYSTEMS BIOLOGY AND DATABASES (1)

Chairperson: **Ana T. Vasconcelos**, Laboratorio Nacional de Computaçao Cientifica, Petrópolis, Brazil

14:30 - 15:10
Amos Bairoch, Swiss Institute of Bioinformatics, Geneva, Switzerland
The UniProt Knowledgebase: to annotate is useful, to annotate well is better, to reannotate is
essential!

15:10 - 15:30
Lubos Klucar, Slovak Academy of Science, Bratislava, Slovakia
phiSITE- a database of gene regulatory networks in bacteriophages

15:30 - 15:50
Ricardo Bringas, Centro de Ingenieria y Biotecnologia, La Havana, Cuba
Data integration and building of biological networks

15:50 - 16:10
Herman Silva, Universidad Andrés Bello, Santiago de Chile, Chile
The Chilean functional genomics approach: towards identifying candidate genes associated with
peach/nectarine fruit quality

16:10 - 16:30
Jingchu Luo, Peking University, Beijing, China
Database of Plant Transcription Factors

16:30 - 16:50 Coffee break

SYSTEMS BIOLOGY AND DATABASES (2)

Chairperson: **Terri Atwood**, University of Manchester, Manchester, UK

16:50 - 17:30
Susana Vinga, INESC-ID, Lisboa, Portugal
Dynamic modeling and control of metabolic networks

17:30 - 17:50
Tomás Pérez-Acle, Pontificia Universidad de Chile, Santiago de Chile, Chile
Scale free architecture of ionic networks in proteins

17:50 - 18:10
Elizabeth Tapia, Universidad Nacional de Rosario, Rosario, Argentina
Sailing towards stable microarray data multi-classification with few binary learners

18:10 - 18:40
Steve Pettifer, University of Manchester, Manchester, UK
Progress with UTOPIA

18:40 - 20:00
POSTER SESSION: "Structural Bioinformatics" & "Systems biology and databases"

21:00 Casual Dinner

Wednesday, 13th June 2007

EDUCATION AND TRAINING

Chairperson: **Pedro Fernandes**, Instituto Gulbenkian de Ciência, Oeiras, Portugal

08:30 - 09:10
Georgina Moulton, University of Manchester, UK
EMBER: an evolving bioinformatics e-learning resource

09:10 - 09:30
Vassilios Ioannidis, Swiss Institute of Bioinformatics, Lausanne, Switzerland
The open e-learning initiative.

09:30 - 09:50
Marta Bunster, Universidad de Concepción, Concepción, Chile
Teaching Bioinformatics to Bioengineers at UEDC

09:50: - 10:10
Alvaro Martínez Barrio, The Linnaeus Center for Bioinformatics, Uppsala, Sweden
BioMacKit: a bioinformatics portable teaching kit

10:10 - 10:30
Guillermo López, Instituto de Salud Carlos III, Majadahonda, Spain
Training health professionals in medical bioinformatics: experiences and lessons learnt

10:30 -10:50
Juan Falgueras Cano, Universidad de Málaga, Málaga Spain
Toward unified schemes for e-learning in bioinformatics: experiencies from the International
University of Andalucia

10:50 - 11:10 Coffee break

COOPERATIVE PROJECTS

Chairperson: **Rosana Chehín**, Universidad Nacional de Tucumán, Tucumán, Argentina

11:10 - 11:30
Erik Bongcam-Rudloff, The Linnaeus Centre for Bionformatics, Uppsala, Sweden
EMBRACE

11:30 - 11:50
Valerie Ledent, Université Libre de Bruxelles, Belgium
SIMDAT

11:50 - 12:10
Ignacio Blanquer, Universidad Politécnica de Valencia, Valencia, Spain
EELA

12:10 - 12:30
Sonia Cattley, University of Sydney, Sydney, Australia
AP Bionet

12:30 - 12:50
Alfredo Hernández-Álvarez, Universidad Nacional Autónoma de México, Cuernavaca, México
GrEMBOSS: EMBOSS over EELA GRID

12:50 - 13:05
Llanos Mora, Universidad Internacional de Andalucía, Málaga, Spain
Opening E-learning facilities

13:05 - 14:30 Lunch

15:00 - 16:00
OPEN ROUND TABLE FOR DISCUSIÓN ON COOPERATION IN E&T
Chairperson: **JR Valverde**, Instituto Nacional de Biotecnología, Madrid, España

16:00 - 16:30 Coffee break

16:30 - 17h30
OPEN ROUND TABLE FOR DISCUSIÓN ON COOPERATIVE PROJECTS
Chairperson: Federico Morán, Instituto Nacional de Bioinformática, Madrid, España

17:30 - 18:00 Closing Ceremony

21:00 Ethnic Party

Announcement

Bioclipse 1.1.1 released

The Bioclipse team is proud to announce the release of Bioclipse beta release 1.1.1. The release constitutes a major step forward for Bioclipse as we have moved to Features. This means the Bioclipse Workbench's download is now 25 Mb, and all functionality must be added using the menu alternative "Add extensions...". It contacts the Bioclipse Update Site

`http://update.bioclipse.net`

and the user can select any features he'd like to install. It is also possible to perform online updates when a new version of Bioclipse is released. Now with MacOSX versions for Intel and PPC.

Course Reports

Valparaiso, Chile, December 2006
Nairobi, Kenya, March 2007

Erik Bongcam-Rudloff and Alvaro Martinez Barrio

Dep. of Animal Breeding and Genetics, SLU and The Linnaeus Centre for Bioinformatics, SLU-UU, Uppsala, Sweden

The teachers in Nairobi: Etienne de Villiers Erik Bongcam-Rudloff, Sofia Burvall and Alvaro Martinez Barrio. at the Amboseli National Parkh

Valparaiso Chile

The authors received funds from the Swedish Linnaeus Palme programme, an exchange programme for university tutors and students, financed by the Swedish International Development Cooperation Agency (SIDA), to organize a course in Valparaiso, Chile. The local organiser was Patricio Velez from the Centro de Neurociencia, Facultad de Ciencias de la Universidad de Valparaíso Chile.

The course was an introductory 3 days course and was given in Spanish. The lectures and tutorials included introduction to Biological Databases, Homology search, EMBOSS and ENSEMBL. The tutorials in the use of EMBOSS programs were done using the Argentinian EMBnet node product "wEMBOSS liveCD". This solution provided a local environment for each student computer and we had no problems with network speed. The problems started when we initiated the Homology search and ENSEMBL tutorials. Using European resources from the computer lab in Valparaiso was a painful experience, the response times for any given query were unacceptable. Most of the students turned to NCBI blast to get any answers at all. The students were very pleased to learn about ENSEMBL but complained that the system was to slow for everyday use. This successful collaboration will continue in 2007 and a new course is planned for December 2007. This time the costs will be covered by funds obtained from "Programa Bicentenario de Ciencia y Tecnologia, Conicyt, Chile".

Nairobi, Kenya

The authors were also financed by a FORMAS/SIDA/Sarec-supported program "Sustainable development bioinformatics project between SLU and ILRI, Nairobi, Kenya" and Alvaro Martinez was partially funded by "Sederholms för utrikes resor", Uppsala, Sweden.

The course in Nairobi was this second year a longer one taking in consideration the Course Evaluation suggestions from the previous course in August 2006 (EMbnet 13.1).

The course was 8 days long and took place from March 5 to March 15, 2007.

The topics for the course were:

- Sequence analysis and alignments
- Comparative genomics using Artemis and Artemis Comparison Tool
- Biological databases and database formats
- wEMBOSS - a web interface to the popular EMBOSS software package for biological sequence analysis
- Bioclipse – a workbench for chemo- and bioinformatics
- Taverna web services

The course participants in Valparaiso Chile. December 2006

The course participants in Nairobi, Kenya, March 2007

The BioMacKit hardware: a MacMini and a LaCie hardisk

- Ensembl - a software system that produces and maintains automatic annotation of eukaryotic genomes

- The Staden package - a series of tools for DNA sequence preparation, assembly, editing and DNA/protein sequence analysis.

To solve the internet speed problems experienced in Chile and in the previous Nairobi course (August 2006) most of the programs were installed locally on the powerful server equipment provided by the Kenyan EMBnet node at ILRI, Nairobi. The course evaluation gave the course 4.6 points out of maximum 5. The response received and the needs for bioinformatics skilled students in Western Africa encouraged us to start planning a new course for 2008.

The successful collaboration between ILRI and SLU will be strengthened with the recruitment of a PhD student who will focus on Genomics and Bioinformatics of pathogenic bacterial species isolated from camel milk.

"BioMacKit" a Bioinformatics Portable Teaching Kit

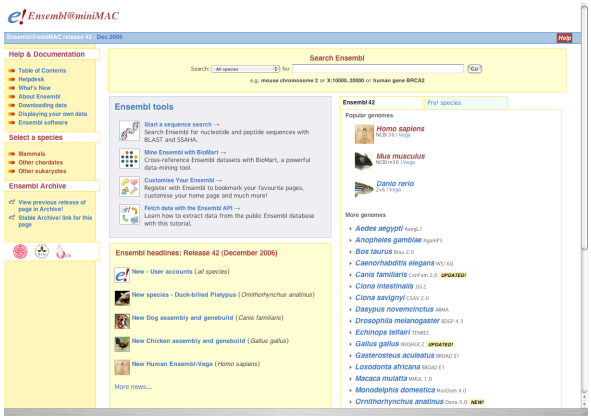
Not all universities and countries have the resources provided by ILRI and the experiences gained during the courses given in three different continents gave us the idea to create a portable bioinformatics teaching kit (BioMacKit) installed on a Mac-mini. This BioMacKit will be presented at the next Workshop on Collaborative Bioinformatics 2007 (RIBIO/EMBnet) in Malaga Spain, June 11-15. A detailed technical article about the BioMacKit will be published in the next issue of EMBnet. News. The BioMacKit contains a complete mirror of the ENSEMBL system, wEMBOSS, MRS with all the major biological databases and the Staden Package among others. The hardware is shown on the above figure and consists of a Mac-mini with an Intel Core 2 Duo 1.8 GHz processor with 2 GB RAM and connected to a LaCie 500MB fireware harddisk. This solution is a highly portable solution that can easily be carried by a teacher. The BioMacKit can also be used by a middle sized lab that needs a bioinformatics workbench or a local area server without the trouble of installing all software with dependencies and databases in different formats.

References and links

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

MRS: A fast and compact retrieval system for biological data. Hekkelman M.L., Vriend G. Nucleic Acids Research 2005 33(Web Server issue):W766-W769; doi:10.1093/nar/gki422.

ENSEMBL: T. J. P. Hubbardet al. Ensembl 2007, Nucleic Acids Res. 2007 Jan 1; doi:10.1093/nar/gkl996. Database issue.



Screen dump example of the pages served by the BioMacKit in-build web-server

Third Annual General Meeting of EMBRACE

Andreas Gisel

Institute for Biomedical Technologies, CNR, Via Amendola 122/D, 70126 Bari, Italy (IT)

Whole group during the EMBRACE AGM session.

During the 24 and 25 April 2007 the third annual general meeting of the European Network of Excellence EMBRACE took place in Lyon, France, after the kick-off meeting in 2005 in Copenhagen, Denmark and the second AGM in 2006 in Uppsala, Sweden.

EMBRACE (European Model for Bioinformatics Research and Community Education, `www.embracegrid.info`) is a wide group of experts throughout Europe who are involved in the use of information technology in the biomolecular sciences, dedicated to optimize informatics and information exploitation resulting in a highly integrated access to a broad range of biomolecular data and software packages. EMBRACE started beginning February 2005 and consists of 18 partners all over Europe and is structured into 5 work packages, data integration (WP1), tool integration (WP2), technology evaluation and watch (WP3), test cases (WP4), and outreach (WP5), lasting for 5 years.

The AGM 2007 was successfully organized under the supervision of Christophe Blanchet member of the French partner CNRS IBCP (Institute of Biology and Chemistry of Proteins).

After the publication of the technological recommendation after month 18, all partners were prepared to apply those guidelines to their data and software developments and therefore during the third AGM, EMBRACE was able to present a drastic amount of results which was clearly appreciated by the representative project officer of the European Community, Fred Marcus as well as the external reviewer Stefan Hohmann.

The key word, not surprisingly, was web services (WS), and nowadays, after 2 years EMBRACE, the partners offer a wide range of tools and data as WSs covering major bioinformatics needs such as tools for sequence analysis, structure analysis, protein domain analysis, and array data analysis or access to various databases such as transcription factor binding sites, protein sequence data, 3D structure data, and everything what is accessible under Biomart from EBI's side. During the meeting it also got clear that next year those resources will drastically increase and more important most of it, if not already, will become publicly available (see regularly the EMBRACE site `www.embracegrid.info`).

Another important part of EMBRACE are the test cases. EMBRACE stated that it will operate "test case oriented" so that the project will not loose the contact with the "real world", the biology. This AGM was the first time the WP4 (test cases) was able to demonstrate the first running test cases formulated by biologists. One test case is the automated build of HCV (Hepatitis C Virus) sequence alignment, a second one is an automated process to identify various classes of proteins belonging to a protein family using the information of the ProDom database. Further, an overview was presented by University College London of a web service which exploit CATH (a protein structure classification database) and Gene3D (a database providing structural assignments genes within complete genomes) for data-mining of functional genomics data. Several test cases are still in the pipeline since

they are more complex. Some of those test cases are using the workflow engine Taverna. EMBRACE is searching continuously for new, complex and interesting test cases and everybody is more than well come to propose a test case on our WP4 web site (https://bioinformatics.bmc.uu.se/WP4/). To support the developments and evaluation of new test cases two external biologists, Goeran Andersson (SLU-LCB, Sweden) and Francesco Di Serio (CNR-IVV, Italy) had been added to the WP4 committee. During the AGM the next stage of test cases was discussed and the first test case to run on the computational GRID was defined. It will be the protein family analysis that will be ported to the GRID infrastructure using the EGEE GRID environment.

Lastly but not least EMBRACE got from the Scientific Advisary Board that consisted of Rita Casadio (University of Bologna), Mathew Woodwark (Cambridge), and Kay Hofmann (Miltenyi Biotech) a positive feedback for the results achieved until now, but would like to see a small set of "real live" test cases solved for the biologist use, with also higher complexity of bioinformatics problems.

Christophe Blanchet from CNRS IBCP

Fred Marcus, project officer of the European Community

p53FamTaG database:
a public resource which integrates genome-wide *in silico* and experimental analyses of p53 family direct target genes

Elisabetta Sbisà

Institute for Biomedical Technologies, CNR, Via Amendola 122/D, 70126 Bari, Italy

elisabetta.sbisa@ba.itb.cnr.it

The p53FamTaG database (p53 FAMily TArget Genes) is a unique integrated public resource developed for supporting and integrating high-throughput *in-silico* and experimental analyses of p53 family target genes. Considering the central role that the p53 family members play in the inhibition of tumour progression and in development and differentiation, this database represents an important reference resource for research groups involved in the fields of oncogenesis, apoptosis and cell cycle regulation.

p53FamTaG database is available free at http://www2.ba.itb.cnr.it/p53FamTaG/ and was developed by **Sbisà E, Catalano D, Grillo G, Licciulli F, Turi A, Liuni S, Pesole G, De Grassi A, Caratozzolo MF, D'Erchia AM, Navarro B, Tullo A, Saccone C, Gisel A**, who contributed equally to the work according to their expertise in the data generation and database construction (1).

