EMBnet.news

Volume 13 Nr. 4
December 2007

- BioMacKit
- The EB-eye
- Protein interactions and drug targets
- A guide to EVALLER (2.0) web server

and more ...

Editorial

This is the last issue of 2007, closing our volume 13. The editorial board of EMBnet.news wishes to salute the many thousands of readers that keep downloading this publication very regularly, thanking them for their attention towards our activities. We have now created a specific discussion forum for readers to talk about the contents of the newsletter (see below). We hope to use the discussions to improve on several aspects that may be escaping our judgement. This number contains a report on the node formed in Brazil, two reports on courses (Athens and Madrid), technical articles on EB-eye (search engine), PairsDB, etc., and an opinion paper on standards in Bioinformatics. We hope that you enjoy the reading and, again, invite you to post your opinions in the forum for discussion, available on `http://www.embnet.org` under "Forums". We wish you a happy and productive 2008, and invite you to follow the celebrations of the 20th anniversary that will take place trough the year and will reach its highest moment with the EMBnet conference that will be held in September in Italy.

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

Cover picture: Green Vine Snake, (*Ahaetulla Nasuta*). Sri Lanka, 2007 [© Erik Bongcam-Rudloff]

Contents

Editorial 2
New presence in the EMBnet Brazilian Node 3
Course Report: Athens, Greece, October 2007 . 5
BioMacKit 6
IBS-ES-07 course: The making of 11
The EB-eye 18
PairsDB protein alignment database 22
Protein interactions and drug targets 25
A guide to EVALLER (2.0) web server 32
MRS version 3 38
Beyond 'The Curse of Babel' in bioinformatics . 40
Protein spotlight 88 42
Node information 44

Editorial Board:

New presence in the EMBnet Brazilian Node




Ana Tereza Ribeiro Vasconcelos, Goran Neshich, Wim Degrave, and Marcia L. Triunfol

National Laboratory of Scientific Computation, Petrópolis, Brazil.
Fiocruz Foundation, Rio de Janeiro, Brazil.
Brazilian Agricultural Research Corporation (EMBRAPA), Campinas. Brazil.

Less than ten years after completing the genome sequencing of *Xylella fastidiosa*, Brazil has now an established and consolidated scientific community devoted to Bioinformatics and Computational Biology that includes a National Genome Network of several well-equipped labs that are spread nationwide. The network concentrates efforts on sequencing ESTs and small genomes of organisms of economical and health interest. Currently, there are more than ten genome sequencing projects going on in the country, along with proteomics and computational modelling projects. Besides the biological data that has been gathered with such projects, the formation of large networks has offered a unique opportunity to train Brazilian scientists in bioinformatics and computational analysis of biological information.

As a country with an active and global participation in genome research and with significant contributions to the development and innovation of bioinformatics tools and software, the Brazilian presence in the EMBnet is now supported by a group of three institutions: the National Laboratory of Scientific Computation, (Laboratório Nacional de Computação Científica- LNCC at `www.lncc.gov.br`), EMBRAPA (Brazilian Agricultural Research Corporation at `http://www.embrapa.br/english`), and Fiocruz (Oswaldo Cruz Foundation at `http://www.fiocruz.br`). The three institutions are represented by Ana Tereza Ribeiro Vasconcelos, Goran Neshich, and Wim Degrave, respectively.

LNCC, the leading institution for scientific computation and computational modelling in the country, has been providing technological support of high performance to the nation's scientific community since its creation in 1980. In 2000, the Brazilian Ministry of Science and Technology created Labinfo (the Bioinformatics Branch of LNCC; see `www.labinfo.lncc.br`) to be the national reference center devoted to research and education in Bioinformatics and Computational Biology. Besides disseminating knowledge in these fields, Labinfo is also responsible for training human resources. As the node of the two major Brazilian networks BRGENE (Brazilian Genome Project) and PIGS (Southern Genome Project), Labinfo has been responsible for the submission, storage and management of nucleotide sequences provided by sequencing labs that participate in these two networks. While BRGENE includes 33 institutions involved with projects such as the genome sequencing of *Chromobacterium violaceum*, *Mycoplasma synoviae*, and *Anopheles darlingi*, among others; PIGS includes a group of labs located in the south of Brazil that have been working on sequencing the genomes of *Mycoplasma hyopneumoniae*, *Mycoplasma hypneumoniae* 7448, and *Mycoplasma hyopneumoniae* 7422.

Since its creation in 2000, Labinfo has developed mathematical and computational methodologies for DNA analysis; it has also built a number of databases for storage of biological information (COLCENTROSUL, CTpedia, MamMIBase, among others) and has created tools for the functional and structural annotation of coding and non-coding genomic regions (SABIA). Besides working with many groups in Brazil, Labinfo collaborates with international organizations to work on several projects. One such project is HAMAP, an initiative taken by Swiss-Prot in collaboration with Labinfo for the annotation of proteins involved with bacterial pathogenicity. Another international collaboration is the one with the Ludwig Institute for Cancer Research to build a database of cancer testicles antigens.

Another member of the new EMBnet Brazilian node is Fiocruz, which is one of the most traditional research institutions in the country. Mainly devoted to biomedical research, technological development and production of drugs and vaccines, Fiocruz has also a long-standing tradition in research and training in Bioinformatics. In-depth research and development of new tools have grown considerably in a variety of fields and comparative genome analysis, assembly and annotation is now the research focus of several groups (http://www.biowebdb.org). Functional genomics, evolution and phylogenomics research projects have emphasis on structural genome organization and its evolution as well as on studies of biochemical pathways and identification of new drug targets for infectious and parasitic diseases.

An additional ongoing line of research is the application of machine learning techniques, genetic algorithms and other statistical methods to solve biological problems. Software tools and new approaches for comparative genome analysis (Bioparser, GenoMycDB, LocalCOG and AenPi; see http://www.dbbm.fiocruz.br/labwim/bioinfoteam) as well as for small laboratory grid computation (Squid), have been developed. The laboratory also interacts with the World Community Grid to build an all-against-all rigorous protein comparison database for studies on annotation and evolution (http://www.dbbm.fiocruz.br/GenomeComparison). The Program for Scientific Computing (see http://www.procc.fiocruz.br) is dedicated to research and development of mathematical models and statistics applied to biological systems, especially those used in epidemiology and public health.

As for the Computational Biology Group (http://www.cbi.cnptia.embrapa.br) at EMBRAPA, projects focus on how proteins interact with their substrates, inhibitors, and other types of macromolecules. A diverse group of individuals, who bring expertise in different fields, work towards developing products and technologies for improving agriculture and human health. Currently, the group is studying proteins that are capable of unsaturating long chain fatty acids in microalgae and which in the future might be used to turn the bioengineering of omega3 production a more economical and sustainable process. Omega-3 is a fatty acid that improves memory and cognition capabilities and which is associated to brain development in infants.

The group also foresees the possibility of growing plants capable of producing specific polymers of desired characteristics such as spider silk-like protein aggregates. Another project undertaken at EMBRAPA is the study of proteins that protect plants against insect attacks at the molecular level. This might open the opportunity for developing plants capable of withstanding attacks of plague insects, which in turn may reduce the use of harmful chemical agents.

At Embrapa Structural Computational Biology, the major research focus is on the relationship between sequence, structure and function of proteins, coupled to the study of protein – protein and protein-substrate interactions. To predict protein interactions, the group developed STING (http://sms.cbi.cnptia.embrapa.br/SMS/STINGm/SMSReport/), which is a database of per-residue reported descriptors of protein sequences, structure, function, and stability (STING has had more than 15 million accesses since launched in 1998).

Supported by the EMBL-Heidelberg staff, and specially by Chris Sander, it was the first out-of-Europe group to host the EMBNet node, back in the early 90's. This event increased the group's international exposure and helped it's establishment as a collaborative liaison.

Together, LNCC, EMBRAPA, and Fiocruz hope to provide a valuable contribution to EMBnet by disseminating bioinformatics information through it's node site (http://www.br.embnet.org/), and ultimately by helping to advance our understanding on how life works and evolves.

Course Report:
Athens, Greece, October 2007

Laurent Falquet

Swiss Institute of Bioinformatics, Lausanne, Switzerlans

Successful collaboration of Swiss, Belgian & Greek EMBnet nodes to deliver an introductory course on sequence analysis in Athens, Greece.

The Bioinformatics Group of the Biomedical Research Foundation of the Academy of Athens (BRFAA, `http://www.bioacademy.gr/bioinformatics`) was appointed the Greek National Contact Point for Bioinformatics during the EMBnet 2007 annual general meeting in Malaga, Spain. The newly elected Greek EMBnet node with the collaboration of the Swiss (`www.ch.embnet.org`) and Belgian (`www.be.embnet.org`) nodes, successfully delivered an introductory course on sequence analysis in Athens, Greece.

The course, from the 1st-3rd of October, was partly funded by EMBnet (`www.embnet.org`) and partly by the registration fees.

Course schedule

Day 1
- Biological databases
- Sequence comparisons: basic methods, scoring matrices, dot plots, pairwise alignments

Day 2
- Sequence comparisons: searches in databases (BLAST, SSAHA, BLAT,…)
- Multiple alignments and patterns

Day 3
- PSI-BLAST, Profiles and HMMs (via Marratech from Switzerland)
- The wEMBOSS package (via Marratech from Belgium)

Course infrastructure

The course was held in the teleconference room of the BRFAA equipped with 10 computers. It combined theory and hands-on exercises.

State of the art e-learning technologies were applied: a Moodle quiz was designed for the first time for one exercise using the `edu.embnet.org` moodle installation and two lectures were given via the Marratech (`www.marratech.com`) video conferencing software.

The IT configuration was the following: two computers with the Marratech client were used during the lectures of the last day (see schedule above). One was localized in the teleconference room to perform "desktop sharing" on the 10 computers intended to the participants using the software "NetOp School" (`www.netop.com`). The second was localized in Switzerland and in Belgium respectively. This course was a good opportunity to test this new configuration that seemed more secure than a configuration in which the Marratech client would be installed on all the computers of the participants due to the bandwidth available at the BRFAA. The configuration proved partially successful since the mirroring software failed to display the video window of the teacher, thus only the slides were displayed. An evaluation from the participants showed that the students were happy with Moodle, however some of them found it difficult to follow the lectures via Marratech. This point should be improved.

Participants

To keep the quality of the course at high standards, the course was limited to only 19 participants. There were 2 participants per computer giving them the opportunity to interact during the exercises. The background of the students varied from biologists, chemists, computer scientists, electrical engineers to physicists. This heterogeneous background was an asset as lively discussions and interesting points were raised during and after the lectures.