Introduction

The p53 gene family is composed of three genes, p53, p63 and p73, with polyhedrical functions in pivotal cellular processes such as DNA synthesis and repair, growth arrest, apoptosis, genome stability, angiogenesis, development and differentiation. p53, p63 and p73 encode sequence-specific nuclear transcription factors which recognise the same responsive element (RE), but

with a degree of specificity for the target genes that is quantitatively distinct. The RE is made up of two or more tandem repeated decamers complying with a specific consensus corresponding to the 5'-PuPuPuC(A/T)(T/A)GPyPyPy-3' sequence, spaced by 0 to 13 nt. The three genes are differentially regulated and carry out specialized, non-overlapping functions. Their inactivation or aberrant expression may determine tumour progression or developmental disease (2).

The identification of the genes transactivated by p53 family members is crucial to understand the specific role of each protein in different cellular processes. With the aim to identify new direct target genes, we combined a genome-wide computational search of p53 family REs and microarray analysis. The huge amount of biological results produced raised a critical need for bioinformatic instruments able to manage and integrate the data and facilitate their retrieval and analysis. Therefore we developed the p53FamTaG database, which contains p53 family direct target genes selected in the human genome searching for the presence of the REs and the expression profile of the target genes obtained by microarray experiments. The data we produced were integrated with other experimental, bibliographic and computational annotation and were made publicly available and retrievable as a web resource

Data model

The p53FamTaG is structured in a relational database schema using MySQL Database Management System. It was designed in a modular way so that data coming from computational, experimental analyses and public resources can be integrated and updated independently as and when needed. The key modules of the database are:

- **InSilicoData**: this module is the core of the database which stores the information of the target hits (RE) obtained by whole-genome *in silico* analysis. The information for each RE includes: the RE sequences, the genomic coordinates, genomic regions were the REs were found (intron, promoter, 5'UTR) and the gene names.
- **RawData**: this module stores the raw microarray data obtained by the AB 1700 Applied Biosystems platform.
- **ExperimentalData**: this module contains the gene expression values obtained by the statistical analysis on the microarray quantile normalized data and the annotation of the genes spotted on the microarray.

Data generation

The genome-wide computational analysis to identify the human p53 family REs was performed using the PatSearch algorithm implemented in the DNAfan tool developed in our Lab (3, 4). DNAfan filter facilities allow the user to analyze specific genome regions (e.g. promoter regions, 5'UTR, introns, etc.). The program therefore automatically generates, on feature key annotation, a specific sequence dataset spanning a given feature key on which the desired analysis program, in our case PatSearch, is executed. PatSearch is particularly suitable for searching sequence data for the presence of complex oligonucleotide patterns, the structure of which was derived from experimental characterization of functional elements. In order to reduce the number of hits and to enhance the selectivity of the pattern searching analysis, we optimised the original RE consensus by introducing new criteria of stringency based on the comparative analysis of 109 REs contained in 83 experimentally demonstrated human target genes of the p53 family members. Using this optimised syntax pattern, we performed a genome-wide search in the ENSEMBL database (release 34) and found 63,384 REs in 18,110 genes, after redundancy cleaning based on their absolute genome coordinates.

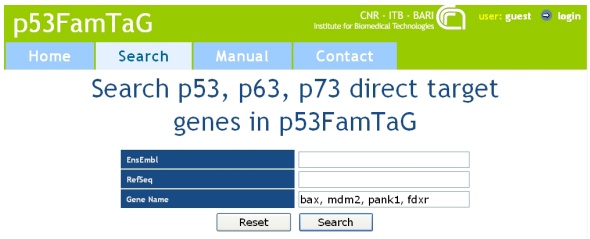


Figure 1. Database query form. The search criteria are ENSEMBL and RefSeq idenfiers, HUGO or alias gene names. All search fields accept lists of items separated by a comma and execute the search in OR mode. All three different identifiers can be used in the same search.

Total Number of Entries: 4

Gene Name	EnsEmbl	RefSeq	Localization	Chr	Strand	REs	Array	UCSC
BAX	ENSG00000087088	NM_138762 NM_004324 NM_138761 NM_138763	INTRON, PROMOTER	19	⟼	3		
MDM2, HDM2	ENSG00000135679	NM_002392	INTRON, 5UTR	12	⟼	6		
PANK1, PANK, MGC24596, PANK1a, PANK1b	ENSG00000152782	NM_138316 NM_148977 NM_148978	INTRON, 5UTR, PROMOT	10	⟼	8		
FDXR, ADXR	ENSG00000161513	NM_024417 NM_004110	INTRON, PROMOTER	17	⟼	3		

Figure 2. Database query report. This figure shows the results of the query in Fig.1. The query matching records are ordered by ENSG ID.

These data were integrated with the microarray results produced in our Lab from the overexpression of different isoforms of p53, p63 and p73 (wild type p53, p53 mutated form p53R175H, TAp73α, TAp73ß, TAp63α, ΔNp63α) in Flp-In-T-Rex-293 isogenic cell lines stably transfected, at 6 h and 24 h after their induction. A stable cell line containing CAT was used as a control.

The p53FamTaG database also provides annotation extracted from different public resources such as: ENSEMBL Gene ID, gene name and aliases, RefSeq accession number, Celera gene ProbeID and links throught PubMed to the papers reporting experimental data produced in different labs. In particular, the database annotates 132 genes experimentally demonstrated as direct target genes by other Authors and 341 p53 high-confidence binding loci obtained using the genome wide ChiP-PET approach (5). These genes have been linked also to UCSC, where the annotation can be found in the ENCODE Chromatin Immunoprecipitation tracks under the p53 ChiP-PET analysis (GIS p53 5FU HCT116 Track Settings).

User Interface

A graphical user interface (GUI) was built to query, in an integrated way, both *in silico* and experimental data contained in the p53FamTaG database. The GUI was developed using PHP Seagull Framework.

The main GUI features are explained in the following sections.

Search options

The user can query the database to find out whether a gene of interest contains a p53 family RE and how this gene is expressed under over-expression of the different members of the p53 gene family in human 293 T-Rex cells. The Figure 1 shows the database query form and the Figure 2 the query result report.

The query report displays retrieved information as a table (Fig. 2). The genes whose experimental data are reported in literature are linked to PubMed references through the clickable book button on the left table side. The "Gene Name" column displays the HUGO and aliases gene names. The HUGO gene name, ENSG ID and NM ID are linked to the HGNC, ENSEMBL and RefSeq databases respectively. The RE localization (intron, promoter, 5'UTR), the chromosome, the strand orientation and the number of the REs found are reported in the relevant columns. More detailed information on predicted REs can be queried clicking on the REs hits. For each RE the database provides the absolute chromosomal position, as it is reported into the ENSEMBL database, size, strand orientation and the gene region localization as well as a graphical representation of the target sequence depicting the pattern as shown in Fig.3.

Sequence Export

One particularly noteworthy feature of the database is the possibility to export the sequences of the REs including full information about structure and localization. Sequences can be selected for export using the check boxes on the left or the

Gene Name	BAX
ENSG	ENSG00000087088
Chromosome	19
Strand	⟼
Start	54149998
End	54156708
RefSeq	NM_138762 NM_004324 NM_138761 NM_138763
PubMed	Articles

Total Number of REs: 3

Select page | Export selected | Deselect page

Select	Start	Size	Strand	Localization	Pattern
☑	54149496	38	⟼	PROMOTER	
☑	54150257	35	⟼	INTRON	
☑	54153092	42	⟼	INTRON	

Figure 3. p53RE details. Information retrieved on REs predicted in the BAX gene. The REs pattern is graphically represented by circles (decamers) and number of spacer bases.

"Select/Deselect" buttons and sequences exported in FASTA format as text file by clicking on the "Export selected" button (Fig. 3).

Microarray data

The magnifying glass-button in the "Array" column of the query report provides the link to the microarray data table (Fig. 4).

For each probe ID detecting the gene, the results of the gene expression of Samples (S) compared to the Control (C) are indicated as "Up regulated" when S>C (fold change value reported as S/C), "Down regulated" when C>S (fold change value reported as -C/S), "Not statistically significant" when the expression value has a PPDE (<p) < 0.995 and as "Negatively filtered" when the expression value has been filtered out by quality control (flag > 5000 and signal to noise S/N < 3).

Finally, the presence of a clickable paper sheet in the UCSC column activates the link to annotations of ChIP-PET binding loci in the UCSC database.

Remarks

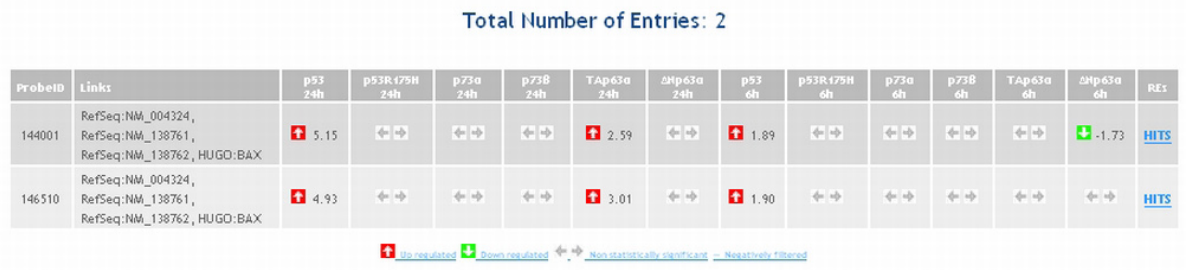
p53FamTaG represents a unique integrated resource of wide genome search of human p53, p63, p73 direct target genes combining *in silico* prediction of p53 family REs and microarray analysis linked to other public databases. p53FamTaG provides the user with a query/retrieval system and allows the export of RE sequences. The p53FamTaG annotates also experimental data produced in different labs; in particular 132 experimentally verified p53 family target genes and 341 p53REs recently identified by ChIP-PET strategy (5) linked to UCSC genome tracks and to PubMed entries.

Acknowledgments

This work was supported by grants from MIUR: Cluster C03 Prog. 2 L.488/92; PON - Avviso n. 68 del 23.01.02 Progetto B.I.G; Contributo Straordinario D.M. n. 1105 del 09/10/2002 (Progetto n. 187);PNR 2001-2003 (FIRB art.8) D.M.199, Strategic Program: Post-genome, grant 31-063933; FIRB 2003 art. 8 D.D. 2187 del 12-12-2003 LIBI. We thank Dr. D. D'Elia for critical discussion

References

1. Sbisà E, Catalano D, Grillo G, Licciulli F, Turi A, LiuniPesole G, De Grassi A, Caratozzolo MF, D'ErchiaNavarro B, Tullo A, Saccone C, Gisel A. p53FamTaG: a database resource of human p53, p63 anddirect target genes combining in silico predictionand microarray data. BMC Bioinformatics 2007,(Suppl I):520.
2. Murray-Zmijewski F, Lane DP, BourdonJC. p53/p63/p73 isoforms: an orchestra of isoformsto harmonise cell differentiation and responsestress.Cell Death Differ. 2006, 13(6):962-72.
3. Grillo G, Licciulli F., Liuni S., Sbisa' E., PesolePatSearch: A program for the detection of patternsand structural motifs in nucleotide sequences.Nucleic Acids Res. 2003, 31 :3608-3612.
4. Gisel A, Panetta M, Grillo G, Licciulli VF, LiuniSaccone C, Pesole G: DNAfan: a softwarefor automated extraction and analysis of userdefinedsequence regions. Bioinformatics 2004,20(18):3676-3679.
5. Wei CL, Wu Q, Vega VB, Chiu KP, Ng P, ZhangShahab A, Yong HC, Fu Y, Weng Z et al: A globalmap of p53 transcription-factor binding sites inhuman genome. Cell 2006, 124(1):207-219.

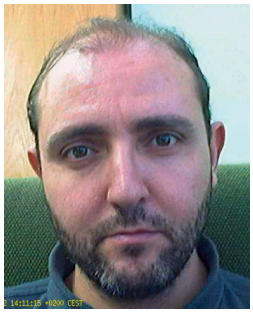

Total Number of Entries: 2

ProbeID	Links	p53 24h	p53R175H 24h	p73α 24h	p73β 24h	TAp63α 24h	ΔNp63α 24h	p53 6h	p53R175H 6h	p73α 6h	p73β 6h	TAp63α 6h	ΔNp63α 6h	REs
144001	RefSeq:NM_004324, RefSeq:NM_138761, RefSeq:NM_138762, HUGO:BAX	5.15	←→	←→	←→	2.59	←→	1.89	←→	←→	←→	←→	-1.73	HITS
146510	RefSeq:NM_004324, RefSeq:NM_138761, RefSeq:NM_138762, HUGO:BAX	4.93	←→	←→	←→	3.01	←→	1.90	←→	←→	←→	←→	←→	HITS

Up regulated Down regulated ←→ Non statistically significant – Negatively filtered

Figure 4. Microarray data. Expression data of BAX gene. The table reports the Celera Probe ID, detecting the gene on the array, the RefSeq ID and the fold change values at 6 and 24h after the induction of the p53 family members in each cell line.

Finding conserved motifs with ANN-Spec



José R. Valverde

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

Abstract:

Finding conserved sequence motifs is a task often needed in transcriptomics experiments to detect regulatory factors common to a given set of sequences. It is also useful to detect relevant domains common to a family of protein sequences. Spotting preserved motifs where there is little sequence conservation is a difficult task for humans and that is why we must resort to specific tools like Neural Networks or Hidden Markov Models. ANN-Spec is a neural network based system developed at the University of Washington. At the Spanish EMBnet node we have developed a web interface to make use of ANN-Spec as easy as possible; this interface is freely available for download.

Introduction

Detecting sequence motifs

As more and more biological data becomes available, we are faced with new challenges made possible by this same data availability. One of them, is identifying common traits to a group of sequences sharing some known biological property (like, e. g. being co-expressed and hence possibly co-regulated).

The most naïve approach calls for making a sequence alignment to identify common stretches. While this approach is intuitively appealing, it has some major drawbacks that make it impractical in most cases: building an alignment is a difficult task to do by hand, and resorting to automatic tools may easily be even worst unless the common characteristic is big enough to clearly stand from the noise introduced by the rest of the sequence. Being more specific, automatic alignment programs will look for the best alignment, and hence will overlook shorter (but possibly more significant) coincidences as well as motives superimposed at different positions on the main sequence (as is usually the case with regulatory sequences); moreover, if the common trait is badly preserved (as is the often the case) it may be overlooked totally, dismissed as noise; and even worst, automatic tools can only detect sparse motives if all subsections are always laid out in the same order (hence no rearrangements are possible).

The advantage of a Neural Network is that it does not suffer of these shortcomings: a neural network may be able to detect badly preserved motives, in any order and -even better- without any need for the sequences to be aligned at all.

It is true that we may do the same work probably using alignments: we could start by detecting the longest aligned sections, removing them, realigning to detect smaller sections, realign, etc... and so on, for both chains, in both senses, and even so we might lose overlapping motives. But it would obviously be a lot of work. A well-designed and well-trained neural network hence can save us time and effort by performing all this in a single sweep.

But wait, there's even more: when we look for motives we want to be sure they are significant, i. e. that they are not due to random coincidences. But this in turn depends on the composition of the sequences and on the distribution of residues (nucleotides, dinucleotides, triplets, etc...) which we know is not uniform, resulting in a need for complex formulas based in detailed statistical studies of your sequence sets to interpret the significance of your findings.

Again, a Neural Network saves us a lot of help, for it may spot non-random stretches and evaluate their significance at the same time it processes the sequences under study. A significant help indeed.

What are neural networks?

Basically, neural networks are a computer simulation of the behaviour of natural neuronal systems. The method was primarily developed by Frank Rosenblatt who described his system as a **perceptron**. Perceptrons were able to process and identify simple patterns, but were forgotten due to some perceived shortcomings at the time. It would take almost 20 years for his work to be vindicated by John Hopfield and become popular in the AI world. Now we know that neural networks are a special kind of a "general problem solver" and can be applied to very complex problems successfully.

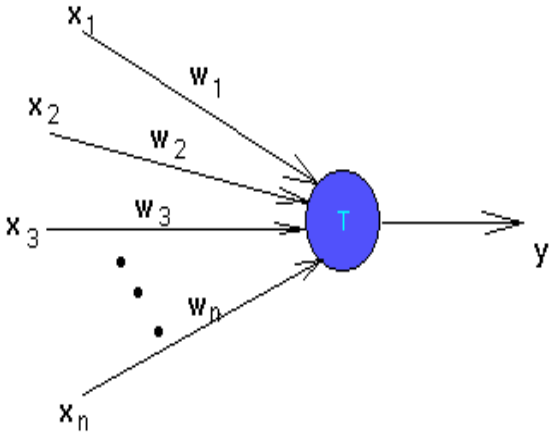


Figure 1: A perceptron takes a number of inputs (xi) weighting each of them appropriately (wi) and uses them to generate a single output value (y).

An artificial neuron abstracts the properties of a real neuron: in short, it accepts any number of inputs, weights them and takes a decision based on their cumulative effect which is sent out as output. We can simulate a neural network (like a brain) by connecting the output of some neurons to the input of others, to any level of complexity we like. In practice, it has been shown that one needs not get too far relating to the number of layered neuron shells one uses: it is enough to just use three layers (an input layer that receives our data -you can like it to perceptive stimulus), an inner layer doing the processing and an output layer providing the results. Of course, using less layers implies needing more neurons, but the simplification is -computer wise- worth the effort.

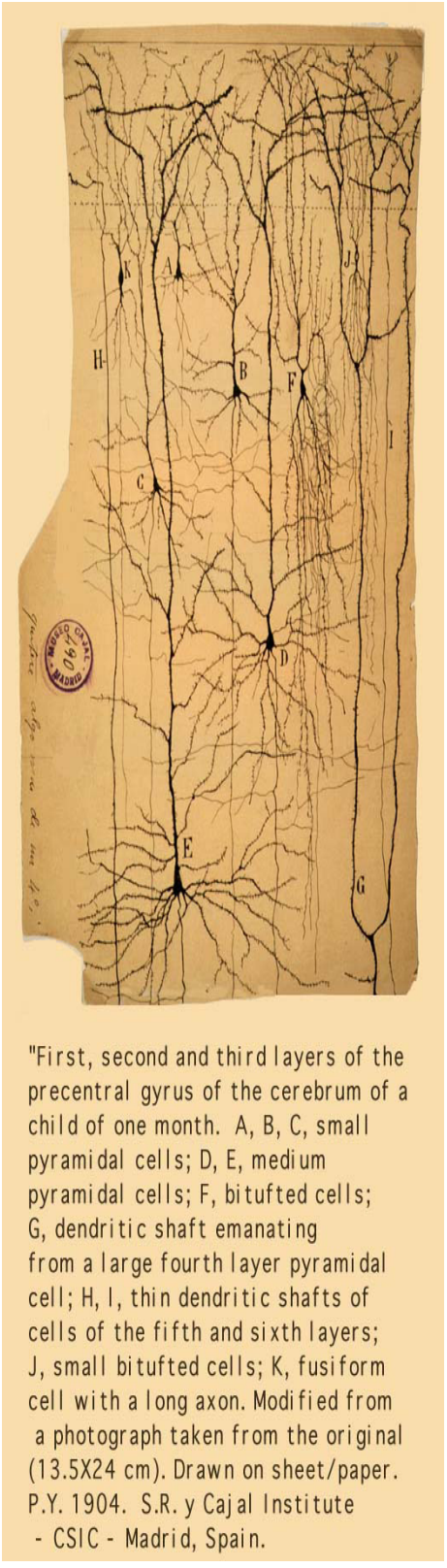


Figure 2: Santiago Ramón y Cajal's drawing of a three layer neural system on the precentral gyrus of a child's brain.

The trick to get a successful neural network relies in selecting the appropriate weight for the inputs received by each neuron. There is no easy, a-priori way to do this, but we already have a large experience in this problem: we have been dealing with it for literally millions of years as we ourselves are neural network based. It is the process of training. What we do is start with a random set of weights and then feed the neural network some stimulus (input), next we look at the results we get and make corrections on the different weights assigned to each communication path based on how good (or close to the right answer) the output obtained was.

This means that we do not just have a single, general purpose neural network: we must build networks to suit each problem, and train them to solve it, which in turn requires us to provide some well-known data (whose answer we know in advance) to start with. We train the network using this data and then we can use it to do the work it was trained for.

Detecting motives with ANN-Spec

ANN-Spec

Original neural networks used in Biology were trained to solve very specific problems: you would get a set of sequences with a well-known property, use them to train the network, and then use it to spot the presence of this property on further, unknown sequences. A good example is the location of splice sites to identify introns and exons: you would start from a set of sequences for a given organism, whose splice pattern was well known, train the network and then use it to identify splice sites in new genomic sequences.

The approach described has been successfully employed in a large number of problems, but has a serious drawback: we need to work with an initial set of sequences known to contain the property we desire and where we also know its exact location (e.g. splice sites). It is not suitable hence for our current problems where we wish to identify unknown characteristics (if we already knew them, we wouldn't need to use the neural network at all) such as identifying novel regulatory regions.

But we can look at the problem from another direction: we are not interested in identifying a *given, known* regulatory section or protein motif, we are interested in spotting regions that are *not due to randomness*, i. e. have some putative biological significance. This is what ANN-Spec does: it is a neural network trained to identify randomness and hence it will chock on preserved non-random stretches of sequence, exactly those you are interested in! And since ANN-Spec has learn to recognize randomness, it can also give you an estimate of how non-random (i. e. significant) the spotted common motif is.

In addition, ANN-Spec has additional practical advantages: it is freely available on the Net at its home web site, and you can get the source code to it so it can easily be ported or adapted to run on any system, plus having the source code you could explore, verify and check it as well as reuse it for other purposes.

Using ANN-Spec

As we have said, ANN-Spec will look for non-random, common stretches of sequence on an unaligned set of sequences. Our main problem, as we laid out at the beginning of the introduction is now how do we define randomness as this is a function of varying sequence properties (like residue distribution).

A first approach is to use a generic distribution function derived analytically. This has the advantage of being readily available, but as we have said may fail to consider biological deviations in the statistical distribution of residues or groups of residues.

A second approach is to train again the neural network to recognize whatever randomness may mean for our set of sequences, using a different set of sequences from the same organism, family, etc.. to show the network what randomness means for the specific problem we have at hand. This allows us to neglect the complex problem of deriving a specific analytic function for each problem we want to solve by letting the neural network derive its own.

So, how do we go about using ANN-Spec? At this point it is worth noting that this is a command-

line program, which is not difficult to use but, certainly, is not the easiest way to work for the average experimental biologist. For this reason we have developed a web interface at EMBnet/CNB to simplify working with ANN-Spec. This web interface is written in C and PHP and is freely available under a liberal license for anybody willing to use it at their own site. We will therefore illustrate use of ANN-Spec using the easy web interface at EMBnet/CNB.

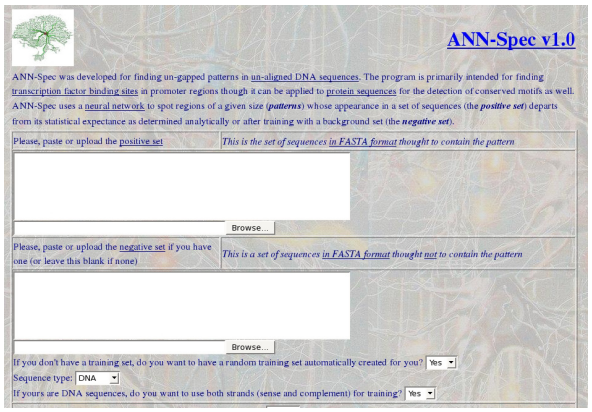


Figure 3: The ANN-Spec web user interface at `http://www.es.embnet.org/Services/MolBio/ANN-Spec/`.

Defining the problem

To start using ANN-Spec all we need is to have a set of sequences that we know or suspect contain a common motif (for example, the 5' regions preceding a set of co-expressed genes -and which we therefore believe are co-regulated by a common transcription factor- where we want to locate the TF binding site). This is called the positive set and we can either paste it or load it directly from a file.

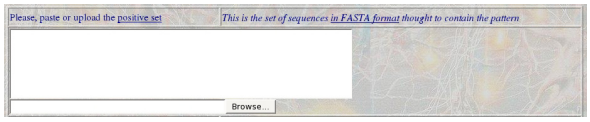


Figure 4: Defining the positive set for ANN-Spec.

Training the network

Next we need to specify a negative set, i. e. a set of sequences we can use as a contrast against which we can highlight non-random, possibly significant hits. This negative set will be used to train the network so it learns what should be actually considered as random in the problem at hand.

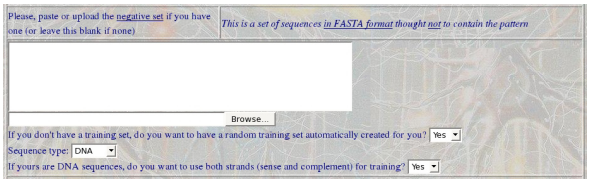


Figure 5: Defining the negative set for ANN-Spec.

If all we have is the positive set of sequences (those we suspect contain a common motif) then we have two choices:

- we may ask ANN-Spec to generate and use an analytical function derived from the properties of our positive set sequences
- we may generate a random set of sequences to use as a negative set (this is done by a separate program).

If we do not generate a set of sequences at random, then an analytical function derived from our input data may be generated by ANN-Spec instead. This is probably better unless we know for sure we want a set of *really random* sequences to use as a negative set (for instance because each sequence comes from a different organism with different residue usage biases and we don't want these to influence the analysis).

The other possibility is that we have at hand a second set of sequences from a related origin which we can use as a negative set. Again we have two possibilities to consider:

- we may know for sure that the negative set *does not* contain the pattern we want to identify (i. e., it is a true negative set of sequences), for example because it contains sequences known for sure *not* to share the same characteristic (regulation, domain..)
- we may be unsure as to whether the pattern searched may be present in the negative set: for example it may be composed of generic genomic sequences which might contain the same feature (we just don't know) or we may want to include the positive set sequences

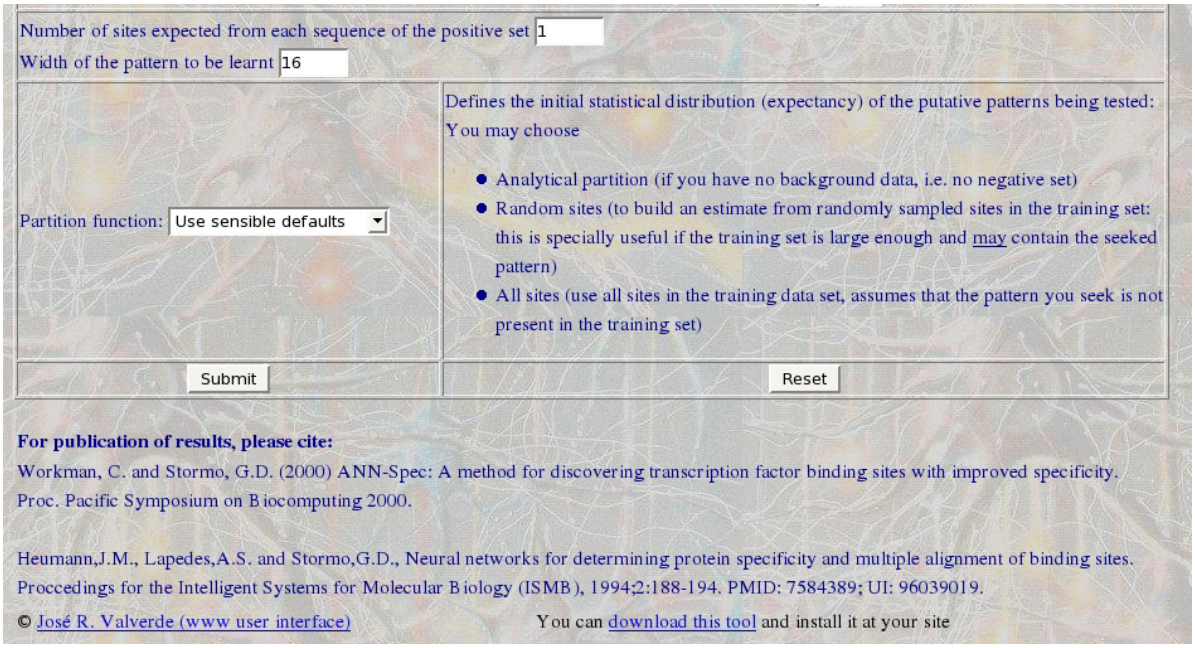


Fig. 6: Defining the parameters for ANN-Spec: number of sites expected, patterns to learn and partition function. At the bottom of the page publication details and a link to the GPL source code is included as well.

in the negative set to provide more training data.

Depending on the situation, we will need to use a different approach to estimate the significance of our findings (as in the second case we need to correct the statistics for the possible incidence of the patterns found). Estimation of statistical significance is done by means of a so called partition function. While the interface provides an option to use sensible defaults, you may force usage of a given partition (statistical estimation) function. The possibilities available are

- use of an analytical function (again, you may derive a generic function from the data provided just as when you only had a positive set and no negative data to contrast against)
- use a random sampling of the negative set: if you are unsure as to whether the pattern may be present, then you can greatly reduce its impact by only picking some residues at random (which will make the possibility of including the pattern in the analysis by chance close to nil)
- use all sites in the negative or training set, which will make for a bigger data set from which the network can learn to discern randomness, but should only be used if we are sure the pattern we seek is not present.

Additional options

There are a few extra options that can be defined to further specify the analysis we want to make. They are fairly self-explaining and obvious:

- the kind of sequence (DNA or protein)
- whether to use both strands if it is a DNA sequence
- number of sites expected from each sequence
- expected pattern length

The last two are estimations that can greatly help speed up the analysis. Of course you may specify a very large pattern length and let the network discover that there are only shorter ones, but this will result on increased work and probably on putative sites being reported noisily. So, if you have an idea about the expected length, it is better to specify one, otherwise, you may start submitting various jobs with a number of reasonable lengths that may cover the patterns and then compare the results.