Course evaluation

The overall evaluation gave 4.7 out of 5 points maximum. The comments could be summarized as follows:

- Most participants were very happy with the course. Although there was too much information in too short time for some of them, they all appreciated the exercise sessions.
- Most participants wanted to learn more and requested more advanced courses in the future.
- Many participants would appreciate if bioinformatics courses were given regularly.

Bioinformatics conference following the course

The Hellenic Bioinformatics & Medical Informatics conference took place immediately after the course. The program of the conference, attended by about 80 participants, is available online at:

`http://www.bioacademy.gr/bioinformatics/meetingOct07.htm`.

Dr Laurent Falquet, as invited speaker, opened the conference with a lecture entitled: "The Swiss Institute of Bioinformatics and the EMBnet organization".

The teleconference room of the BRFAA during the practicals.

BioMacKit:
a bioinformatics portable teaching kit

Álvaro Martínez Barrio
PhD Student

Erik Bongcam-Rudloff
Associate Professor

The Linnaeus Centre for Bioinformatics, SLU, Uppsala Sweden

Introduction

Our group is currently involved in different educational projects involving nations located in various parts of the globe – including Sri Lanka, Kenya and Chile [1] [2].

According to our experiences, to achieve successful results, courses should be taught using the latest methods and demonstrating a wide

Figure 1: Classroom at ILRI, Nairobi, Kenya.

range of bioinformatics resources. The contents of the course highlight the integrative discipline of bioinformatics, with test oriented hands-on tutorials after each theoretical section. Students then practice their skills and realize the power of what they just have learned. If courses are not interactive enough or without a practical tutorial after each module, students become lost and

fail to understand the ideas involved. The reasons that contribute to make this situation worse, vary from low internet speed connections to lack of convenient and organized resources.

Methods

To solve these problems we opted to engineer a portable bioinformatics teaching kit that would locally provide all the necessary resources, solving the speed related problems. The close locality of the connection to our server will avoid latency problems with the network.

This solution will be far more complete and powerful when compared to other solutions. It contains some of the most important biological databases and tools, which are not available on comparable systems, e.g., LiveCDs.

Our kit tries to address several other important issues. The kit is based on a collection of the most well known tools released under open source licenses, and therefore avoids paying the excessive prices related to software licensing. Furthermore, the operating system chosen, though proprietary, is based on a good mixture of user friendliness and a high-end Unix development environment for bioinformaticians.

Some of the software packages included in BioMacKit comprises:

• eBiotools [3] contains more than 150 useful bioinformatics tools, among them: EMBOSS suite, embassy, HMMER-2.1.1, Domainatrix, Phylip, NCBI Toolbox, Primer3, ClustalX, Infernal, sim4, smile, Vienna RNA Package, meme. A new release of ebiotools is planned to be shipped within the BioMacKit. The next version will update most of the existing applications whilst also adding some new goodies.

• MRS 3 [4] presented some initial configuration problems but once adapted with the `configuration.pl` installation script and installed Boost libraries from the Fink [5] project, a functional MRS site is in place.

Figure 2: UML schema of installed parts for MRS.

• wEMBOSS [6]

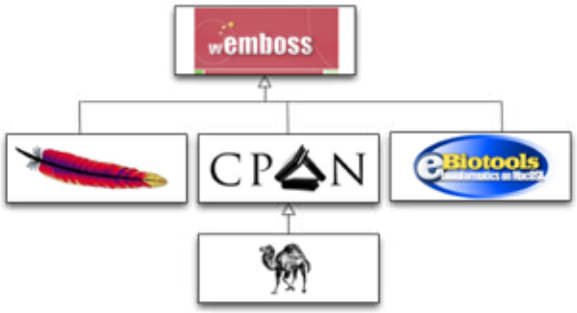


Figure 3: UML schema of installed parts for wEMBOSS.

• wrappers4EMBOSS [7] [8]. The new release of wrappers4EMBOSS 2.0 and its `mrsseqget.pl`, will make possible to retrieve data within wEMBOSS using the MRS system.

• EnsEMBL [9]. The biggest problem installing EnsEMBL is the large amount of hard disk space consumed by the downloaded databases, the amount of records contained in some of them

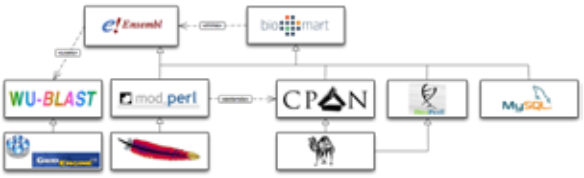


Figure 4: UML schema of installed parts for EnsEMBL.

(i.e: **compara**) and the space consumed by the sequence data and parsing step for BLAST.

• Web services development (servers and clients): gSoap and SOAP::Lite.

• Queuing system: Xgrid (proprietary) and SGEE (optionally installable). DRMAA interfaces accompanied.

The hardware consists of a mini computer and an external disk that only weighs 4.10 kilograms and measures 16.5 x 16.5 x 9.58 centimetres.

Figure 5: Dimensions of the teaching-kit when assembled.

	Mac mini [10]	LaCie min Hard Drive & Hub [11]	Total
Processor	1,83 to 2.0 GHz Intel Core 2 Duo		
Memory	1 GB 667 MHz DDR2 SDRAM (PC2-5300), support until 2 GB		
500 GB	580 to 620 GB	Hard-disk	580 to 120 GB
Interface	Serial ATA	USB 2.0 & FireWire plus multiple hub ports	
Weight	2.65 kg	1.450 kg	4.10 kg

Table 1: Specifications of the teaching-kit hardware parts.

When both components were assembled and the BioMacKit was switched on, the disk remained unmounted so we could never reach the data from the applications (if no user was logged in). A simple XML file (in Apple "jargon" called a plist) make it possible to have the disk mounted as long as the computer is running (even when all users are logged out).

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE plist PUBLIC "-//Apple Computer//
DTD PLIST 1.0//EN"
"http://www.apple.com/DTDs/PropertyList-
1.0.dtd">
<plist version="1.0">
<dict>
<key>AutomountDisksWithoutUserLogin</key>
<true/>
</dict>
</plist>
```

Listing 1: autodiskmount.plist.

To be able to connect to the X11 server from another hosts, X11 forwarding should be enabled by following the steps below [12]:

```
alvaro$ sed 's/#X11Forwarding\ no/
X11Forwarding\ yes/' /etc/sshd_config >
/tmp/sshd_config
```

```
alvaro$ mv /tmp/sshd_config /etc/.
```

Listing 2: Enabling X11 forwarding.

There is ongoing work to enable the rapid cloning of kits and we are working with new ways of easily updating the databases. For the cloning purpose, the ideal tool to use is Apple Software Restore. Apple Software Restore, or `asr`, is a command-line tool built-in to Mac OS X. It can create bootable clones of your volumes: `asr` essentially functions like `ditto` in it's file-copy mode, it also has the ability to clone a volume at the block level, preserving every last bit of data on the volume.

The "`-erase`" argument is optional, though recommended when cloning an operating system. To use `asr` in blockcopy mode, unmounting both the source and target is needed. In other words, your boot volume can't be block-copied.

`asr` can also accurately clone volumes without the use of an intermediate disk image. But, at the same time, `asr` efficiently copies disk images onto volumes, either directly or via a multicast network stream.

```
sudo asr -source / -target /Volumes/
Backup
-erase -noprompt
```

Listing 3: Cloning your boot disk into a Backup volume with `asr`.

Results

This kit enables several interactive approaches:

- **web-browser:** The most common and probably easiest way of accessing the BioMacKit applications.

- **ssh:** A useful way to access the BioMacKit both for administrative purposes as well as for students who may use their username to login and use command-line applications or X11 forwarding.