The reasoning for sending various jobs is simple: if you select a length too short, you will probably find the pattern, but *not the full* pattern, and so you will need to extend the length, plus if the pattern allows for intervening tracts, you may only detect one end and miss the other; on the other hand, starting with a length that is too long you may join various separate patterns into a single one or get short patterns reported as large ones with large tracts of 'X' residues, or even miss them. Thus, you will have to experiment a bit with each problem you have.

Interpreting results

As a result of running the program you will get a very long listing. This will include first your input data sets as seen by the program (so you can verify it saw them correctly) and the options used to run it (for your reference later when comparing analysis results). Then you will find the actual program output. The output will contain all the patterns detected, as weighted alignment matrices. The resulting output is labeled with special tags to facilitate interpretation. Most probably, the results you will be more interested in, and the ones you will want to look at first are the following:

- STR: *site results*, provides the alignment with the best weights for each identified site, annotated with their sequence, position and orientation, and followed with the average score obtained

- ALR and WMR: *alignment and weight matrix results*, show the counts for each residue at each position (alignment) and the relative weights derived for each residue and position. The alignment is followed by additional complementary information, including the consensus sequence, its information content (Shannon entropy) which serves as a measure of how much it deviates from random expectation (or how meaningful it is), and the sequence that reached the best score (in arbitrary units) during the search.

We have already mentioned the main caveats when interpreting the results, which derive by common sense from the expected number of sites and their length. As with any other analysis, you should always consider the parameters you used and put them into the appropriate biological context, something you only as a Biologist can do or, in other words, make extensive use of common sense and repeat your analysis with different parameters whenever in doubt or guaranteed by you biological expectatives.

Final comments

It is worth noting that the ANN-Spec program actually offers many more possibilities that those described here when used on the default command line, among them the ability to search for a defined alignment matrix (such as those it reports) in a sequence or set of sequences. Not all the features available in ANN-Spec have been implemented in the web interface since the goal was to provide an easy tool to identify preserved patterns such as regulatory motifs and since there are already many programs that may be used to search for the incidence of weight matrices or consensus sequences on a query sequence.

Bibliography

Rosenblatt, Frank (1958), The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, Cornell Aeronautical Laboratory, Psychological Review, v65, No. 6, pp. 386-408.

Parallel Distributed Processing. David E. Rumelhart, James L. McClelland and the PDP Research Group. The MIT Press. Cambridge, Mass.,1987

Workman, C. and Stormo, G.D. (2000) ANN-Spec: A method for discovering transcription factor binding sites with improved specificity. Proc. Pacific Symposium on Biocomputing 2000.

Heumann,J.M., Lapedes,A.S. and Stormo, G.D., Neural networks for determining protein specificity and multiple alignment of binding sites. Proccedings for the Intelligent Systems for Molecular Biology (ISMB), 1994;2:188-194. PMID: 7584389; UI: 96039019.

ANN-Spec web site: `http://www.cbs.dtu.dk/~workman/ann-spec/`

ANN-Spec web interface: `http://sci.cnb.uam.es/Services/MolBio/ANN-Spec/`

UTOPIA:
User-friendly Tools for OPerating Informatics Applications

S.Pettifer, T.Attwood, P.McDermott, J. Sinnott and D. Thorne

The University of Manchester, Oxford Road, Manchester, M13 9PL

http://utopia.cs.manchester.ac.uk

Introduction

UTOPIA is the application of modern visualisation, interaction and knowledge management techniques to the problem of analysing bioinformatics data. It is a software suite consisting of a set of friendly interactive and interoperable graphical tools combined with seamless access to distributed workflows, web services and databases. Its key aim is to protect from the user all the technological complexities of accessing these facilities, hiding them behind familiar desktop metaphors such as drag-and-drop and cut-and-paste, without trivialising the problems of data integration and limiting the kind of functionality available. It is free software, available to download for OS X, Linux and Windows from http://utopia.cs.manchester.ac.uk.

Architecture

The system's architecture is broadly separated into three layers, as shown in Figure 1. A user of the system sees only the visualisation and analysis tools, which appear as independent applications but are interoperable, sharing data much as wordprocessors, spreadsheets and databases would in an office software suite. Behind the scenes lies a networking infrastructure that connects the tools, and a sophisticated semantic model that invisibly manages the transformations necessary to move data seamlessly between applications. Finally, a series of conduits connect the system to external resources that provide data and algorithmic functionality via workflows and web services.

UTOPIA is a high level implementation of the 'Model View Controller' design pattern, which hinges on a clean separation between architectural components representing the underlying data (the 'model'), mechanisms for displaying the data in various ways (the 'views') and an eventbased infrastructure for linking these together (the 'controller'). This approach has the advantage that new views can be introduced to the system without affecting the underlying model, and similarly the data model can be populated from multiple sources without disrupting the behaviour of the viewing tools. A key feature of the system's design is *semantic* rather than *programmatic* integration, which is to say that the components are loosely integrated in terms of the code used to make them communicate, and instead achieve interoperability by sharing a semantic model.

The Visualisation Tools

The UTOPIA suite currently has three front-end applications. These are shown in Figure 2.

CINEMA [9] is a fully-featured multiple sequence alignment tool. Alignments can be shown at different scales within a particular view, from a close-up suitable for detailed editing tasks through to a pixel-per-residue overview. Multiple views of an alignment can be open simultaneously allowing, for example, many different regions of an alignment to be shown at once, or for the same region to be compared at different levels of detail, using separate colour schemes if desired. CINEMA can visualise more than just the sequence-level information traditionally presented by alignment editor applications. Having access to the system's semantic model, it is able to relate other information associated with the sequence and display this in a a suitable graphical format. For example, annotations representing a set of contiguous extents on the residue sequence, such as conserved sequence motifs, could be drawn as coloured bars below the appropriate section of sequence. Similarly, annotations that consist of continuous values associated with individual residues could be drawn as a graph or spectrum of colours.

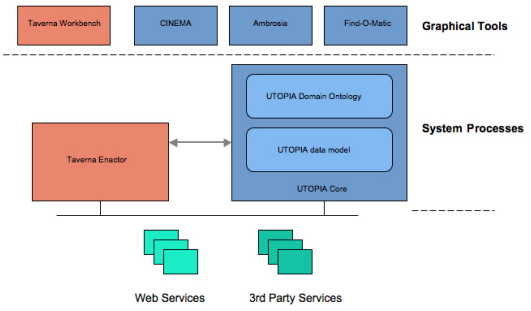


Figure 1. Combined architecture.

Ambrosia [2] is a 3D structure viewer, exploiting modern Graphical Processing Unit techniques to accelerate high quality high fidelity rendering of very large molecular models in real time. Ambrosia currently supports a number of representation styles, including Space Fill, Backbone, and 'Cartoon' rendering, and is able to overlay annotations from the semantic model on all of these.

Find-O-Matic provides an iTunes-like interface for discovering services and data objects. Simple keyword based queries are submitted to multiple databases and tools, with results being returned in a clean and unified format, which can then be sorted, arranged into 'playlists', and moved interactively to the other visualisation tools.

The Data Model

At the core of the UTOPIA system – and key to it's functionality – is a data model designed to be rich enough to capture the semantics of bioinformatics data in such a way that it can be exchanged between applications, and at the same time sufficiently light-weight such that it can be interrogated in real time to extract the objects required by the interactive visualisation tools [7]. To achieve this balance between richness and efficiency – and also for conceptual elegance – the model is split in to a number of orthogonal spaces.

First, a distinction is made between *structure* and *annotation*: concepts that are accepted as 'fundamental facts' within a domain, and concepts that annotate or enrich the knowledge of the structure in some way but are in themselves either 'received wisdom', fuzzy, or refer to a process or collection of structural concepts. Unlike in the physical and mathematical sciences where discoveries are axiom based, very few of the concepts in the biological domain can be thought of as absolute truths: beyond such things as atoms, bonds, residues and sequences the majority of biological features contain degrees of uncertainty or ambiguity that must be somehow represented within the model in order that it can be rendered as a visual object. UTOPIA's **structure space** is therefore quite small, and consists of four types of node: bonds, atoms, residues and sequences. All other concepts are mapped as *annotations* that project onto this structure space, and comprise **annotation space**. Each annotation may map to a single node in structure space, or to a set of nodes. An annotation may also have associated provenance. Optionally, a set of annotations may have an ontological structure projected onto it from semantic space to give it meaning in a particular domain or context and so that annotations can be classified and grouped in a hierarchy if appropriate. Finally, **variant space** represents uncertainty, conflict, and alternatives within a data set. A *variant* node maps onto a set of structural nodes that all maintain to represent the same data, and provides a mechanism for making any identifiable ambiguity or conflict explicit in the model.

The separation of these spaces allows their implementation to be tailored for their most common use in visualisation: a certain amount of computational reasoning may be required to infer that an Enzyme is-a-kind-of Protein so that it can be viewed in a sequence viewing tool; however the data structures and algorithms to support this must not interfere with the need to rapidly extract 10s of 1000s of annotations that form a systems biology graph, or the 100,000 or so atoms 30 times a second in order to be able to render a ribosomal complex as an interactive 3D structure.

This underlying model allows UTOPIA to gather and integrate data from a wide variety of heterogeneous sources and to generate a canonical internal representation that can be visualised by any of the front-end tools. Tools negotiate with the model using terms from the semantic space, e.g. 'can render sequences of residues with re-

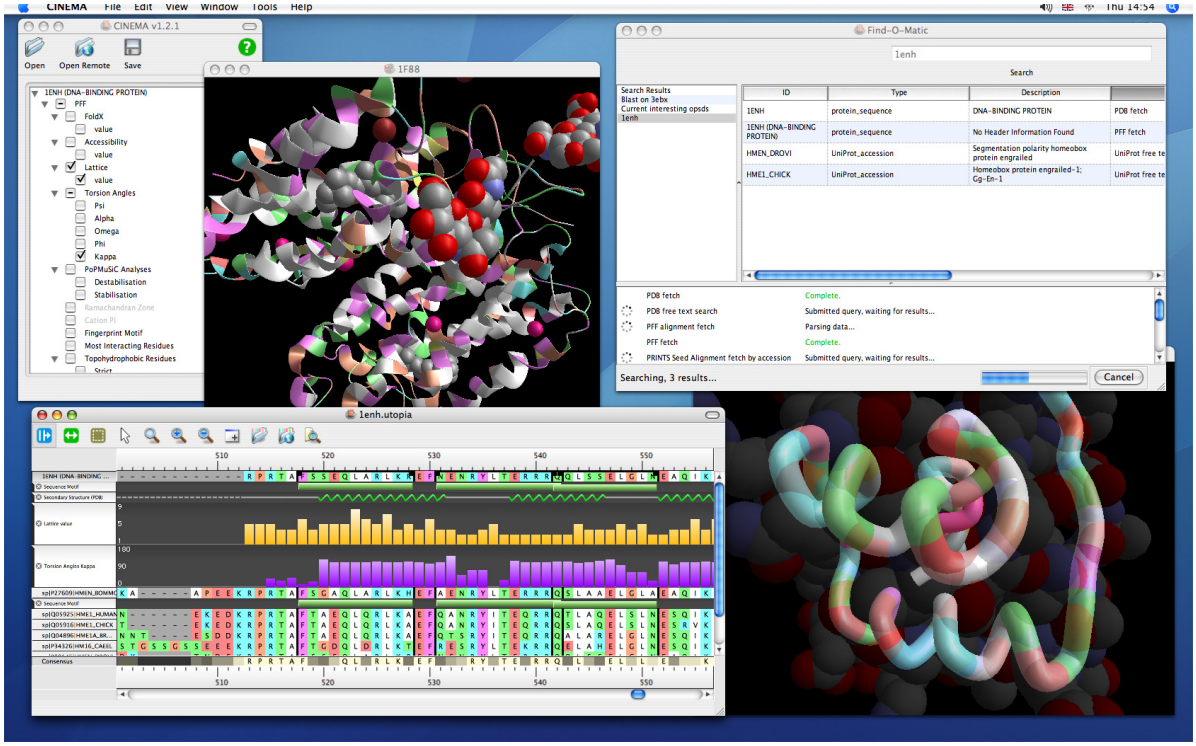


Figure 2: Using UTOPIA to examine a sequence motif.

gional annotations', 'can show a fingerprint motif' or 'can display a structure of atoms with regiona annotations', and thus do not have to be aware of file formats or the means of accessing remote sources of data. The richness of the model has two additional important features:

1. Multiple UTOPIA tools are inherently aware that they are viewing the same biological concept, albeit potentially in radically different forms (e.g. as a residue sequence, as a molecular structure, and as a frequency plot). Thus modifications made to the data in real time by one tool and injected in to the underlying model are immediately reflected in any other.

2. Biological concepts are exposed as 'first class citizens' in the interface itself, thus the tools are aware that the user has selected 'a sequence', 'an alignment of sequences', 'a signaling pathway', a 'cell compartment' and so on. This is especially important in terms of UTOPIA's integration with myGrid (`http://www.mygrid.org.uk`), and is explained in more detail in [10].

Access to Remote Resources

UTOPIA provides tools for retrieval, visualisation and interactive manipulation of biological data it does not of itself provide any algorithms or mechanisms for performing computational analysis of the data it manages: all such features are provided by third party software, which can be installed locally but is more easily accessed remotely via freely available web services or workflows provided via Taverna[10]. Plugin components called *conduits* connect UTOPIA to other sources of data and computation such as scripts, executable programs, web services and workflows. A controlled vocabulary is used to describe these, allowing UTOPIA to expose the functionality in appropriate parts of its tools' interfaces and to inject concepts into its model from the input and output formats used by the third party software.

UTOPIA in use

The kinds of functionality provided by UTOPIA are best described by an example. We'll start by finding some sequences. Lets say that were only interested in rhodopsins. We could just type rhodopsin into Find-O-Matic's 'search' box and see

what happens, but this is likely to return far more results than we care about, since Find-O-Matic searches multiple databases. Instead, we'll limit the results a little by searching for sequences with Swiss-Prot rhodopsin identifiers – to generalise the search, we can use a wildcard: i.e., OPSD*. This search returns just over 100 results from UniProt and PDB; from the list, we identify OPSD SHEEP as the sequence of interest. We select OPSD SHEEP and, from the context menu, see that there is an option to find its homologues using BLAST[1]; accordingly, we invoke this service and wait for the results. Behind the scenes, a workflow is enacted that fetches the OPSD SHEEP sequence, sends it to a BLAST service and retrieves the results. From the returned list, we select the top few matches and drag them into the CINEMA alignment tool; this causes UTOPIA to fetch the sequences and any associated annotations from their respective databases. As were interested in relating sequence to structure, we also need to fetch an atomic model that represents the rhodopsin family. We know that the structure of bovine rhodopsin has been solved and was deposited in the PDB with identifier 1F88; so, using Find-O-Matic again, we can retrieve the combined sequence and structural information by dragging and dropping 1F88 into the set of sequences already collected in CINEMA (note had we not known the PDB identifier at this point, we could, instead, have used Find-O-Matic to search for OPSD BOVIN, which returns around 20 PDB codes). A quick inspection of the sequences reveals that they are already well aligned, owing to the high level of similarity between them, but that the C-terminal region is misaligned. In fact, only 1F88 is out of alignment, so we could use CINEMA to manually slide its C-terminus into place; instead, we'll invoke a service to do this for us. CINEMAs context menu knows that we are now looking at a set of related proteins, so it offers various alignment tools, like ClustalW[4] and MUSCLE[3]. We choose one of these; the sequences are transformed, sent to the remote service, and the results are retrieved and applied before our eyes to give a neat alignment all this without ever seeing a file format or Web service interface description. Now we want to analysethe aligned sequences in more detail: e.g., we want to know where the predicted transmembrane (TM) regions are and how these compare with the known tertiary structure, and we want to select the TM domains to form a

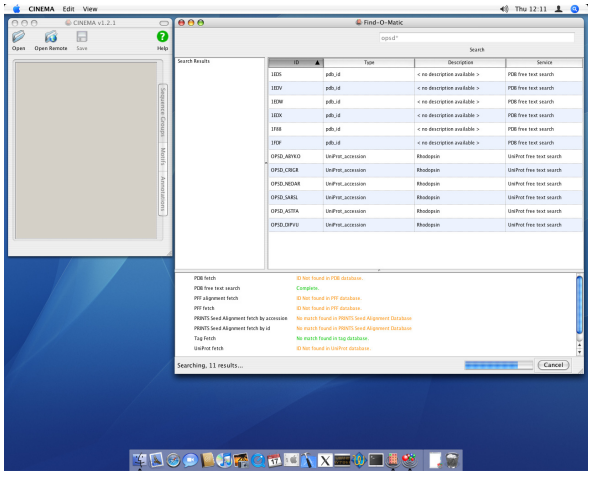


Figure 3: Search using Find-O-Matic.

characteristic signature or fingerprint for the rhodopsin family. Using the motif tool, we identify 7 hydrophobic regions in the alignment the colour scheme allows us to do this quickly and easily, as aliphatic residues are coloured white and aromatics are coloured purple (the colours chosen deliberately mimic those of standard 3D modelling components). We can now compare these regions with other features of our sequences by using UTOPIA's annotation manager to switch on graphical representations of features of interest, which have been retrieved from their source databases. Looking at 1F88, for which structural information is available, the context menu reveals that additional services are available. We invoke both the DSSP [5] program, which derives the molecule's secondary structure from the 3D coordinates, and the 'tmap' TM-prediction tool included in EMBOSS [11]. Use of these tools adds extra annotations to our graphical view, allowing easy comparison of data from multiple sources; more importantly, they allow us to compare predicted features with the experimentally derived 3D structure and hence to evaluate how good the predictions actually are. To do this, we invoke the Ambrosia molecular viewer and, from within CINEMA, instruct it to display the structure of 1F88, which appears in a new window. Although separate, the two tools are, however, linked, both

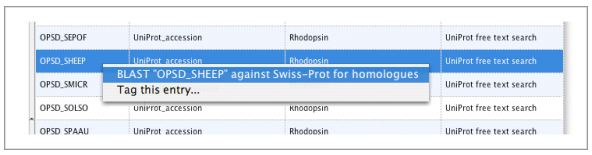


Figure 4: Running a service from Find-O-Matic.

viewing the same underlying data. Selecting one of the annotation lines in CINEMA demonstrates this point – e.g., Tmap's predicted TM domains are now highlighted on the 3D structure displayed in Ambrosia, and we instantly see the difference between the predicted and manually identified regions. Finally, aside from exploring specific features of the sequences and structure of rhodopsin, we wish to find out more about the function of these proteins, their disease associations, their wider kinship, and so on. To do this, we again invoke CINEMAs context menu (remember this offers tools that specifically work on multiple sequences, so there is no need to select all the sequences) and choose PRECIS[8]: this distils the annotation from each of the Swiss- Prot sequences in the alignment, and returns a structured report detailing the name of the protein family, relevant cross-links to information in related databases, literature references (linked to PubMed), keywords, and a set of notes describing the function of the proteins, relevant disease information, family hierarchy and additional structural information, and so on

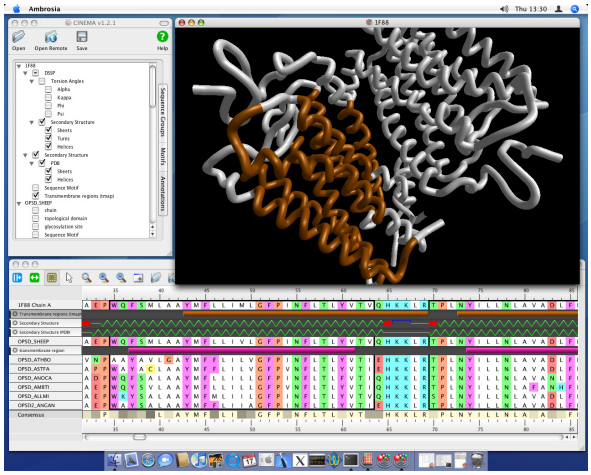


Figure 5: Comparing predicted features with structure.

Interfacing with UTOPIA

Interfacing applications and data sources with UTOPIA is done via plugins that connect to 3rd party tools and data sources, and translate their content into manipulations of the system's semantic model. Plugins simple to write but still very powerful, and currently can be coded in C++ or, more commonly, Python. The following Python excerpt uses the SMART web service[6] to anno-

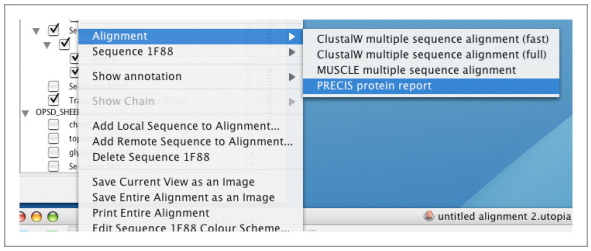


Figure 6: Generating a PRECIS report

tate a protein sequence and return it to UTOPIA. The 'description' method at head of the code advertises this plugin's functionality to UTOPIA using terms from its ontology, which allows UTOPIA to work out where to put this feature in its interface, and when to display it to the user (in this case, in any context menu where 'annotating' an object of type 'protein sequence' would be useful), and what inputs and outputs are required to invoke the service. A full tutorial on writing plugins is available on the UTOPIA website.

Future plans

Our current plans include the extension of CINEMA to deal with the display of genomic data, of Ambrosia into the area of chemoinformatics and of Find-O-Matic to further simplify and manage the fetching and storing of data. Additionally we are working on the development of new tools fo displaying metabolic networks and pathways, and for managing bibliographic information. We are very keen to work with the developers of tools and resources to help them integrate their data and applications with UTOPIA; please contact the development team via the website if you

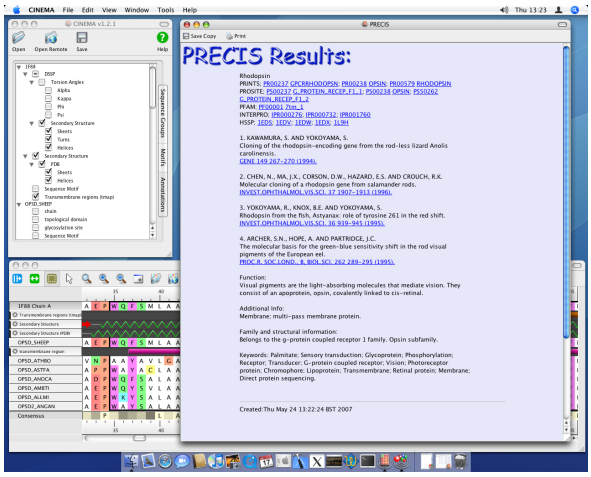


Figure 7: Displaying a PRECIS report.

```
from SOAPpy import WSDL

def description():
  return 'protein\ _ sequence', 'annotating',
'protein\ _ sequence', 'Annotate sequence
using EMBL SMART'


def invoke(source):

  # Get source > complex > protein >
sequence
  seq = source.childAt(0).childAt(0).
childAt(0)

  # Serialise sequence
  residues = []
  sequence = ''
  for i in range(0, seq.childCount()):
    residue = seq.childAt(i)
    residues.append(residue)
    sequence += residue.getAminoAcid().
getSymbol()

  # Get web service proxy
  proxy = WSDL.Proxy('http://smart.embl-
heidelberg.de/webservice/SMART _ webservice.
wsdl')

  # Run the SMART service
  features = proxy.doSMART(protein _
sequence=sequence)

  # Apply features
  for feature in features.feature:
    annotation = source.add('annotation')
    annotation['name'] = 'SMART features'
    annotation['class'] = 'extent'
    annotation['width'] = 1 + int(feature.
end) - int(feature.start)
    annotation['description'] = "Name: %s\
nType: %s\nE-Value: %s"
      % (feature.name, feature.type,
feature.e _ value)
    annotation.add(residues[int(feature.
start) - 1])

return source
```

Figure 8: A UTOPIA plugin written in Python

have software that you think would benefit from this kind of integration.

Acknowledgements

The development of UTOPIA was initially funded by a grant from EMBNet, the UK Engineering and Physical Sciences Research Council, and The Department of Trade and Industry, and work now continues under the auspices of the E.U. funded EMBRACE Network of Excellence (http://www.embracegrid.info).

References

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Mol. Biol.*, 215:403–410, 1990.

[2] D.Thorne and S.Pettifer. Unifying abstract and physical molecular model interaction. In *Proceedings of EGUK Theory and Practice of Computer Graphics Conference 2005*, pages 75–82. Eurographics Association, Jun 2005.

[3] R. C. Edgar. MUSCLE : multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.

[4] D. Higgins, J. Thompson, T. Gibson, J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positionspecific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673– 4680, 1994.

[5] W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 2004.

[6] I. Letunic, R.R. Copley, B. Pils, S. Pinkert, J. Schultz, and P. Bork. SMART 5: domains in the context of genomes and networks. *Nucleic Acids Research*, pages 257–260, 2006.

[7] P. McDermott, J. Sinnott, D. Thorne, S. Pettifer, and T. Attwood. An architecture for visualisation and interactive analysis of proteins. In *Proceedings of 4th International Conference on Coordinated and Multiple Views in Exploratory Visualization*, pages 55–65, Washington, DC, USA, 2006. IEEE Computer Society.

[8] A. L. Mitchell, J. R.Reich, and T. K. Attwood. PRECIS : Protein reports engineered from concise information in SWISSPROT. *Bioinformatics*, 19(13):1664–1671, 2003.

[9] S. Pettifer, J. Sinnott, and T.K.Attwood. CINEMA - a UTOPIAn sequence editor. *EMBNet.news*, 10(3), Jan 2004.

[10] S. Pettifer, K. Wolstencroft, P. Alper, T. Attwood, A. Coletta, C. Goble, P. Li, P. McDermott, J. Marsh, T. Oinn, J. Sinnott, and D. Thorne. myGrid and UTOPIA: an integrated approach to enacting and visualising in silico experiments in the life sciences. *Lecture Notes in Bioinformatics*, 06 2007. accepted for publication.

[11] P. Rice, I. Longden, and A. Bleasby. EMBOSS: The european molecular biology open software suite. *Trends in Genetics*, 16(6):276–277, June 2000.

Linux for bioinformatics:

dedicated distributions for processing of biological data – Part 1: Live distributions

Antonia Rana[1]

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

antonia.rana@jrc.it

Introduction

The number of active, widely used and valuable bioinformatics projects at open source software repositories such as bioinformatics.org and sourceforge.net is constantly increasing. Besides historical tools used for the analysis of biological data before Linux became a viable option on the desktop, new projects are started and new tools are being made available to improve the ability to collect, analyse and integrate large collections of data. The development of new or improved algorithms for the analysis of genomic data are fostering the development of new tools. On the other hand, the availability of open source tools with the full access to algorithms and source code and the possibility to modify and improve them, is encouraging a sort of good scientific practice in providing new tools and promotes reproducible research [1]. Important efforts in the past years have been dedicated to making access to data easier and facilitate their analysis improving our knowledge of biology. These efforts have been targeted, for example, at harmonization through the definition of ontologies [24], at integrating databases [23], at devising mechanisms to overcome the typical pattern of creating manual ad-hoc connections among software tools and databases, cutting and pasting queries, creating temporary files and taking notes by providing workflow mechanisms and single access point portals which link smoothly the actions and facilitate reproducible research. [3] [5] [7] [8] [20]. The availability of these "enabling" tools allows biologists to focus on their research without being distracted by computer problems.

In parallel to these efforts, the availability of easy to use live distribution of the GNU/Linux operating system, has facilitated the development of ready made desktop (and server based) solutions which collect in one single CD or DVD an entire operating system equipped with tools for bioinformatics analysis as well as with development environments enabling the further development of new tools.

This paper is the first of a two parts survey which reviews the most popular solutions available in the Linux arena providing a desktop environment equipped with applications and development libraries for life scientists. The first part of this paper is organised as follows: the first section provides an introduction to live distributions dedicated to bioinformatics and the second section provides a review of eight different live distributions; the second part of the paper will cover the packages repositories providing ready-to-use bioinformatics applications for linux, complete distributions, including those which may be deployed in a cluster based environment and will discuss the results of the review.

Ready made Linux solution for a bioinformatics workbench

Most bioinformatics software has been available historically on Unix platforms and has naturally migrated to Linux. Most of these programs are available in source code and need to be compiled and installed. Installation, particularly with software with graphic user interface, can be a painful exercise for non experienced-users. They can have dependencies on libraries or include files which are installed in different places on dif-

1 Disclaimer required under the terms and conditions of use of the Internet and electronic mail from Commission equipment:
"The views expressed are purely those of the writer and may not in any circumstances be regarded as stating an official position of the European Commission."

ferent unix flavours. Having a program available as a package ready made for the Linux distribution of choice, makes installation of a new application an easy exercise. For this reason, many have started to make available bioinformatics software in package format and also to pre-package the whole Linux distribution with the inclusion of popular bioinformatics applications.

After a brief search on the Internet, we have found out that there are at least nine such dedicated Linux distributions, plus a number of package repositories which provide ready made bioinformatics packages for a number of selected, popular Linux distributions, which are being used routinely by researchers. Most of these distributions are so-called "liveCD" whose main characteristic is that they do not require installation on the hard disk, but can be run directly from a CD or DVD.

The choice of Linux as the underlying platform on which to run bioinformatics applications seems to be the obvious one: Linux is free, freely available and free to study, modify, redistribute, it is highly scalable and modular, it supports multiple hardware architectures, has a rich development environment and a strong user community. There is a huge number of scientific applications available on this platform: sourceforge.net, the largest development and download repository of open source code and applications, lists currently more than 15000 projects on scientific applications and more than 1000 dedicated specifically to bioinformatics.

In this review we will explore the features of eight Linux distributions dedicated to bioinformatics, and we will briefly describe some of the package repositories. All the distributions will be examined here from the point of view of the desktop user. In the second part of this article, we will examine some distributions that can be installed on the hard disk and used in cluster environments.