```
alvaro$ ssh -X -l user biomackit
/usr/X11R6/bin/xclock
```

Listing 4: Execute xclock into your X11 server using X forwarding.

• **GUI:** Proprietary programs like Apple Remote Desktop (ARD) [13] may be used to administer the BioMacKit easily within the Aqua environment. Our team can help with support issues if the ARD ports are open through the institutional firewall. Other applications can be run or scheduled remotely from other Mac OS X desktops as well.

Figure 6: xclock window on your own desktop.

By virtue of the double card (ethernet and wireless) installed in our BioMacKit, it is possible to provide a wireless secured connection for a pool of computers. The mini kit can also be used as a bridge. Therefore, laboratories and tutorials can be prepared seamlessly in rooms not properly equipped with sockets. Note that performance is not tested in this configuration (it logically will be reduced to some extent).

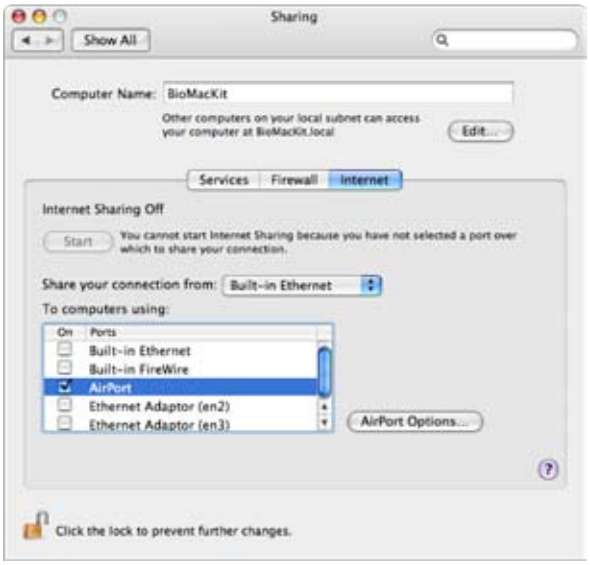


Figure 8: Activating the share network connection on Mac OS X.

The BioMacKit was presented and demonstrated in the past RIBIO/EMBnet collaborative workshop [14].

Conclusions

The BioMacKit described by this article is a mini-server that fits perfectly into small spaces due to its reduced dimensions. At the same time, it allows easy connectivity due to its multiple network cards. Even if the kit lacks screen and a keyboard,

Figure 7: ARD while controlling another computer.

Figure 9: BioMacKit creating a WLAN for two computers.

it can be administered remotely and controlled effectively. Furthermore, there are also protocols installed which allow both the system and available data to be rapidly updated. This will ease logistics and updates.

Therefore, our educational kit permits easy setup, transportability and a user friendly environment for groups aiming to teach bioinformatics in countries with an underdeveloped Internet infrastructure.

This "Bioinformatics portable teaching kit" will greatly facilitate the teaching of courses and the organization of workshops. The kit can also provide a small to medium sized laboratory with basic tools and necessary databases in a very affordable money-value return rate.

Furthermore, our experience of working with many different departments and their IT technicians makes us believe that our solution is highly configurable and compatible with a wide range of security policy rules.

References

[1] Course Report: Nairobi, Kenya, August 2006 Erik Bongcam-Rudloff and Etienne de Villiers EMBnet news Volume 13, Nr 1

[2] Course Reports: Valparaiso, Chile, December 2006 Nairobi, Kenya, March 2007 Erik Bongcam-Rudloff and Alvaro Martinez Barrio EMBnet news Volume 13, Nr 2

[3] eBiotools Sept. 1st 2005 (new release) Anders Nistér and Erik Bongcam-Rudloff EMBnet news Volume 11, Nr 3

[4] MRS: a fast and compact retrieval system for biological data M. L. Hekkelman and G. Vriend Nucleic Acids Res. 2005 July 1; 33(Web Server issue): W766–W769.

[5] Fink - Home http://finkproject.org/ [6] wEMBOSS: a web interface for EMBOSS Martín Sarachu, and Marc Colet Bioinformatics 2005 21(4):540-541

[7] wrappers4EMBOSS: a fast-and-easy way to integrate BLAST and other 3rd party software under EMBOSS Guy Bottu EMBnet news Volume 11, Nr 1

[8] MRS version 3, EMBOSS and wrappers4EMBOSS Guy Bottu EMBnet news Volume 13, Nr 4

[9] EnsEMBL 2007 T. J. P. Hubbard et al. Nucleic Acids Res. 2007, Vol. 35,(Database issue): D610-D617

[10] Apple - Mac mini - Technical specifications http://www.apple.com/macmini/specs.html

[11] La Cie - mini Hard Drive & Hub http://www.lacie.com/se/products/product.htm?pid=10462

[12] Technical Q&A QA1383: Enabling X11 Forwarding http://developer.apple.com/qa/qa2004/qa1383.html

[13] Apple - Remote Desktop 3 http://www.apple.com/remotedesktop/

[14] CBI07: Workshop on Collaborative Bioinformatics 2007 (RIBIO/EMBnet) Torremolinos, Málaga (Spain) 11th - 13th of June 2007

COPYRIGHT:

IBS-ES-07 course: The making of

José R. Valverde

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

This is a report on the recent course on "Introductory Biostatistics" at EMBnet/CNB. This course had some interesting novelties which are discussed and used as a way to introduce some basic concepts in the usage of the EMBnet education web site. General advice on creating and running a basic course (with a heavy traditional trend) is presented, leaving out advanced capabilities that are of reduced interest for the newcomer to using an e-learning facility.

The course

This has been an internal course organized by EMBnet/CNB for local users. Its aim was to serve as an introduction to Biostatistics with an outlook to professional usage. To achieve our goal we decided that we needed to meet several seemingly conflicting constrains: users had to be introduced to general Biostatistical analysis using a professional tool of general acceptance in advanced Bioinformatics, the tool should use a trivial GUI and should be available on standard operating systems (Unix [1], Linux [2], MacOS X [3] and MS-Windows [4]).

The course covered contents treated by many undergraduate Biostatistics courses (although due to time constrains we only dealed with a few key topics). This allowed us to concentrate on the practical approach: from this point of view it was intended to act more as a refresher of the methods and their application than of their mathematical foundation. In this way we intended to show attendants how they could perform non-trivial statistical analysis easily using a professional tool and induce them to carve for more advanced knowledge.

As a matter of fact, we have been surprised by the success of this approach: users soon got past basic analysis and started asking for more advanced features, accepting the command line surprisingly well. From this point of view, the course was a success. But getting to it proved to require us to pull some tricks from the hat.

Using R for teaching Introductory Biostatistics

We decided to use **R** [5] as it is most popular in advanced Bioinformatics analysis, but as R is a command line application driven by a rich programming language, it would soon become frightening for any lay user. There are many GUI interfaces for R; some of them like **RKward** [6], **Rgtk** [7] or **Rgnumeric** [8] have limited OS constrains, and so we decided to use **Rcmdr** [9, 10] which is available on all R supported OSes. By using Rcmdr users could get many seemingly complex tasks done easily and start using a few trivial written commands to lose fear against the full application.

Rcmdr provides an enhanced interface to R analysis: while the command line treats each simple step independently, Rcmdr agglomerates several conceptually related steps in one single menu selection. This makes it easy to teach the usual approaches to problem analysis as several related tasks can be done simultaneously, reflecting the usual approach taken in practice. As a result we could concentrate on teaching students the dynamics of normal analysis without needing to pay much attention to all the R commands involved.

Providing a professional environment

As we started testing the course we soon realized that sadly enough there are some differences among the support for statistical functions in Rcmdr on different systems, with Windows lagging behind and Linux providing the Gold Standard. In order to make the course run as smoothly as possible it was desirable to use Linux as the base platform, which is also the ideal evo-

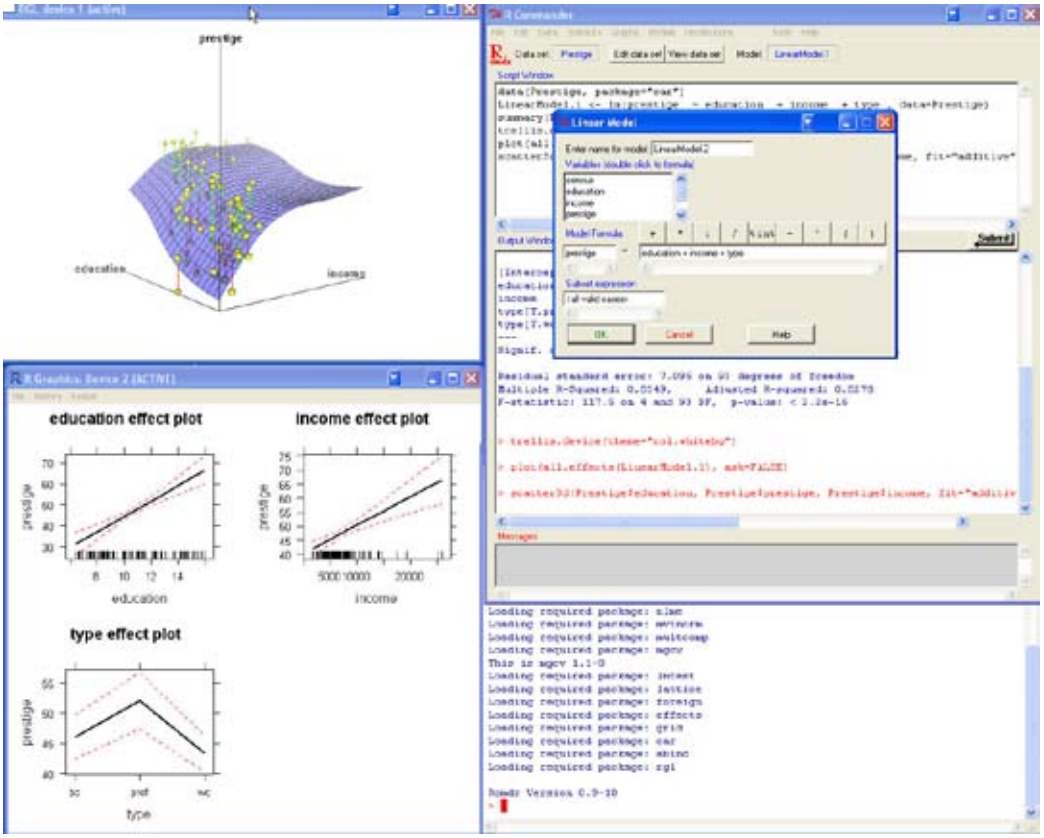


Figure 1: R commander.

lution for them if they later want to get deeper into Biostatistics. This left us with an unease feeling as students might not be able to fully reproduce the course at home (or the lab) if they used MS-Windows.

Our approach has been to create a virtual machine using **Qemu-Puppy-Linux** [11,12] enhanced to include R and Rcmdr. This way users can get acquainted with a full professional environment and not just R, and see that there is no reason to fear Linux any more than to fear R. We have developed two machines, one based on Qemu-Puppy-Linux 2.1.3 and the other on the more recent 2.1.7. They were deployed on Windows machines for the course using Qemu acceleration (virtualization) for efficiency. Our experience during the course was that the latter version (2.1.7) tended to run noticeably slower than the earlier one (possibly due to a more advanced desktop environment), hence although we provide both versions for public download we currently advise use of the 2.1.3 based machine.

Using the virtual machine proved to be surprisingly well accepted by students, who easily understood its basic principles. While we still advised all them to install the native version of R and Rcmdr on their own machines, it proved to be a useful tool for teaching. Plus, it provided a new environment "neutral" to both Windows and Mac based students. *These machines are publicly available for free download* [13]

Gathering course contents

We had originally intended to create the full contents for the course from scratch so we could have full control over them. For a number of reasons our workload has increased steadily and we finally decided to compromise and yield to third party auxiliary sources. To our delight we were able to find on the Net a wealth of information on Biostatistics that could be used for training; to the extent that our next problem became one of choosing which sources were most adequate to our needs (e.g. [14-18]).

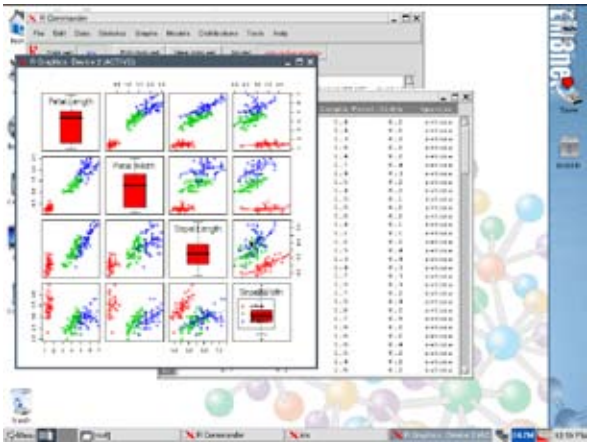


Figure 2: the Statistics virtual machine.

Obviously we decided to start off with resources available from within EMBnet: the **Swiss EMBnet node** [19] had preceded us on teaching this topics, although their approach differed significantly from ours: they devoted much more time to their course (we only had planned for 6 hours), where they stressed advanced applications (like microarray data analysis) we wanted to remain within basic applications and while they were using bare bones R we wanted to use an easy GUI tool. Still, some of their presentations were well aligned with our goals and we decided to rely on them [21].

But by far, the most useful resource that turned out on a web search was without doubt **Supercourse** [22]. Supercourse is "a global repository of public health and prevention" encompassing a network of over 45000 scientists in 174 countries and providing access to over 3000 lectures in 26 languages. The lectures in Supercourse cover a wide range of topics (Biostatistics among them), are free to share and use without restriction, and use a style described as "**hypertext comic books**" [23]. Striking as it may sound, the hypertext comic style has relevant advantages for training, like the inclusion of extended comments and hyperlinks to additional information associated as comments to presentation slides. While the basic format for Supercourse diffusion is as HTML based presentations, most of them are also available in MS-PowerPoint format, being thus easy to adapt, translate or modify. As a matter of fact we found Supercourse so useful that we decided to host a mirror site of it locally [24] as an additional service to our users and the community.

Designing the course for electronic access

As we were gathering the information we realized that we were relying heavily on electronic materials and tools. Rather than printing them and handing copies, printouts and presentation handouts to students we felt it was only natural to use an electronic site for it. We considered following the Swiss EMBnet node lead and offer a set of static web pages, but being overworked we wanted to automate other tasks (like assignment submission) as well to reduce our work. The obvious choice was to use an e-learning tool like the one installed at EMBnet education web site (Moodle[25]).

Creating the course is very simple, after logging in all that we needed was to go to the course listing and, at the bottom of the page press the "Add a new course" button and fill in the general information. In this case we decided to create one section for each hour/topic plus a wrap-up

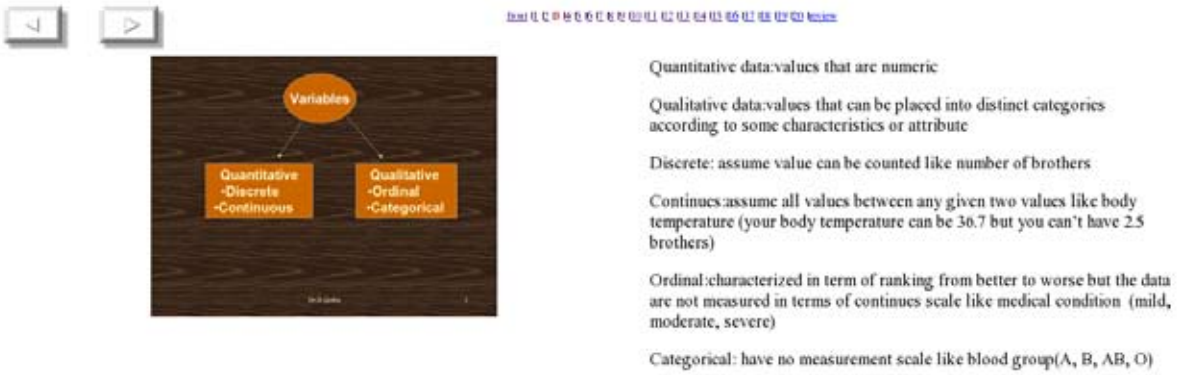


Figure 3: Supercourse example.

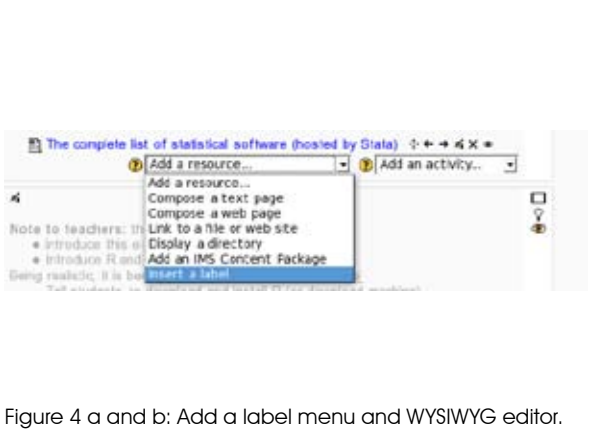


Figure 4 a and b: Add a label menu and WYSIWYG editor.

module for final advice and a terminal section for gathering student feedback. We also set an access password for students to enrol and disallowed guest access during the course period.

Our next step was to fill in a description for each section. While the immediate approach may seem to just edit the section headers, we preferred to create new labels independently. This is so because section titles are not carried over when a course contents are imported into a new course (e.g. to build an extended version) while all other labels are. This way we ensure that section titles and comments are preserved in the future.

Once in the course page, we switched our role to teacher and turned editing on. Adding section comments was trivial: just select "Insert a label" from the pull-down menu labelled "Add a resource" in each section. This opens a WYSIWYG editor were we could add the section title and comments.

Next came adding the contents. Although Moodle provides a rich environment for adding course materials and activities, we had decided to take it easy and stick with the traditional approach (i.e. Providing presentation files and external links). Once more, adding them was trivial: we used again the "Add a resource" pull-down menu and selected "Link to a file or web site".

It might seem that this menu would just provide links, hence requiring us to store files somewhere else, but actually it is more powerful than that: not only can you provide a title and comment for the link, and make a range of decision on how it will be displayed, but you can also use this form to *upload* files into the course contents. These files can be arranged in folders and be of any type. Thus, if you have an external link, you can just provide it, but if you want to include a file in your course, you simply upload it, organizing your files as you wish, and chose the uploaded file for linking. Moodle will insert the appropriate internal link to the file within the course for you.

Not only that: Moodle will realize what kind of file is linked (web page, PDF file, Word document, etc...) and set the correct icon for it automatically in the course page.

With just these two devices (inserting labels and links to files or web sites) we easily included all course materials in our course. We then went on to automate assignment submission for evaluating students. It was easy again: pulling down the "Add and activity..." menu we soon spotted an "Assignment" option which produces a form with an easy WYSISYG editor. Using this form it is easy to provide a description of the tasks to be performed by students, and what's even more useful -for us- we can also control the time they have to hand out their assignment, define what kind of assignment it is and how it will be evaluated.

Tuning the course for teachers

Moodle offers a wide choice of tools for teachers and their coordination, like forums, chat rooms, grading systems, etc... As this is a small course we decided to do without most of these help tools which we felt were not needed. At the same time we did not want to give up our original goal of building a better, more complete course for the

Figure 5: Link to a file or web site.

future. To achieve this we took advantage of a few trivial aids provided by the system.

First, we wanted to know what we had done well and what could be enhanced. For this we needed to run a poll on students and obtain their feedback (this is, by the way, customary practice in our courses). Using Moodle **Survey** activity solved our problem. Clicking on "Add an activity..." and then choosing "Survey" allows us to generate a full student satisfaction survey automatically. We did not even need to worry about the questions, as Moodle comes by default with a number of standard student satisfaction evaluation forms (ATTLS, COLLES) to collect user preferences, actual course performance or both. All we had to do was select the kind of survey (COLLES preferred and actual), give it a name and accept the default survey description.

Secondly, we wanted a way to provide additional **information to teachers** for driving the course. This entails descriptions on how to conduct the practical sessions or a general introduction to the course which may help other teachers willing to

Figure 6: Assignment.

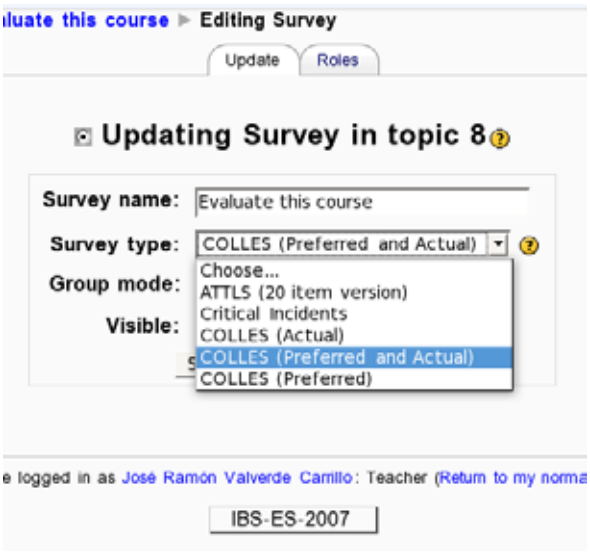


Figure 7: Survey.

use the course or even remind ourselves later on when we try to give it again. We could collect this advice and our experience in a number of ways, but for this time we went for the easiest one: we simply used a "label" (online text) to add these comments were appropriate on the course sections. This in turn brought up a new concern: this information is intended for teachers, not for students. Hiding it totally was undesirable, but so was leaving it open. Thankfully, Moodle allows us to easily hide materials from users' view while keeping them visible for teachers: all we needed to do was click on the "eye" or "view" icon attached to these labels to gray them out for teachers and hide them for students.

Planning for the best (worst)

Finally, we were optimistic that our work might be used by others to impart their own courses. Eager to facilitate this we decided to make a backup of the course so others could download all the contents easily and replicate it at their sites.

Building a backup was easy as cake: just click on the "backup" link in the course "Administration" box. It allows one to backup course contents with a large degree of granularity. Our main concern here is to make sure that everything worth saving is included (which is easy as everything is selected by default) but *nothing* private is exported (so user details do not get out of the server). Once more the task proved trivial as all we had to do

was to select to export *None* of the *User data*, and make sure (at the bottom of the form) that student details, files and logs were not included.

Publishing the backup was a bit more tricky. Backup files are stored in a folder within the course named "backupdata". In principle one can just add a "Link to a file or web site" and select this newly generated backup file to publish it (the folder and the backup file appear on the list of course files brought up by the "Link to a file or web site" form).

One can actually do it and get a link to the backup published with the course, but it will not work. The reason is that the "backupdata" folder has special permissions and does not allow its contents to be accessed by non-teachers. That may be OK for internal use, but not if you want to make the backup available to others. The trick, though, is trivial: if a file within "backupdata" is not readable, all that is needed is to *move the backup outside its folder*. To do this, open the file browser and click on the checkbox by the backup file. Then, on the pull down menu "*With chosen files...*" select "*Move to another folder*". Finally move up to the *Parent Folder* and click on the button labeled "*Move files to here*". The backup file is now outside the "backupdata" folder and fully accesible to users. You can now Choose it and make a link to the file from the course. Now everybody will be able to download your backup and replicate the course at their site.

That was an optimistic scenario (assuming our course is good enough for others to use), but there is a pessimistic one as well and that is the main reason for backups: what if our server crashes? And indeed, the disk array holding our local (CNB) server did crash on the morning of the course day, yet we did not have much to worry. What happened was that we had been cautious enough to save a copy of the backup on our machines and -this was the most beautiful and satisfying thing- we had replicated the whole course contents at EMBnet education site. When our server crashed, we had nothing to worry as students could be redirected to the replica at EMBnet main server seamlessly. Truly, their local user names were not recognized by the EMBnet main site, but they could register there freely and get the contents. That saved the day for us until we finished recovering the disk array for next day.

Finally, we added some copyright information and announced the availability of the course for others to use on the Moodle Exchange[26] and EMBnet sites.

Discussion

We have introduced the basic methodology to use the EMBnet education web site to support a traditional course, showing that it is easy to build a useful course site collating all the materials and to add some extra bonus with very little effort.

Compared with the traditional approach of using static web pages, using Moodle is not comparatively more difficult, and although the richness of the environment may seem discouraging at first, it is actually trivial to use if one limits oneself to use the most basic tools to reproduce traditional approaches. Hence converting an existing course to Moodle is indeed trivial and requires nothing different from what has traditionally been done. Actually, since the tools included provide an integrated, WYSIWYG environment, one can easily argue that it is even easier to add content to a Moodle course than it is to add contents to a traditional static web page.