LiveCD(DVD) distributions

A LiveCD or Live Distribution is an operating system that is executed upon boot from a bootable removable storage media, without being installed on the hard disk. Live distributions are usually available on CDs or DVDs, but recently they are being made available also on USB flash drives. They are called "live" because they are in fact a complete and runnable instance of the operating system rather than a collection of packages that must first be installed in order for the operating system to be used.

An important characteristic of live distributions is that they do not modify in any way the operating system already installed on a computer. When they start, they perform auto-detection of the system hardware and place the files that would be stored on the hard disk into temporary memory, the so called ramdisk. Of course, this means that the amount of memory available to run applications is reduced and therefore, in order to perform well, these distributions should be run on computers with at least 256 MB of RAM.

Because they do not require installations and have powerful auto-detection mechanisms, live distributions are not only a good way to demo or preview an operating system, but also a good solution for users who do not have much technical expertise and would not embark into installing an operating system. As such, they are useful in involving more potential users to test drive the system and can be uses on borrowed computers. In addition, they have become more appealing with the increase in the speed of CD/DVD drives.

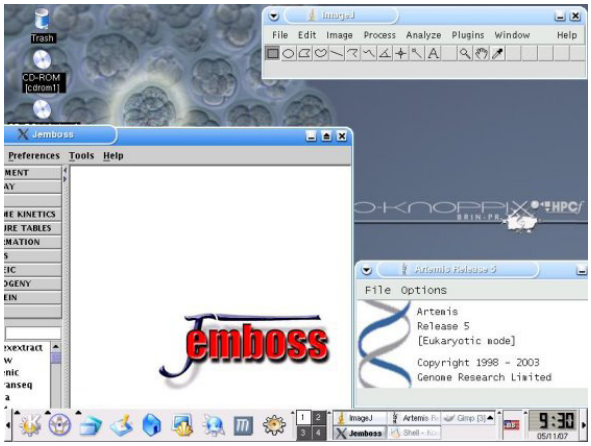
All but one of the live distributions dedicated to bioinformatics that will be examined here are remastering of the well-known, Debian-derived Knoppix distribution. "Remastering" is the process by which a base live distribution is modified with the installation of additional packages (and possibly the removal of packages useless for the specific needs). Knoppix was first released in 2003 and has found usage both as a rescue disk system to repair operating system installations and as a primary distribution in its own right. Since 2003, the popularity of LiveCDs has increased substantially and has become popular in other application field than the initial system and network administration use.

Being run in RAM, one might wonder how the work being done, the results of the analysis, can be saved. Knoppix scans available peripheral devices at boot time and provides for a mount point for partitions on the hard disk, as well as

floppies and USB memory devices. So, saving the results of one's work means to copy files on USB flash cards or floppy disks. If the computer is networked, another option is to save on network resources. As of version 5, knoppix also mounts safely NTFS partitions in write access as well as read (this option should be used with caution, though) In addition, knoppix provides scripts to create a "portable" home directory, as well as to save configuration or extensions options, on a USB memory device or on a hard disk partition. Booting with the USB device plugged in, causes knoppix to load the personalised configuration and the user's own home directory. Knoppix based systems can also be installed on the hard drive. Although this is not recommended, this feature provides for a means installing a Debian-based system in an easy and painless way.

One word of caution on liveCD is related to security updates: in order to install new versions or patches to the systems, the liveCD must be remastered. This aspect can be important for systems connected to an open network.

BioKnoppix



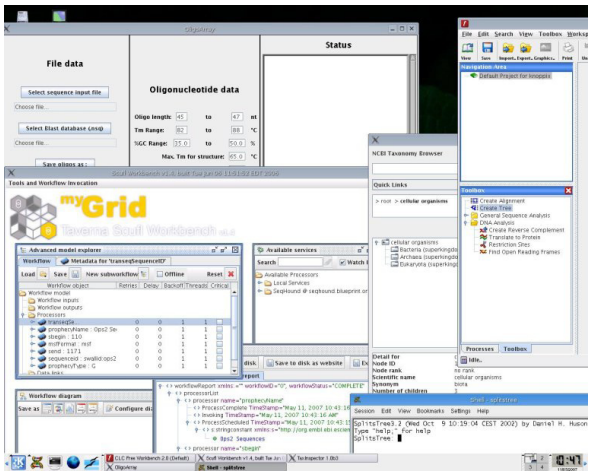
Bioknoppix is a Knoppix based LiveCD developed by High Performance Computing Facility at the University of Puerto Rico, with focus on molecular biology and bioinformatic. This distribution, available on CD, is listed as an official knoppix customization on the knoppix web site and can also be purchased for a nominal fee at cheapsites.com or linuxcd.com. The objective of its author was: *"To have a working environment attractive to the life science community"*. As most of the distributions described here, it has been released in 2004. It is documented on a plone-based portal featuring a short tutorial with the list of applications installed on the CD, a user forum and news, however, activity on this web site does not seem to be very intense. According to the new section provided in the main page, the last update was made in 2004.

Bioinformatics applications can be started using the "**Biology tools**" menu item which is available directly under the main KDE menu. Some of them have been installed as packages and can be listed using the KDE standard Kpackage package manager available in Linux system using the KDE desktop environment, not all of them can be started from the menu. When the system starts, a web page is opened with information on knoppix, while no introduction or "getting started" information is provided to the novice user. It would be more useful to have information about the contents of this distribution how to use the bioinformatics programs which it provides rather than information on knoppix in general. However, the definition of a menu item through which most of the bioinformatics applications can be started is very useful. Bioknoppix is also useful to those scientists who need to implement their own programs for their work, as it includes a full development environment (C, C++, Python, perl, BioPerl). Not all dedicated distribution provide a full office suite, bioknoppix provides OpenOffice

Bioinformatics application installations are located in several directories (e.g. `/usr/local`, `/usr/share`), not all of them can be started from the graphical menu. Bioknoppix is a popular distribution, although a bit out of date. However it provides the most popular bioinformatics programs and access to the applications via the graphical interface certainly helps the novice user.

Home page: `http://bioknoppix.hpcf.upr.edu/`
Current version: 0.2.1 last update: 25-9-2005
Base system: knoppix 3.3
Kernel version 2.4.22
Media: CD
Wireless: not recognized

Bio-Linux



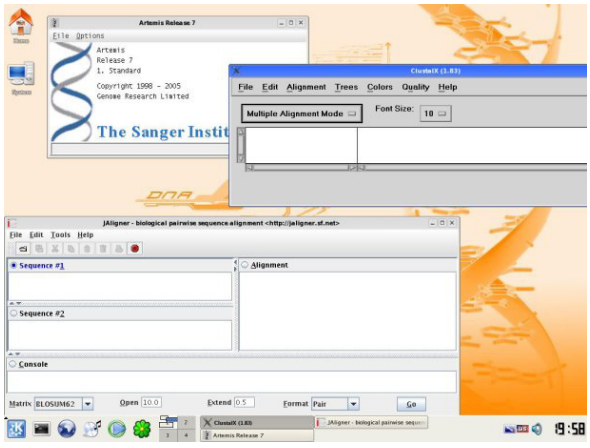
Bio-Linux is distributed as a LiveDVD by the NERC environmental bioinformatics centre in Oxford. It is a rich distribution based on Debian with the addition of about 60 popular bioinformatics packages, including programming libraries. A lot of attention has been paid to user friendliness, through the provision of a comprehensive, categorised and searchable documentation system for bioinformatics software which can be started by clicking on the "Bio-linux" icon on the desktop. The documentation system is served by the apache web server, is categorized and information on the packages can be searched by name or by category. For each application, information such as name, description, homepage, access to documentation available locally and on the internet (remote documentation) are provided. In most cases information is available locally and documentation pages are very informative, which is quite useful in case the PC is not connected to the internet. The main page also contains a section on tutorial and courses. However, most of these sections, at least for the LiveDVD that was examined, were empty.

Access to the bioinformatics applications is also quite easy. They can be started from a customized menu, that is located on the right of the standard KDE menu at the bottom of the screen. Within this menu, the EMBOSS suite programs are accessible via a hierarchical submen. This kind or organization avoids crowding one single menu item with too many programs making it easier to locate the application of choice. The same list of applications is also accessible from the main KDE menu under the submenu *"education -> science"*.

Most of the bioinformatics software is located in the directory `/usr/local/bioinf`. Generic applications installed on Bio-linux, include OpenOffice 2.0beta and a tool to configure the TV card. The system starts automatically at startup the apache service which serves the pages of the documentation system, the postgresql and ssh services are also started. USB device is mounted even if a flash card is not plugged in when the system is started. This is a complete and well organized distribution, which is easy to use also for novice users thanks to the well thought organization of applications and documentation. The website is also very informative and full with very detailed information on the use of the packages provided in the distribution. Bio-Linux is also available in an installable version, which comes with a list of default packages and a list of optional packages that can be downloaded from the web site and added subsequently to installation. The Bio-Linux LiveDVD v1.4 has also been tested on a laptop with wireless access and the wireless card was configured without any problem.

Home page: `http://envgen.nox.ac.uk/bio-linux.html`
Current version: 1.4
Base system: Debian
Kernel version: 2.6.12
Media: LiveDVD
Wireless: Supported

DNALinux



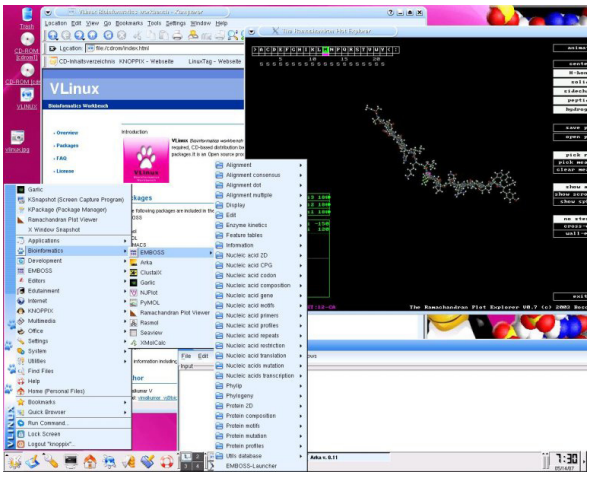
DNALinux is developed and distributed by the Universidad Nacional de Quilmes in Argentina. The current release is 0.5beta, released on October 2005. Unlike the other distributions reviewed in this paper, it is not based on the popular knoppix distribution but on Slax, which is a live distribution based on Slackware. On their website, they also distribute a server version containing the full genome of H5N1 Influenza A virus in BLAST format. A Support Forum for discussion on

DNAlinux is provided on on LinuxQuestions, a free, general linux forum, where the DNAlinux Project Managers reply to questions (it can be accessed at this URL: `http://www.linuxquestions.org/questions/forumdisplay.php?f=53`). DNAlinux is more text-based than the other linux distributions dedicated to bioinformatics. The system starts in text mode and requires login. Root login credentials provided are provided in the start screen together with an indication of the directory where the bioinformatics software is located (`/biosoft`). After logging in, it is anyway possible to start the graphical desktop environment based on KDE, by typing the `startx` command.

Within the graphical desktop environment, a number of applications can be started from the *Bioinformatics* menu added to the main KDE menu (Abacus, Apollo, Arka, Artemis, CLustalX, JAligner, Njplot). For the remaining applications available in `/biosoft`, the path to execute the programs does not seem to be set, therefore it is necessary to type the full path or set the path variable. This distribution shows some evolution and attempt to update the bioinformatics applications installed, and seems to privilege those users who are familiar with the text based environment. This is important if memory resources are not sufficient to run the system in graphical mode with an acceptable performance. Some efforts have also been made at improving the performance of the on-the-fly decompression of the system files on the CD. The distribution would benefit from the addition of "getting started" introduction to the features of the applications installed in this system.

Home page: `www.dnalinux.com`
Current version: Desktop version 0.5beta released on Dec 2005, but updated recently with recent versions of blast (2.2.10), emboss (3.0.0) e clustal (1.8.3) e JRE (1.5).
Base system: Slax 5.0.6
Kernel version: kernel 2.6
Media: LiveCD
Wireless: device is detected, but not configured properly

Vlinux



Release 1.0 of *Vlinux "Bioinformatics workbench"* is distributed under the GPL license on the well known bioinformatics.org website. A new portal providing information, including support forum, mailing list and news introduces the distribution. However, it does not seem there is a lot of activity on it. Like most of the live distributions dedicated to bioinformatics, it is based on knoppix, on a version which is by now a bit old, however, the hardware auto-recognition and configuration procedure performed better than the other distribution based on knoppix 3.3 providing a graphical desktop environment with higher resolution.

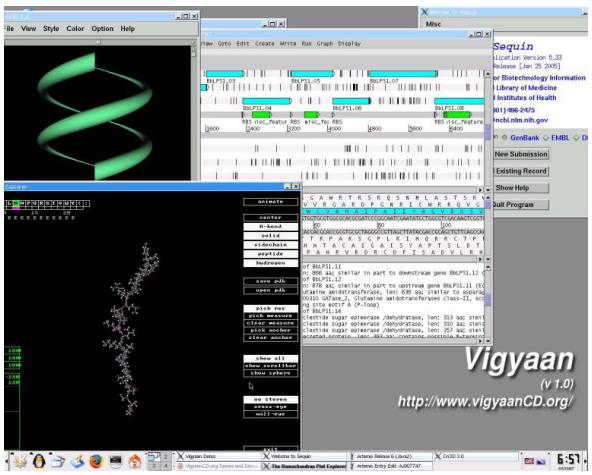
When the system starts, the konqueror KDE browser opens the file `/cdrom/index.html` which displays a web page with information on Vlinux and the bioinformatics packages installed on it. In addition, a brief description of each package installed is also available in `/cdrom/KNOPPIX/packages.html`. This page is also linked from the initial presentation page displayed at system startup. For each application a brief description consisting of name, version, usage and links to tutorials. Where available, a link to user manuals in HTML or PDF is also given. The description page

also has an FAQ which contains information on how to use the CD. This kind of info/presentation is very useful when loading the system for the first time as it provides a way to navigate through the possibilities offered.

Bioinformatics software can be easily found as it is located in `/usr/local/bioinfo`. Access to the tools is facilitated by the dedicated menu item *"Bioinformatics"* in the KDE menu which includes the applications: EMBOSS, which is a submenu giving access to all the programs within the EMBOSS suite, Arka, ClustalX, Garlic, Njplot, PyMol, Rachmandarn Plot, RasMol, SeaView, XMolCalc. Easy and well identified location for the bioinformatics packages as well as the documentation provided at startup which also functions as a map of the software available are very useful in this distribution, which, although not frequently updated, is quite easy to use.

Home page: `http://bioinformatics.org/vlinux/news.php`
Current version: 1.0
Base system: knoppix 3.3
Media: LiveCD
Kernel version: 2.4.24
Wireless: not recognised

Vigyaaan



Vigyaan is introduced as an *"Electronic bench for bioinformatics, computational biology and computational chemistry. Designed to meet the needs of both beginners and experts"*, and indeed, this distribution is quite well presented and documented in comparison to the others examined here. The LiveCD is distributed under the GPL license by Pratul K. Agarwal on the www.vigyaancd.org website. The website itself is very informative containing quite detailed description of the distribution and its software contents and features.