We also show that it is not needed to make full use of Moodle capabilities to run a successful course, and that simply limiting use to basic inclusion of presentation files and web page links may be good enough (obviously, since that is the tune of most current courses). In addition one may enhance a course and its management with a few mouse clicks to add student satisfaction surveys or include teacher advice hidden from students.

Although we have not mentioned it, it should also be obvious that the same hiding mechanism can be used to hide temporarily from students resources (such as assignments) until needed, at which time they can be made visible as if they appeared fully anew out of nothing.

Moodle also provides a default news forum which can be used to notify users of relevant news, as well as an integrated calendar that will notify

them when assignments are due, thus releasing extra work from teachers, and a discussion forum for students to comment on the course. These are a bonus added without need of teacher actions.

The backup facility has proven useful both as a system recovery mechanism and also as a course dissemination method. The ability to replicate a course in other sites is invaluable not just for sharing and dissemination purposes but also as a way to have a backup server in case of disaster.

Connected with the latest we have shown that using existing resources can save a lot of work when preparing a new course. We have introduced Supercourse, an ambitious initiative for Public Health training, and demonstrated its utility and, by extension, the utility of public course repositories such as the one recently setup by EMBnet [27] or Moodle Exchange. We can hardly stress the utility of these initiatives to save trainers work and to deliver quality materials for training and wholeheartedly encourage everybody to join and support them. There are certainly many other public sources of course materials (witness MIT OCW site [28]) but most of them are closed to its hosting institution; the availability of public spaces for sharing and exchange training materials is paramount to reduce training costs and promote e-learning development.

Finally, we have developed a reference -simple yet powerful- platform for teaching Statistics using a lightweight virtual machine, and have verified through our experience its good acceptance among students and its utility in the classroom.

All the tools, materials and contents used are publicly available for use at EMBnet (subject to the corresponding copyright and licensing restrictions by their respective authors).

We want to thank EMBnet [29] for making publicly available its education web site [30], and the EU for its support to projects EGEE (INFSO-RI-031688) and EMBRACE (LHSG-CT-2004-512092) which have allowed us to free internal resources to do this work.

References

1. UNIX is a registered trademark of the Open Group
2. Linux is a registered trademark of Linus Torvalds
3. Mac OS X is a registered trademark of Apple Inc.
4. Windows is a registered trademark of Microsoft Corporation in the US and other countries
5. http://www.r-project.org/
6. http://rkward.sourceforge.net/
7. http://www.omegahat.org/RGtk/index.html
8. http://www.omegahat.org/RGnumeric/
9. John Fox (2005) The R commander: a basic-statistics graphical user interface to R. Journal of Statistical Software, 2005, vol 14, no. 9
10. http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/
11. http://www.erikveen.dds.nl/qemupuppy/
12. José R. Valverde (2007) Taking education beyond the classroom. EMBnet.news (2007) vol 13, no 3. pg 13.
13. ftp://ftp.es.embnet.org/pub/sic/machines/
14. Introduction to biostatistics. Ann Emerg Med 1990
15. Basic statistics for clinicians. CMAJ 1995
16. Statistics for Clinicians. Mayo Clin Proc 1981
17. http://www.statsoft.com/textbook/stathome.html
18. http://www.stata.com/links/stat _ software.html
19. http://www.ch.embnet.org/
20. http://www.ch.embnet.org/pages/courses2.html
21. http://www.isrec.isb-sib.ch/~darlene/EMBnet/
22. http://www.pitt.edu/~super1/
23. Global Health Network (1998) The reincarnation of biomedical journals as hypertext comic books. Nature medicine (1998), vol 4
24. http://www.es.embnet.org/Doc/Supercourse/
25. http://moodle.org/
26. http://moodle.org/course/category.php?id=1
27. http://elearning.embnet.org/course/view.php?id=40
28. http://web.mit.edu/ocw/
29. http://www.embnet.org/
30. http://edu.embnet.org/

The EB-eye

Mickael Goujon, Franck Valentin, Teresa Miyar, Hamish McWilliam and Rodrigo Lopez

EMBL Outstation
European Bioinformatics Institute

http://www.ebi.ac.uk/support/

INTRODUCTION

The biological sciences are generating data faster than it can be interpreted. Reaping the benefits in health care, agriculture and industry needs to be done in a fast, precise and consistent manner. The EMBL-EBI is one of the largest providers of data for the biological sciences and thus faces many challenges that need to be resolved by careful combination of its IT and scientific resources. One of these challenges is to provide fast, efficient and reliable searches of all the EBI data resources. This includes ArrayExpress, a repository and database for microarray data; ENSEMBL, that maintains automatic annotation on selected eukaryotic genomes; EMBL-Bank, Europe's first and largest nucleotide sequence archive; UniProt, the Universal Protein Sequence knowledgebase; MSD/PDB, The European arm of the Protein Structures database, and many more that can be accessed and explored from the EBI web pages starting from http://www.ebi.ac.uk/.

At the end of 2006, the EMBL-EBI unveiled a new search engine capable of performing fast searches across all its databases in a very simple and semantically uniform manner. This development is known as the 'EB-eye'. In this short piece we describe in some detail what lies 'under the bonnet'.

EB-EYE FRONT-END

EB-eye development started in 2006 with the main objective of improving the global EMBL-EBI user

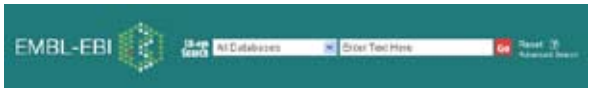


Figure 1. Simple search box is available at the top of every page at the EMBL-EBI web site.

experience. Based on the Lucene open-source search engine (http://lucene.apache.org/), the EB-eye user interface offers a simple and efficient way to search more than 170 million entries, from more than 24 databases. It is designed to cater for all types of users, from novices to experienced researchers, regardless of if they are already familiar with the EMBL-EBI databases.

To use the EB-eye, a simple search box is available at the top of every page at the EMBL-EBI web site (see Figure 1). The user simply enters search terms in the box. Any term can be used, be it database identifiers, author names, gene names, descriptions, titles, or a set of common words. The EB-eye will find and display all the entries in the system that match these search terms.

Following the principle of searching wide and then narrowing down the search, the results are first presented in a summary page with the different databases organised into several categories, each one representing a knowledge domain currently maintained and supported at EMBL-EBI. For each category and database only the number of matching entries is displayed. This gives an overview of the results covered by all the data resources. The user then has the choice of refining the search by adding new terms to the query or to select a category (or a database) to view the corresponding entry results. The latter will lead to a new page similar to the result pages which can be found in most common web search engines (e.g. Google, Yahoo, Live/MSN, etc.). The results are displayed as a list of entries with short descriptions and links to more information; typically the database's main web site (see Figure 2).

A user can refine their search at any time by adding new terms to their query. Additionally, EB-eye also uses an open source search clustering engine called Carrot2 (http://www.carrot2.org) to extract additional information from the result and display potential query refinement terms (see Figure 3).

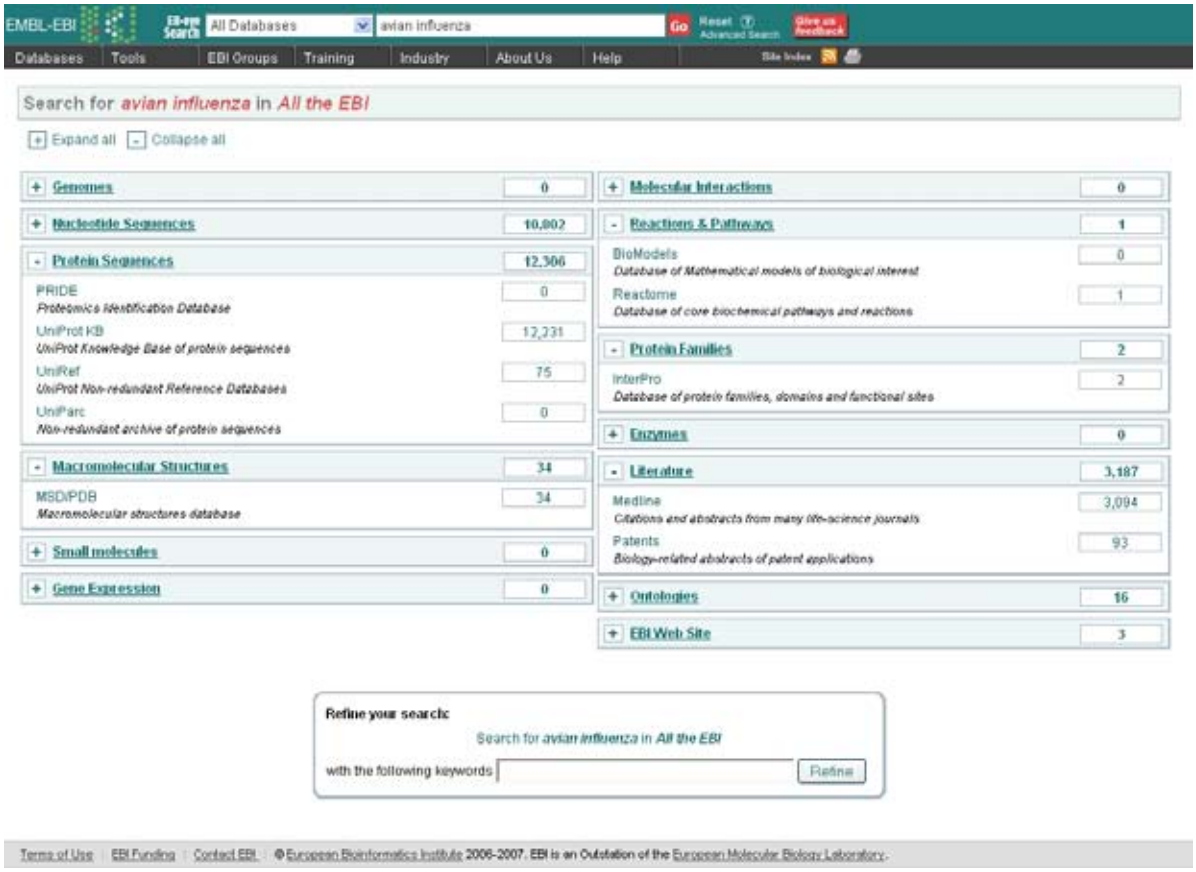


Figure 2. The results are displayed as a list of entries with short descriptions and links to more information.

Another important feature is the possibility to navigate within the network of database cross-references. EBI databases maintain a list of database cross-references for each entry containing identifiers of entries in other databases. These cross-references make it possible to inter-connect databases. EB-eye uses these cross-references to build a network from the databases indexed in the system and makes it available to the user and these are displayed for each entry on the result page (see Figure 3).

Searching data often requires greater specificity in queries and using boolean operators (i.e. AND, OR and NOT) and restrictions to particular fields of a database provide this granularity. An advanced search is available with a basic query form that uses natural language instead of boolean operators and a simple two step wizard-like process that allows the user to select the databases and fields to search. As a result, the advanced search is easy to use, while exposing the powerful search capabilities of the search engine.

Details of how to use the EB-eye can be found in the dedicated help pages at: `http://www.ebi.ac.uk/inc/help/search_help.html`.

WEB SERVICES

The modular architecture of the EB-eye provides flexibility in the addition and support of new features. For example, one important feature added recently is a SOAP-based Web Services interface. Web Services enable developers to use the functionalities of remote services from inside their own applications without the need for a local installation or being restricted to a particular programming language. Being relatively easy to use, Web Services technologies have become popular, notably in bioinformatics.

The EB-eye Web Services API consists of a set of simple methods covering most of the functional-

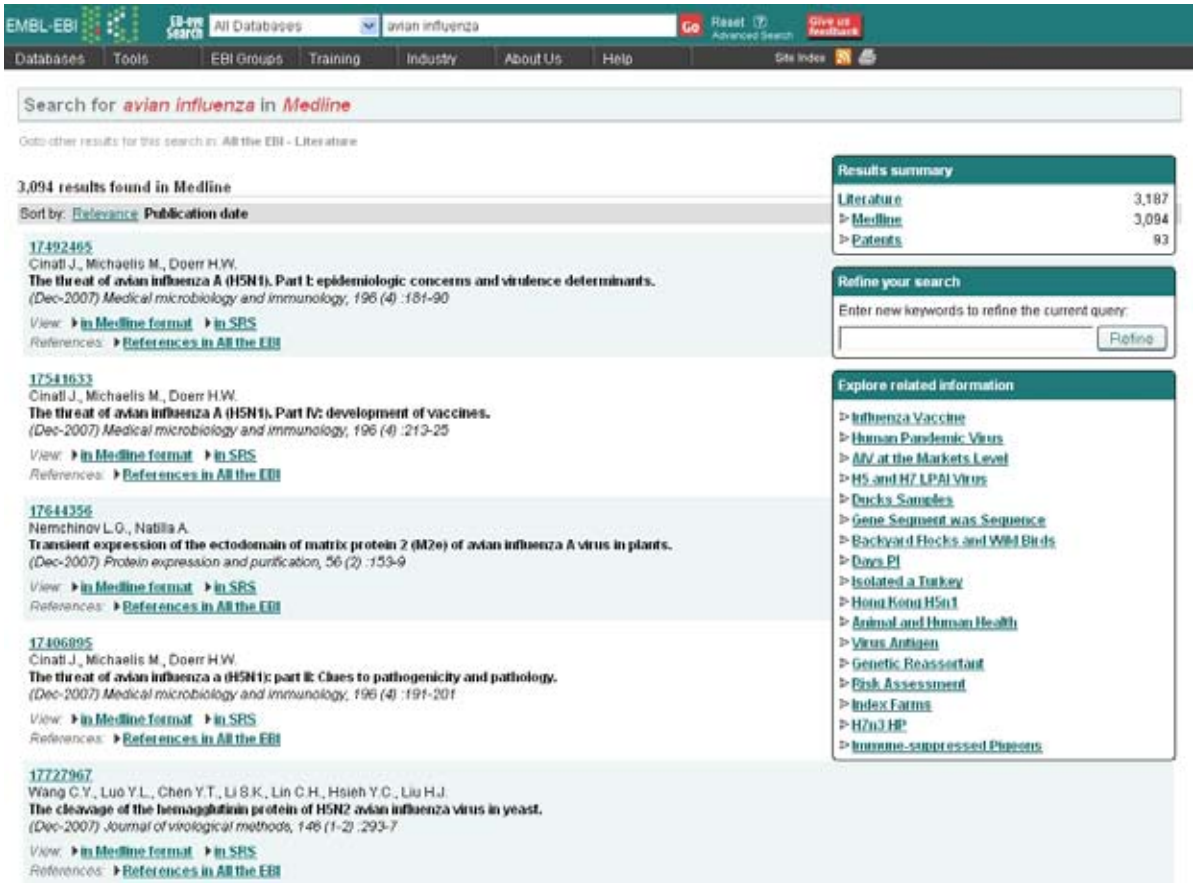


Figure 3. EB-eye also uses an open source search clustering engine called Carrot2 (`http://www.carrot2.org`) to extract additional information from the result and display potential query refinement terms.

ity described available in the web interface. A developer can query a database indexed in the EB-eye, retrieve the results and navigate the cross-references network with a few calls. Although the web service API is basic, it is possible to reproduce the functionality available on the web interface by combining several methods calls.

One aspect of the Web Services implementation that is particularly interesting is the possibility of daisy-chaining several calls together to create a workflow. In this context, several functionalities of the EB-eye, like the text-based search or the cross-reference network navigation, have already proven to be useful for database annotators and cross-reference maintainers.

To start using the EB-eye Web Services, the WSDL (Web Services Description Language) descriptor can be found at: `http://www.ebi.ac.uk/ebisearch/service.ebi?wsdl`

INDEXING DATA FOR THE EB-EYE

As outlined earlier, the main aim of the EB-eye is to provide fast and efficient searches of all the major data resources hosted at the EBI. At present there a more than 170 million documents, or entries, that need to be indexed and kept up-to-date. The indexing engine available in the Lucene distribution takes care of this task. However it does not provide tools that deal with very large datasets or take into account release cycles. Since Lucene is a free text indexing engine and it does not, by default, take into account the structure of a database. End-users require increased granularity when searching the EBI databases. In the EB-eye development, specifications that take into consideration data availability and triggers to copy and index new data, parallelised indexing and granular parsing have been factored into the indexing application:

Configuration

The application is driven by a set of configuration files described in the configuration module. Each one describes all the information needed about a domain: where to find the data, how to index it, how to search it and how to display the results. This module is an essential part of the system, enabling new databases to be added easily or existing ones modified without requiring redeployment of the application.

Data update

A Groovy script (`http://groovy.codehaus.org`) is created for each database to check for updates and, if needed, to copy and uncompress the data. The use of independent scripts enables us to easily address the difficulty of dealing with data from different locations and running specific post processing. Each script is responsible of creating a metadata signature (e.g. version number, release date, number of entries) for further Quality Control. In the same manner as the configuration, modifying a script does not require redeployment of the application.

Data splitting

The indexing is done concurrently on several machines. To equally share this task and to take into account the heterogeneity of the file sizes, the source data are split beforehand into chunks (sets of entries). New chunks are pulled by each indexing machine as soon as possible avoiding any idle time and thus making the most of each machine's capabilities. The information needed to split the files and the sizes of the chunks are defined by the configuration files.

Parsing and indexing

The data files need to be parsed to extract the information to index. How the fields are indexed (dates, cross-references, authors, etc.) is defined in the configuration files. In order to avoid coding cumbersome parsers and to offer a clear view of the data's format two open source parser systems are used: ANTLR (`http:// www.antlr.org`), for flat files; and ANTXR (`http://javadude.com/tools/antxr/index.html`) for XML files. Both parser systems use grammars to describe the format of the data. Actions are attached to these grammars that define which information is extracted for each field to be placed in the index.

When all the chunks for a database are indexed, the partial indexes are merged into one large index and basic verification of the data is performed automatically, based on the metadata signature.

Deploying

In the last step the completed indexes are copied to a new location and the web application is notified of the update. This triggers index redeployment on the servers and the up-to-date indexes become searchable by the user.

Automatic updating

Each one of the previous steps can be run manually but an automatic update mechanism has been developed that checks for data updates on a daily basis and, if necessary, retrieves and re-indexes the new data. A report is generated after each run detailing which databases where indexed, what errors occurred and the duration of the indexing. If a failure occurs, the administrator can resolve the issue and re-launch the update process at the indicated stage.

CONCLUSIONS AND FUTURE WORK

The new search engine provides a single point of access to all data resources available at the EMBL-EBI. The system has been designed to scale and is easy to extend and to maintain. At present, the development effort has been directed toward making it easy to access data from distinct and semantically different resources and empowering users with scientific knowledge without additional specialised training. In the future, further functionality will be added to the system that will allow the user to launch analytical tools such as BLAST, InterProScan and text-mining tools, such as Whatizit, directly from the EB-eye, thus providing greater exploration and exploitation capabilities to the end-user.

PairsDB protein alignment database

Kimmo Mattila

CSC - Scientific Computing Ltd., Finland

Introduction

Sequence similarity searching with the BLAST algorithm is a cornerstone of molecular biology.

The new PairsDB service provides access to pre-calculated BLAST and PSI-BLAST based alignments for a comprehensive set of protein sequences. The service allows you to explore protein sequences and their similarity relationships quickly and easily. The web interface for the PairsDB-service can be found in:

`http://pairsdb.csc.fi`

Structure of the PairsDB database

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

Next a sequence set containing less than 90% identical sequences, the NRDB90 set, is created. This pruning step is done with CD-HIT program. CD-HIT sorts all sequences by their lengths in decreasing order. Starting from the longest sequence, the procedure removes all sequences from the set that align over their full length and are more than 90% identical to the selected sequence. The procedure then takes the second longest sequence and does the same. The procedure continues, until all sequences have been processed. Because of the high similarity threshold most alignments need not be calculated explicitly, but instead a fast tuple lookup algorithm is sufficient. As a result the NRDB100 sequences are clustered into sequence families that contain a long representative sequence and group of shorter family members that are more than 90 % identical compared to the representative sequence. These representative sequences form the NRDB90 sequence set

For the NRDB90 set a BLASTP analysis is run in an all-against-all fashion. The results from this massive BLAST analysis step are stored into a relational database. Using these BLAST results non-redundant databases are created also for 80%, 70%, 60%, 50%, 40%, and 30% sequence identity. As a final step an all-against-all PSI-BLAST analysis is run using the NRDB40 sequence set.

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

PairsDB WWW interface

Finding name for your sequence

PairsDB interface is operated using the UniProt, PDB or ENSEMBL sequence names like CYC_HUMAN or 1J3S-A (this refers to the A-chain of PDB entry 1J3S). If you do not know the name of your sequence you can use the "*Sequence Space Filter*" to check it. Sequence space filter is found in the top bar of the PairsDB interface. With this search tool you can try to find the sequence name by searching the sequence descriptions finding sequences that match 100% to your query sequence or a fragment of it. Often already a fragment of 10-20 amino acids is enough to identify your sequence. If the sequence is not found, the reason may be that it was not yet in

Figure 1. The BLAST query interface of PairsDB service.

the public databases when the last PairsDB data set was collected.

Sequence Space Filter can also be used to collect sequence data sets using combination of several search criteria. For example you could easily collect all sequences that are from an organism and contain a given InterPro domain.

BLAST and PSI-BLAST based searches

PairsDB provides two ways to look for similar sequences for your query sequence. BLAST in nrdb90 level and PSI-BLAST in NRDB40 level. Both of them use the same logic to construct the sequence relationships from the database. Here we discuss only about the BLAST search interface but the same features exist also in the PSI-BLAST interface.

The BLAST search interface can be opened from the BLAST link in the top bar of the interface. There are two ways to do the search. The more simple "*BLAST results in NRDB90*" collects the NRDB90 level BLAST results for the query sequence or its NRDB90 level representative. Remember that you should feed the name of the sequence to the "*Query sequence*" filed, not the actual query sequence.

The second search option, "*BLAST results expanded to NRDB100*" retrieves also the family members of the NRDB90 level hit sequences. This search checks first the NRDB90 representative sequence for the given query sequence. BLAST hits for the representative sequence are then collected at the NRDB90 level. After this the hit list is expanded to NRDB100 level so that also those sequence neighborhood members that have overlapping match region with the query sequence are selected. The list of hit sequences can be filtered using following features:

- e-value (can vary between 1 - 0)
- fragments, hypothetical or transmembrane proteins
- source database (UniProt, PDB, RefSeq or ENSEMBL) or certain NRDB hierarchy level
- domains from InterPro, SCOP, CATH or ADDA domain databases. For InterPro and ADDA standard database identifiers are used. For SCOP and CATH domains PairsDB uses coding system, that can be checked from help pages of PairsDB
- any taxonomy ID number

The data retrieval can be started with the "Search" button. Typically retrieving and filtering the data takes 5 -15s.

BLAST Results

The BLAST results page starts with information about the query sequence and the corresponding representative sequence in NRDB90 level. Detailed information about the query or representative sequence can be obtained using the links in Shortcuts column. Note that the actual BLAST results are computed using the NRDB90 representative sequence, not the original query. Except in those cases, where the query is also the NRDB90 representative.

Match Overview