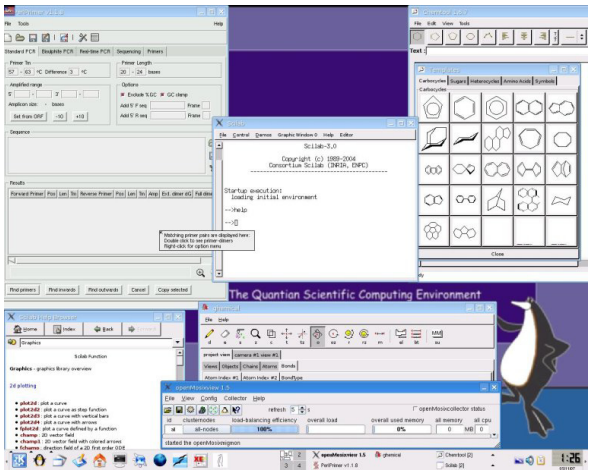
At startup it presents itself as the *"biochemical software workbench"*. The *"About Vigyaan (start here)"* icon on the desktop starts the documentation page opening the Firefox browser on the page `file://cdrom/index.html`. This page contains a description of VigyaanCD, very similar to the information provided on the home page on the Internet. The software installed is separated into categories for easy accessibility (*"Biology tools"*, *"Chemical tools"*, *"Other tools"*) and for each application, description, link, version, license and notes are provided in this page. In addition to the documentation, demos and various tutorials are provided on the CD, which help novice users at learning how to use the tools. Each demo is available in a separate directory associated to the relative tool. A readme file on the demos and an icon which starts a shell to try the demos are available on the desktop. Another icon on the desktop helps users in creating a persistent home directory on a USB flash memory or on a partition on the hard disk so that saving the scientists work on persistent memory is facilitated.

Bioinformatics applications are accessible via three menus under the *"vigyann"* menu which is, in turn, a submenu of the KDE main menu. The three menus map to the categories used in the documentation, i.e.: a submenu named *"Bio Software"* provides access to biology tools NCBI Tools, Ramachandran plot viewer, Arka/GP, Artemis, ClustalW/ClustalX, Cn3D, Garlic, Ghemical, Jmol, NJPlot , PyMOL, Rasmol, Seaview, EMBOSS tools; a submenu named *"Chem Software"* provides access to chemical tools: Ghemical, Jmol, RasMol and XDrawChem; and a submenu named *"More tools"* provides access to other scientific tools available in the distribution such as: R, Gnuplot, Octave, the GIMP, Xmrg and gnumeric. A separate menu item is provided for EMBOSS. *NCBI Tools* and *EMBOSS Tools* are again submenus items which provide access to

all the programs contained in the two program suites. Compilers for C, C++ and forrtan, as well as python and perl with the bioperl libraries are also installed making this distribution a complete environment for development. In order to make room for all these tools some disk space hungry applications usually found in knoppix based distributions such as OpenOffice are not installed, which makes this distribution really compact and focused on scientific applications as well as very useful and user-friendly for beginners.

Home page: `www.vigyaancd.org`
Current version: v1.0, released 20050907
Base system: knoppix 3.7
Kernel version: 2.4.27
Media: LiveCD
Wireless: not recognised

Quantian



Quantian is a bit different from the other distributions analysed here. Its main focus is on numerical and quantitative analysis and it also includes computer algebra systems. It started back in 2003 with version 0.1 with the initial proof of concept release based on knoppix 3.3, has gone through several releases until its current version 0.7.9.2 dated 26 February 2006 and based on knoppix 4.0.2. More precisely, the current version of Quantian is based on a derivative of knoppix, clusterKnoppix, which is a modified knoppix distribution that includes the openMosix extension to the Linux kernel that turns a network of ordinary computers into a supercomputer for Linux applications. The support for openMosix allows Quantian to be deployed in a cluster environment, making it easy to turn a network of old computers into a cluster farm for CPU intensive applications. However for openMosix to work properly, version 2.4 of the kernel needs to be loaded instead of the default 2.6. version. This requires users to indicate the kernel option at boot. Documentation on how to do this change is provided on the Quantian mailing list.

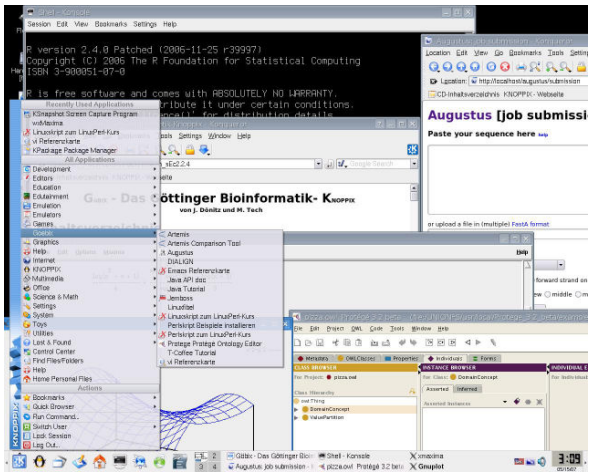
Information on the website is very exhaustive and includes support forum (blog), a mailing list and various HOWTOs as well as todo list and changes history, giving the idea that this distribution is very well supported. Quantian has mailing list discussing topics related to the distribution, which seems to be quite active. Issues related to the functioning of the distribution in cluster environment are often discussed and replies to requests are prompt.

Added packages include R and Octave, the GNU Scientific Library, the Maxima, Pari & Ginac computer algebra systems, BioConductor. The system starts with a clean desktop which only contains icons for mounted file systems (hard disk partitions, USB memory stick and floppy). No information or documentation page is displayed on startup or made available from a link on the desktop. It is difficult also to locate the bioinformatics tools installed, as no specific directory is used to collect them. However, the bioinformatics applications can be started from the KDE menu, under the submenus *"education -> science"*. All applications are accessed via this menu without any hierarchical submenus structure. Bioinformatics tools seem to be a recent inclusion in Quantian, which includes the whole DebianMed packages. Its initial focus was numerical computation and linear algebra and the choice of packages in this field is very impressive. It includes the free software environment for statistical computing R with all packages from the CRAN repository, all packages from the BioConductor projects, Octave, the GNU Scientific Library, the Maxima, Pari & Ginac computer algebra systems. Among the popular bioinformatics packages, EMBOSS seems to be missing. Being distributed on DVD, it does not suffer from space limitations, therefore, it includes the complete OpenOffice suite in addition to many other generic applications and

games distributed usually with knoppix. Quantian is also available on a commercial basis from third party distributors.

Home page: `http://dirk.eddelbuettel.com/quantian.html`
Current version: 0.7.9.2, release date: 2006-02-26
Base system: knoppix 4.0.2
Kernel version: 2.6.12
Media: LiveDVD
Wireless: supported

GöBIX



GöBIX is the most recent of the distributions dedicated to scientific computing examined here, being released in 2007. It is provided by the Göttingen University and used as an aiding tool in teaching. The documentation in the website is available only in German. It describes the distribution and the packages for scientific computing that have been added to the knoppix 5.1 distribution upon which it is based. Bioinformatics applications include the popular Artemis, BLAST, Emboss with the Jemboss java interface, PHYLIP, T-Coffee, BioPerl, BioJava, BioPython, ClustalW, but also tools for numerical computations such as Octave, SciLab, and an editor for ontologies, Protégé, are included.

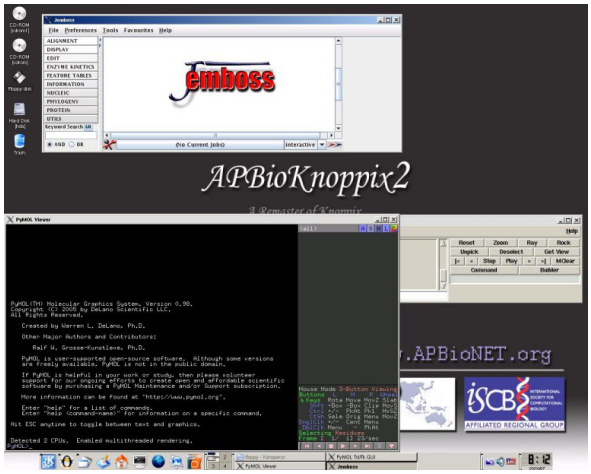
When the system starts, the Konqueror web browser opens the page `file://cdrom-index.html`, which provides an introductory description to the system containing basically the same information as the page on the website in German. Some of the applications can be accessed from the *GöBIX* menu available under the main KDE menu (i.e. artemis, Augustus, DiAlign, Emacs, Jemboss, Protégé, t-coffee).

Other applications are distributed in more than one different menus, this is the case, for example, for clustalX, RasMol and Ghemical, which can be started from the submenu *"education -> science"* or GNU R which can be started from the submenu *"utilities"*. This can be a bit confusing for beginners.

The collection of software in this system is really impressive, the distribution uses all the space available on the DVD being more than 4GB in size. An impressive collection of development environments is also installed, including eclipse, the perl, python and ruby interpreters and libraries, the MONO development platform, the Quanta plus web development environment, as well as the usual C, C++, fortran, java, etc. compilers. Office production suites are also included (OpenOffice 2.1 in addition to Koffice). Finally, applications for numerical computation and linear algebra including R, Octave, Maxima and Ginac are also available. However, one feels a bit lost in the maze of the hundreds of applications provided, without the guide of an introductory "getting started" documentation. But this is a young distribution, developed in support of some teaching activity, and, in fact, it is not clear whether it is also meant to be used by a more general audience and it is likely that it is supported by local documentation.

Home page: http://www.bioinf.med.uni-goettingen.de/teaching/goebix-dvd/goebix-inhalt/
Current version: unknown
Base system: knoppix 5.1.1, releasedate: 2007-01-04
Kernel version: 2.6.19
Media: LiveDVD
Wireless: supported

APBioKnoppix2



APBioKnoppix as been developed and released as part of the AsiaPacific BioGRID initiative. It is a remaster of knoppix 4.0.2 and is being used primarily for teaching and training on bioinformatics and biocomputing in several Universities in South East Asia. Its current version, which updates APBioKnoppix, is 1.0.2. The information on the website provides very detailed instructions on how to install the live CD on hard disk, on different options for running the liveCD with different levels of speed and performance and with instruction on how to remaster the CD. Remastering a live distribution is an operation which can be useful for those users who wish to upgrade and update the list of packages in a permanent way. The detailed description on the available mechanisms to run the system including running from hard disk in order to improve performance are useful if one considers that knoppix based system operate on the fly decompression from the CD to RAM before running any software, which is why there is a short delay from the moment an application is launched to the time when the application starts on the screen.

APBioKnoppix provides a wiki based documentation system that requires that services such as apache, webmin and mysql run on the system. These services are in fact started at boot time. However this page is not started automatically at the system start, so it is quite difficult to find it. In fact it can be found by starting the Mozilla Firefox browser and opening the "Edit GTD Wiki" tab which opens the starting page for the documentation, located at `/home/knoppix/public_html/course.html`. Unfortunately, the documentation system does not seem complete: the links often lead to empty pages.

Bioinformatics applications can be started from the menu *"Education"* under the main KDE menu. They include artemis, ClustalX, Jalview, Jemboss, Pymol and RasMol. As it often happens also in the other distributions, not all applications are installed as packages, some of them are installed with the *"configure-make-make install"* mechanism and are located in several different directories. This distribution adopts a familiar windows-like, look and feel.

Home page: `http://www.apbionet.org/grid/apbioknoppix2/`
Current version: 1.0.2
Base system: knoppix 4.0.2
Kernel version: 2.6.12
Media: CD
Wireless: supported

Package repositories

Package repositories which distribute bioinformatics applications packaged for easy installation on the most commonly available Linux distributions are an interesting alternative for those who have already an installed Linux base. The most common of these will be examined in the second part of this article. They include initiatives such as BioLinux, Debian-Med, BioRPMs and others.

Complete systems

In addition to live distributions and packages repositories, there are a few complete distributions that are meant to be installed and work as the main (only) operating system. These too will be reviewed in the second part of this article.

Conclusion

As a conclusive remark, the second part of this article will include a comparative discussion on all linux-based environments tailored for bioinfor-

matics including live distributions, package repositories and complete systems.

References

[1] Quackenbush, J. Open-source software accelerates bioinformatics, A report on the Wellcome Trust/Cold Spring Harbor Genome Informatics meeting, Cold Spring Harbor, USA, 7-11 May 2003, http://compbio.dfci.harvard.edu/pubs/CSHL_Bioinf_report.pdf

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

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

[4] Eddelbuettel, D. Quantian: A Scientific Computing Environment, Proc 3rd Int Workshop on Distributed Statistical Computing (DSC 2003), March 20–22, Vienna, Austria ISSN 1609-395X

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

[6] Knoppix, http://www.knoppix.org

[7] Fristensky , B. BIRCH: A user-oriented, locally customizable, bioinformatics system, BMC Bioinformatics 2007, 8:54

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

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

[10] Gearing Up for Bioinformatics, Canadian Bioinformatics Helpdesk Newsletter, March 3, 2005, http://gchelpdesk.ualberta.ca/news/03mar05/cbhd_news_03mar05.php

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

[12] BioKnoppix, http://bioknoppix.hpcf.upr.edu/

[13] Bio-Linux, http://envgen.nox.ac.uk/biolinux.html

[14] DNALinux, http://www.dnalinux.org

[15] Vlinux, http://bioinformatics.org/vlinux

[16] Vigyaaan, http://www.vigyaancd.org/

[17] Quantian, http://dirk.eddelbuettel.com/quantian/

[18] GöBIX, http://www.bioinf.med.uni-goettingen.de/teaching/goebixdvd/

[19] APBioKnoppix, http://www.apbionet.org/apbioknoppix

[20] Shannon, P. T., Reiss, D. J., Bonneau, R., and Baliga, N. S, The Gaggle: an open source software system for integrating bioinformatics software and data sources, BMC Bioinformatics. 2006; 7: 176

[21] Gilbert, D., 2004, Bioinformatics software resources, Briefings in Bioinformatics, Vol 5, No 3, 300-304

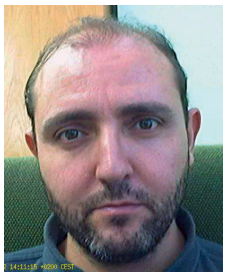
[22] Wren, J. D., 2004, 404 not found: The stability and persistence of URLs published in MEDLINE, Bioinformatics 20, 668-672

[23] Stein, L. D., 2003, Integrating Biological Databases, Nature Review, Vol 4, May 2003

[24] Gene Ontology Consortium, The Gene Ontology (GO) project in 2006, Nucleic Acids Research, 2006, Vol. 34,

Keeping your data up to date.

Part I: mirroring data



José R. Valverde

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

Introduction

For years, one of the key roles of EMBnet has been to provide access to updated databases for our community, both to end users and to other institutions by acting as mirrors for the original sites. At EMBnet/CNB we need maintain an up-to-date copy of the main databases that is updated daily. This work is not that much different from the one that systems administrators and users have to do at their own sites to keep their local data updated (either from their local EMBnet node -preferred- or from original sources).