The Match Overview table lists the found BLAST/PSI-BLAST hits. The first column displays the location of the matching region between the hit and representative sequence. The original query sequence id is represented by a red bar and its NRDB90 representative sequence as a green bar. The matching sequences that originate from NRDB90 are shown as dark yellow bars while the corresponding NRDB100 level family members are presented as light yellow bars. Using the shortcuts (I,B,P) you can directly go to the sequence

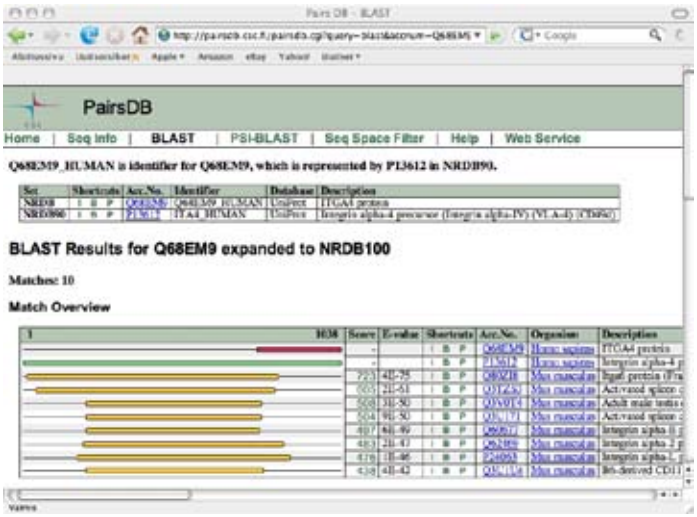


Figure 2. BLAST result page of PairsDB service

info, BLAST or PSI-BLAST page of any of these sequences.

Note also that one hit in NRDB100 level can represent several entries in the source databases. Thus if the result list seems to lack a UniProt entry or PDB structure that should be there, it may be presented by some other sequence name. E. g. UniProt entries CYC _ GORGO, CYC _ HUMAN and the A chain of PDB entry 1J3S have identical sequences so they are presented by only one hit, in this case named as 1J3S-A.

Stacked Multiple Alignment

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

Pairwise Alignments

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

Using the section options in the BLAST query page you can also choose to show the stacked multiple sequence alignment or hit sequence list in FASTA format.

Benefits of PairsDB

PairsDB is very useful tool when you need to do several slightly modified BLAST searches. E. g., when you want to get familiar with a protein sequence with which you have not yet worked before you right want to quickly see if there is some structural data available for the query sequence or are there known homologues in certain taxonomic group? All this you could of course do with normal BLAST too, however this would be much slower. One normal BLAST search may take several minutes and more specific queries may require laborious filtering of BLAST results or construction of users own BLAST databases. PairsDB produces essentially the same results in few seconds.

If the WWW interface of PairsDB does not suite all your needs, you can utilize the Web Service interface of the service or install the PairsDB to your local MySQL server. The files needed for building a local installation of PairsDB can be found at `ftp://ftp.funet.fi/pub/sci/molbio/pairsdb`.

Acknowledgment

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

Protein interactions and drug targets



Charalampos Moschopoulos*, Athina Theodosiou* and Sophia Kossida

Bioinformatics & Medical Informatics Team, Biomedical Research Foundation of the Academy of Athens, Soranou Efesiou 4, Athens, Greece
www.bioacademy.gr/bioinformatics
{cmoschop, atheodosiou, skossida}@bioacademy.gr
*The authors contributed equally to the work.

ABSTRACT

Designing a new drug is not a trivial project and very often it demands large amounts of money and time. The proteins and their interactions play important role in this scientific sector. The main approach of drug designing is to create new molecules which could affect a biological mechanism with the minimum side effects possible. In this review, we present the different kinds of protein interactions and their main depositories which are the online protein databases. Moreover, the main computational techniques that are being used in designing drug targets are presented. The combination of computational approaches with the biological ones demonstrates the important role that bioinformatics is playing in drug designing and targeting.

Introduction

Bioinformatics approaches have key role in the analysis of genomic, transcriptomic and proteomic data, in order to understand molecular mechanisms involved in diseases and in order to identify new targets for drug discovery. Nowadays, we are able to handle vast amount of data and reach valuable conclusions concerning disease mechanisms with the help of bioinformatics.

Proteomic data and more specifically those of protein interactions, provide insight in how diseases are developed. However, the nature of both the proteins and their interactions varies depending on the specific environmental and developmental conditions as well as the cell type. Therefore, it is quite challenging to obtain reliable protein interactions data.

Bioinformatics offers a great asset to the reduction of the cost of the drug developing projects. Various computational techniques can be applied in order to reduce the number of the laboratory experiments or the time that is needed to determine the most appropriate drug target.

In this report, we give a small review on the role that protein interactions play on drug designing projects and on how a number of various algorithmic strategies can help the drug developers to their research. Furthermore, we present a small introduction on the protein interactions and we refer to the online protein databases that are hosting many different kinds of information about proteins such as their structures, their interactions and their functions.

The remaining of this report is organized as follows: the first section reviews the nature of protein interactions as well as the different kinds of protein interactions. The second section focuses on the human protein interactome, while the third one presents the most important on-line protein databases. Three databases: HRPD, OMIM and Protein Data Bank are presented in more detail thanks to their importance and their common usage. In the forth section two examples, where the drug designing approaches are based on specific protein interactions, are provided. Finally, in the last section, we give a small overview of the computational techniques that are being used in designing drug targets. The overview focuses on the different nature of the algorithmic techniques, proving that drug designing is a research area that bioinformatics can be of great assistance.

Protein Interactions

The research on protein interactions has been very important in order to understand how proteins function within the cell, where proteins interact with other proteins, metabolites and nucleic acids. More specifically, protein interactions are crucial for forming structural complexes, for extra-cellular signalling, for intra-cellular signalling, for cell communication and for several other aspects of cellular function.

The characterization of protein interactions is really important to understand the molecular mechanism of biological pathways and disease processes. Complete knowledge of these pathways will help us to understand how diseases, such as cancer, are developed. Since almost all processes are regulated by multiple complexes, the absence of some interactions or the complete absence of physical interactions can be the cause of disease in humans [1].

The following section will focus on the general information about protein interactions. A brief description of physical protein interactions is presented and it is divided into two major groups: protein-protein interactions and protein-DNA interactions.

Protein - protein interactions

Most proteins live and function in very complex environments and have many potential binding partners. Some proteins are very selective on their binding partner, while other proteins are more "open-minded" and can interact with different kind of proteins making the binding more competitive. This, so called, multi-specific binding between two protein families is very common in regulatory pathways and networks [2]. There is an important distinction between the types of protein-protein interactions. They can be classified according to the proteins involved in the interactions, (structural or functional) or they can be classified based on their physical properties. From the structural point of view, protein-protein interactions can occur between identical or non identical chains (homo or hetero-oligomers). In addition, depending on the stability and mechanism of the formation of a protein-protein complex, they can be subdivided into non-obligated (short living) complexes and obligated (stable) complexes. Furthermore, they can also be divided into transient and permanent, based on the lifetime of the complex. Last but not least, protein-protein interactions can be classified based on their functional role. Common functional classes are the enzyme-inhibitor complexes, antibody-protein complexes and protein-receptor complexes.

Protein - DNA interactions

Protein – nucleic acid interactions play important role in various important cellular processes such as transcriptional regulation, recombination, genome rearrangement, replication, repair and DNA modification. A classification of protein-DNA complexes was attempted by various authors [3].

The process of transcription is mediated by a number of protein–protein and DNA-protein complexes. The protein factors modulating gene transcription are the transcriptional regulators which bind to specific DNA sequences named promoter sequences. Several transcription factor - DNA interactions have revealed new insight into the molecular basis of cancer and other human diseases. Genome-wide protein-DNA interactions may be measured using chromatin immunoprecipitation (ChIP) in conjunction with expression from microarray data. In contrast with protein-protein interactions, protein–DNA interactions are not obligate, as both the proteins and the DNA exist in isolation [4].

Human protein interactome

When proteins interact with other proteins they form complexes. Finally these complexes can be part of an extensive network. The so-called interactome network is the complete collection of all physical protein-protein interactions that can take place within a cell. The first large scale protein interaction studies were done in yeast but have more recently been done in the fly and the worm. As it can be expected, the ultimate goal for the research community is the reconstruction of the human interactome. A comprehensive and accurate mapping of human protein interaction network [5] has been constructed. Interaction maps were constructed from literature and from

experimental approaches. A catalogue of all human protein-protein interactions is seen as a crucial prerequisite to understand how cells function and to decipher the general principles governing this function. Importantly, such information should also enhance the understanding of complex disease processes such as cancer. In various bioinformatics analyses, the authors collected information, concerning human interactome, and constructed maps by identifying conserved orthologous interactions [6]. However, transferring interaction information from model organisms to humans has been shown to be a difficult task.