In our case, we have a special user (**netadmin**) in charge of updating the databases. This is a lot of work best dealt with automatically. In this article we'll reveal the methods employed at EMBnet/CNB to maintain the databases. It has been split in two parts, the first, published in this journal issue, gives you an overview of the strategy used to keep data up to date. The second part, which will be published in the next issue, deals with the task of maintaining processed data and indexes. We are aware that the approach we'll describe is prone to numerous enhancements (mostly it should include mailing the user in case of errors), but should be enough to get you on track.

Mirroring the databases

We originally started using "mirror", a Perl script that became at one point the standard replication mechanism in the Internet. Although we still use 'mirror' for some things, we currently tend to favor mechanisms that are easier to set up and maintain.

One of the simplest methods available is **wget**, which we'll describe in more detail next.

Getting the code

You may get the code for our wget mirroring setup and scripts from our site if you wish. Just download
`http://www.es.embnet.org/~jr/embnet/scripts/wget-mirror-system.tgz`
please, note that it may be outdated by the time you get it (database setups have an habit of changing continuously).

Mirroring with wget

Our **wget** based replication system uses a special script to launch the downloads from the target sites. This script simplifies handling mirroring tasks with wget:

- **do_mirroring** takes a single argument: the name of a directory containing files with URLs. Each file in this directory will be read and the URLs contained within will be replicated locally.

We normally run this script as a periodical **cron** job using various periodicities, e.g.:

```
30      6       *       *       *
(cd /u/sysadmin/netadmin/wget ; ./
do_mirroring daily >> /u/sysadmin/
netadmin/logs/daily-wget.log 2>&1)

30      6       *       *       6
(cd /u/sysadmin/netadmin/wget ; ./
do_mirroring weekly >> /u/sysadmin/
netadmin/logs/weekly-wget.log 2>&1)

30      6       1       *       *
(cd /u/sysadmin/netadmin/wget ; ./
do_mirroring monthly >> /u/sysadmin/
netadmin/logs/monthly-wget.log 2>&1)
```

These entries belong to user **netadmin**, the user in charge of maintaining all local network services. Here, `/u/ssyadmin/netadmin` is the home directory of this user. The *wget replication system* is maintained within a subdirectory called 'wget' at his home.

As you can see, the script is run with a single argument: 'daily', 'weekly' or 'monthly'. This argu-

ment is just the name of a directory and has no other effect on the script behavior. We simply use these names so it is easy to remember with which frequency the URLs contained are updated.

Additionally, the result is sent to a log file so it may be reviewed later in case of trouble. It should be noted that the script maintains its own set of independent logs for each replicated URL in a special directory named 'log'.

The meaning of this 'crontab' file is:

1. line 1: run the script every day at 6:30 using the contents of directory 'daily'

2. line 2: run the script every week using the contents of directory 'weekly'

3. line 3: run the script every month using the contents of directory 'monthly'.

In other words, the script will be run periodically updating the replicas specified by the chosen URLs (stored at each directory).

Where does data go?

To keep the system as simple as possible, we didn't want to deal with stating where data should be stored. We just wanted to indicate the origin URLs, maintained in files under the appropriate directories. Data goes all to a common site.

This requires us to consider two aspects: how do we separate the different replicas? and how do we know -by looking at the contents- where does each dataset come from?

Our solution was to use the full name of the original host and directory to identify the directory hierarchy of the replica, storing all replicas in a single common directory. This way each replica gets its own independent hierarchy, and the hierarchy reflects the original host and path. E.g.:

URL
`ftp://ftp.es.embnet.org/pub/some _ place`

Saved as
`ftp.es.embnet.org/pub/some _ place`

URL
`http://www.es.embnet.org/Doc/some _ document`

Saved as
`www.es.embnet.org/Doc/some _ document`

The only thing to be done is sending these directories to wherever we actually want to save them. This is easy to do by creating the hierarchy by hand the first time -when we add the URL to the system- and making sure the last directory in the hierarchy is instead a symbolic link to the place we actually want the replica data stored. E.g.:

URL
`ftp://ftp.es.embnet.org/pub/some _ dir`

Saved as
`ftp.es.embnet.org/pub/some _ dir`

Symbolic link
`some _ place → /data/ftp/pub/some _ dir`

URL
`http://www.es.embnet.org/Doc/some/document.txt`

Saved as
`Www.es.embnet.org/Doc/some/document.txt`

Symbolic link
`some → /data/www/EMBnet/Doc/some`

In order to make it easy to remember that we maintain a listing of base hierarchies leading through symbolic links to the actual replica location we have called the directory where the replicas are dumped 'linkdirs'.

Avoiding conflicts

Some sites are bigger than others, and some are huge. So huge that downloading or updating their replica may take longer than one day or the frequency used in some extraordinary cases (e.g. the complete EMBL release).

To avoid conflicts, the script creates a lock file that signals a replica is already going on. If the script is run before a previous instance has finished, it will detect the lock file and refuse to start the replication process.

The lock files are maintained in a subdirectory called '**lockdir**'. The exclusion mechanism is nothing extraordinary. Our example does not take many precautions to correct problems like an ongoing replication being aborted before termination (i.e. dies without removing the lock file). You can add this using the **trap** instruction on `sh` or `onintr` on `csh`. In the sample implementation, however, the lock will remain forever after an interrupt unless it is removed manually.

Finally, how can you use it?

1. Extract the contents of the package anywhere (and not this 'where').

2. To use this script all you need to do is adapt the sample 'crontab' shown above to suit your local setup and add it to your **cron** tables.

3. Whenever you want to replicate a URL, just go to the directory that reflects the frequency you want to use and create there a file containing the URL

4. Then go to the 'linkdirs' directory and make a hierarchy until the but-last directory you want to replicate (e.g. using `mkdirhier`). Change to this directory and create there a symbolic link that redirects the data replicated to the directory you actually want to use.

5. That's all, folks!.

Locally, we tend to follow the conventions of the 'mirror' package, naming the file containing the URL after the host (and maybe the packages) that will be mirrored. A file may contain one or more URLs (see the **wget** manual page). For instance, assuming that you want to replicate this same script from `http://www.es.embnet.org/~jr/embnet/scripts/wget-mirror-system.tgz`

1. Start by downloading it by hand and extracting it somewhere in your home directory

2. Adapt the crontab file to include the above lines (providing for you user home and the location of your installed wget-mirror-system) and update it:
```
# crontab my-new-crontab
```

3. Change your working directory to the directory of the mirror system:
```
# cd wget-mirror-system
```

4. Create the URL file:
```
# cat > daily/www.es.embnet.org-wms
<< END
http://www.es.embnet.org/~jr/embnet/
scripts/wget-mirror-system.tgz
END
```

5. Create a symlink to the actual location (e.g. `/src/net/mirror`)
```
# mkdirhier linkdirs/www.es.embnet.
org/%7ejr/embnet
# cd linkdirs/www.es.embnet.org/
%7ejr/embnet
# ln -s /src/net/mirror scripts
```

6. And that's all.

Troubleshooting

First, look at the crontab log and see what happened

- Did it run? Verify the last modification date of the log file (`ls(1)`)

- Did it finish? Verify the log contents (`tail(1)`)

- Is there any error?

If this is not enough, or if you need more details, then you have to look at the URL-specific log for the file we were replicating and that gave us trouble. This log is located in a subdirectory of the replication system named 'log' and has the same name as the URL terminated in '.log'. It is just an **wget** log file and so it should be easy to read and interpret.

If the above do not shed any light, try to replicate the **wget** command by hand and see what happens. As a last resort, try to download the file(s) by hand using a different tool (e.g. **firefox** or **ftp**).

The power behind pain

Vivienne Baillie Gerritsen

We feel pain for a reason. Either to be informed of something that is likely to hurt us more unless we turn our backs on it, or of something that has gone wrong inside us. It is a sensation that has been evolving over millions of years, from yeast to man. Pain is multiple. Understanding its vocabulary and intricate syntax can shed light on what it is, why it is and how it could be countered. Detected by receptors, the sensation of pain can be kick-started from any part of our body. The TRP receptors are a family of such receptors, activated by an array of pain stimuli. They can detect hordes of different noxious chemical compounds but also environmental sensations such as extreme heat and cold. One particular TRP receptor – TRPA1 – comes as a surprise because, unlike many of the other TRP family members, it can detect multiple sensations leading to pain, as opposed to only one.

"Hellbound", Matt Sesow

http://www.sesow.com

Over time and as a means of defense, Nature has devised the most diverse ways of hurting. Snakes spit venom. Nettles sting. Bacteria puncture. And dogs bite. However, deprived of the resources to sense pain caused by venom, or a nettle's sting or a dog's fangs, we wouldn't understand the warning that goes with it. Likewise, pain which is caused by something inside us has to be detected so that our attention can be drawn to it. To this end, pain receptors line our body's every nook and cranny, ready to send out a signal which will be relayed to our brain and translated into pain.

The Transient Receptor Potential (TRP) channels – or receptors – are pain receptors. Pain receptors are an essential part of the process which leads to the actual sensation of pain. The signal triggered off by a TRP receptor is sensed by nerve fibres which release neuropeptides that, in turn, inform the brain both of pain and inflammation. TRPA1 is one such receptor, known to be directly stimulated by the pungent components of mustard oil and garlic, which cause the familiar burning and pricking sensation we have all experienced. Besides mustard and garlic though, TRPA1 is also stimulated – though not directly – by other substances such as volatile irritants found in vehicle exhaust, tobacco products and tear gas, or components unleashed during chemotherapy treatment, or even environmental stimuli such as heat, cold. Why can TRPA1 relay so many signals of discomfort, while other TRP receptors deal with only one at a time?

Like all TRP receptors, TRPA1 is membrane-bound and most likely acts as a heterodimeric voltage-gated channel. TRPA1 has a particular secondary structure: its N-terminus is lined with a large number of ankyrin repeats which are believed to form a spring-like edifice. Most receptors have intricate pockets which are specific to a certain kind of ligand, and the slightest alteration of either the pocket or the ligand has drastic effects. Since TRPA1 can

respond to a variety of stimuli, it must have another system. Indeed, instead of presenting a pocket into which a ligand can lodge, the TRPA1 receptor forms covalently linked adducts with electrophilic compounds. The difference with other 'pocket-binding' TRP receptors is that TRPA1 ligand-binding persists for hours. The physiological response – i.e. pain in this instance – is greatly prolonged because the electrophile cannot readily dissociate from its receptor. Consequently, the receptor remains activated.

TRPA1 reacts to a variety of compounds, and does so in a variety of ways. It is directly stimulated by isothiocyanate and thiosulfinate compounds which give mustard oil and garlic their specific pungent qualities. This was discovered when mice, which didn't carry the receptor, turned out to be insensitive to both. Volatile irritants such as those found in vehicle exhaust, tear gas and tobacco smoke, as well as heat and cold stimulate TRPA1 indirectly. How? TRPA1 is part of a sensory pathway – or a number of sensory pathways – and is most likely activated or even modulated downstream of other neurotransmitter or growth-factor receptors. Funnily enough, although TRPA1 is not specific to only one component, it is surprisingly fine-tuned. As an example, the receptor is stimulated by acrolein – a compound in tear gas – yet it is insensitive to acrolein's corresponding saturated aldehyde: propanal.

Historically, mustard oil has been used extensively as a paradigm to study the mechanisms underlying inflammatory pain. Thanks to mustard oil, and a number of other substances, it has now been demonstrated that TRPA1 is capable of translating diverse signals of hostility into a singular sensation, i.e. pain. TRPA1 does seem to be on the crossroads of a number of sensory pathways. Consequently, it could prove to be an exceptional therapeutic target and may help to find ways of relieving those that suffer from chronic pain, or the secondary effects of chemotherapy and medication used in the treatment of arthritis for example. Likewise, engineering TRPA1 could help to counter the effects of noxious compounds in airways that induce not only inflammation but also chronic cough or asthma, which are a discomfort to many. Interestingly though, besides TRPA1's multiple talents, one laboratory has shown that sexual dimorphism can be an essential factor in the modulation of pain, and how it is sensed. Yet another difference between man and woman.

Cross-references to Swiss-Prot

Transient receptor potential cation channel, *Mus musculus* (Mouse) : Q8BLA8
Transient receptor potential cation channel, *Homo sapiens* (Human) : O75762

References

1. Bautista D.M., Jordt S.-E., Nikai T., Tsuruda P.R., Read A.J., Poblete J., Yamoah E.N., Basbaum A.I., Julius D.
TRPA1 mediates the inflammatory actions of environmental irritants and proalgesic agents
Cell 124:1269-1282(2006)
PMID: 16564016

2. Peterlin Z., Chesler A., Firestein S.
A painful Trp can be a bonding experience
Neuron 53:635-638(2007)
PMID: 17329204

3. McMahon S.B., Wood J.N.
Increasingly irritable and close to tears: TRPA1 in inflammatory pain
Cell 124:1123-1125(2006)
PMID: 16564004

National Nodes

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

Australia
RMC Gunn Building B19, University of Sydney,

Austria
Vienna Bio Center, University of Vienna

Belgium
BEN ULB Campus Plaine CP 257

Brazil
Embrapa Informatica Agropecuaria, UNICAMP-CP, CampinasT

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

China
Centre of Bioinformatics, Peking University, Beijing

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

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

Cuba
Centro de Ingeniería Genética y Biotecnolgía, La Habana

Finland
CSC, Espoo

France
INFOBIOGEN, Evry

Hungary
Agricultural Biotechnology Center, Godollo

India
Laboratory of Computational Biology & Bioinformatics facility, Centre for DNA Fingerprinting and Diagnostics (CDFD), Hyderabad

Israel
INN (Israeli National Node) Weizmann Institute of Science, Department of Biological Services, Biological Computing Unit, Rehovot

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

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

The Netherlands
Dept. of Genome Informatics, Wageningen UR

Norway
The Norwegian EMBnet Node, The Biotechnology Centre of Oslo

Pakistan
Department of Biosciences, COMSATS Institute of Information Technology, Chak Shahzaad Campus, Chak Shahzaad, Islamabad

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

Portugal
Instituto Gulbenkian de Ciencia, Unidade de Bioinformatica, Oeiras

Russia
Biocomputing Group, Belozersky Institute, Moscow

Slovakia
Institute of Molecular Biology, Slovak Academy of Science, Bratislava

South Africa
SANBI, University of the Western Cape, Bellville

Spain
EMBnet/CNB, Centro Nacional de Biotecnología, Madrid

Sweden
Uppsala Biomedical Centre, Computing Department, Uppsala, Sweden

Switzerland
Swiss Institute of Bioinformatics, Lausanne

Specialist Nodes

EBI
EBI Embl Outstation, Hinxton, Cambridge, UK

ETI
Amsterdam, The Netherlands

ICGEB
International Centre for Genetic Engineering and Biotechnology, Trieste, Italy

IHCP
Institute of Health and Consumer Protection, Ispra. Italy

ILRI/BECA
International Livestock Research Institute, Nairobi, Kenya

LION Bioscience
LION Bioscience AG, Heidelberg, Germany

MIPS
Muenchen, Germany

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

for more information visit our Web site

www.embnet.org

EMBnet.news
ISSN 1023-4144

Dear reader,

If you have any comments or suggestions regarding this newsletter we would be very glad to hear from you. If you have a tip you feel we can print then please let us know. Please send your contributions to one of the editors. You may also submit material by e-mail.

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

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

Submission deadline for the next issue:
August 20, 2007