In order to understand disease mechanisms and signalling cascades, smaller protein interaction networks, representing part of the human interactome, were generated. For instance the interaction network for Huntigton`s disease included 186 interactions and the network for the transforming growth factor-ß signalling pathway contained 755 interactions. Moreover, a study of the interaction attributes of all known human cancer genes has been attempted, where it was shown that cancer proteins display a different global topology from non-cancer proteins [7] . This study clearly demonstrated the central role of cancer proteins within the human interactome. The human protein interactome has revealed information about potential new target genes responsible for genetic diseases.

Protein interaction databases

Many databases have been created in order to store protein-protein interactions. These data have been derived from high-throughput approaches, automated text mining techniques, and/or manually from the scientific literature. The most popular databases that include data concerning human protein interactions are HPRD [8], BIND [9] , MINT [10] and IntAct [11]. A more comprehensive review of protein interaction databases to date is presented in Table 1.

There are some other databases where protein interactions are predicted by computational methods and not obtained from experimental methods. The most significant one is called Online Predicted Human Interaction Database (OPHID) [12] and combines the data that are stored in HRPD, BIND and MINT databases with *in silico* predicted data. The STRING database has integrated known and predicted interactions from a variety of sources as well [13].

The HPRD, the OMIM and the PDB databases, due to their importance in relation to this report, will be presented in more detail in the following paragraphs. PDB is not a protein interaction database but a database that provides information about the structure of the proteins.

OMIM

Online Mendelian Inheritance in Man database (OMIM) [15] is an online database focusing on human genes and genetic disorders. Initially, it was based on Dr. Victor A. McKusick's book entitled "Mendelian Inheritance in Man". Today, the online database OMIM is distributed electronically by the National Center for Biotechnology Information (NCBI). It is updated daily and provides links to a variety of related resources. OMIM catalogues all the known diseases with a genetic component and it links them to the relevant human genes, when this information is available. Each entry has textual information and it is accompanied with references. Many other databases, including HPRD, are based on the entries provided by OMIM which strengthens the trust in the quality of the data included within the database.

HPRD

The Human Protein Reference Database (HPRD) [20] is an online database providing information about human proteins. This database was developed from Dr. Akhilesh Pandeyat's team in Johns Hopkins University and the Institute of Bioinformatics. The database includes domain architecture, protein functions, protein-protein interactions, post-translational modifications, subcellular localization and disease association of genes. HPRD also reports interactions of proteins with other nucleic acids and small molecules. HPRD is a curated database, where data are derived manually by expert biologists reading the published literature. The larger part of HPRD data is derived from in vitro methods.

The property that makes HPRD a very important database is that it contains information concern-

Databases	Features	Web links	References
BIND	Binary molecular interactions, molecular complexes and pathways	http://www.blueprint.org/bind/bind.php	[9]
DIP	PPI data manually curated from literature	http://dip.doe-mbi.ucla.edu/	[14]
HPRD	Human PPIs, information about post-translational modifications, subcellular localization, protein domain architecture, tissue expression and human disease associations	http://www.hprd.org/	[8]
OMIM	Information on human genes and genetic disorders	http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM	[15]
PDB	3D structural information, relationships to sequence, function, and diseases	http://www.rcsb.org/pdb/home/home.do	[16]
MINT	Experimentally verified protein interactions	http://mint.bio.uniroma2.it/mint/Welcome.do	[10]
MIPS	Mammalian interaction data manually curated from literature	http://mips.gsf.de/proj/ppi/	[17]
IntAct	Interactions, experimental methods and literature citation of human proteins. No species restriction	http://www.ebi.ac.uk/intact/site/index.jsf	[11]
PDZBase	PPIS involving protein with PDZ domains, confirmed in vitro and in vivo experiments	http://icb.med.cornell.edu/services/pdz/start	[18]
Reactome	Pathways and biochemical reactions in humans	http://www.genomeknowledge.org/	[19]
STRING	Known and predicted protein-protein interactions from various organisms	http://string.embl.de/	[13]
OPHID	Predicted human protein- protein interactions	http://ophid.utoronto.ca/ophid/index.html	[12]

Table 1. The most important protein interaction databases.

ing the connection of many proteins with diseases. This kind of information is retrieved from OMIM database, where the disease genes form these proteins are annotated. The HPRD is a database that connects proteins with diseases based on OMIM database information. Moreover, HPRD has information about protein modifications which are very important as they are related with diseases. The identification of protein modifications can lead to the design of new and more effective drugs.

PDB

The Protein Data Bank (PDB) [16] is an online database that provides a wealth of information about the structures of biological macromolecules and their relationships to sequence, function, and disease. The PDB was established in 1971 at Brookhaven National Laboratory and is continuously updated with new structures. Furthermore, a big collection of tools is provided to help the study of biological macromolecules structure. The function of this online database is ensured by the RCSB (Reasearch Collaboratory for Structural

Bioinformatics), which is a member of the wwPDB [21]. RCSB ensures that the PDB archive remains an international resource with uniform data.

The PDB is the single worldwide depository of information concerning the three-dimensional structures of large biological molecules, including proteins and nucleic acids. Therefore, its importance in drug designing project is indisputable.

Druggable protein interactions

Protein interactions appear in every single living cell. They are crucial for function and growth and are involved in various cellular pathways. Abnormal behavior of protein interactions and protein complexes play a key role in various diseases. Therefore, the identification of molecules preventing the formation of the complex or interaction of the proteins of the under question complex could be valuable drug targets. The inhibitors design is a very hot topic in drug discovery nowadays and one of the major goals of many drug design projects.

Below two examples are provided, a known and interesting example of a drug target and another example of a new challenging target in drug therapeutic agents. Great effort has been made in order to design an inhibitor for the interaction of the complex formed by the transcription factor p53 and the murine double minute 2 (MDM2). Our second example focuses on the new promises and challenges in applying gamma-secretase inhibitors as therapeutic agents for cancer.

Inhibitors for p53-MDM2 complex

P53 is a transcription factor known to be involved in various biological processes such as cell-cycle regulation, apoptosis, DNA repair, and differentiation [22]. The mutation or deletion of this transcription factor has disastrous consequences since such phenomena have been correlated with human cancer [23]. MDM2 is a negative regulator of the p53 tumor suppressor. Overexpression of MDM2 found in several tumor cases can lead to inactivation of p53, since MDM2 constantly inhibits p53. These interactions were taken into consideration, in order to design molecules that would prevent the p53-hmd2 interaction and therefore cancer.

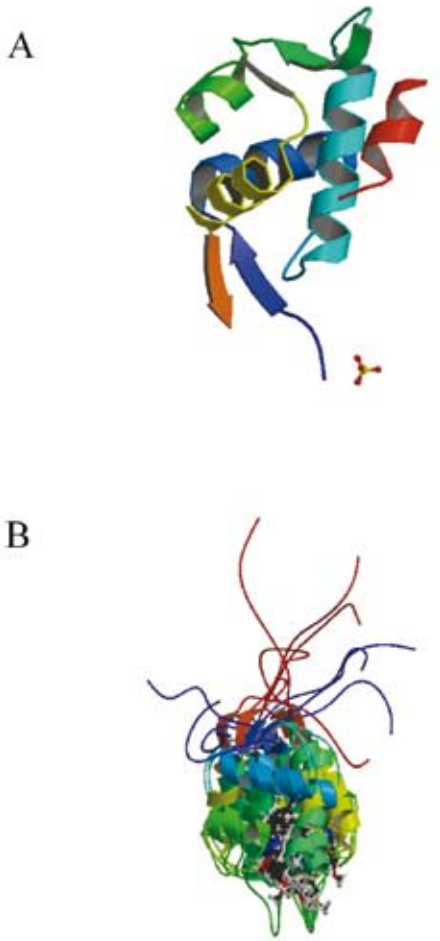


Figure1. A) X-RAY Structure of human MDM2 in complex with an optimized p53 peptide: PDB: 1T4F [24]. B) NMR Structure of a Complex Between MDM2 and a Small Molecule Inhibitor: PDB: 1TTV [25].

G-secretase- Cancer therapeutic agents

Gamma-secretase is a protease with catalytic activity, and cleaves mostly type I membrane proteins such as Notch receptor and amyloid beta peptide precursor. Several g-secretase inhibitors have been developed for this enzyme for the treatment of Alzheimer due to its role in cleaving beta-amyloid precursor in the brain. Inhibition of amyloid ß-peptide (Aß) production by blocking γ-secretase activity is at present one of the most

promising therapeutic strategies to slow progression of Alzheimer's disease. Gamma-secretase inhibitors (GSIs), as well as various biopharmaceutical or genetic Notch signaling inhibitors have been suggested as potential novel cancer therapeutic strategies.

Applications of computational techniques for drug designing

One of the relatively new approaches concerning drug designing, is to find / build molecules targeting proteins involved in protein-protein interactions which have a vital role in the appearance of a disease. There is a variety of target selection schemes, ranging from focusing on only novel folds to selecting all proteins in a model genome [26]. Nowadays, there are many different data sources that provide researchers with information about the function of proteins, their molecular features and so on. Therefore, the combination of these data resources with various computational techniques can be promising for drug designing.

Machine learning techniques are suitable to be applied on the drug designing area. Genetic and evolutionary algorithms have been applied to predict the most suitable starting drug target [27]. The molecule conformational search and the handling of the chemical structure of the new molecule are two areas, where the above mentioned algorithms find application. Genetic and evolution algorithms have been used in conformational analysis, in ligand docking and in de-novo design. Besides this, neural networks can be combined with genetic algorithms to analyze the similarity or the diversity of combinatorial libraries or mapping the molecular surface into a 2D plane [28]. Additionally, other machine learning methods such as classifiers and more accurately Support Vector Machines (SVM) are used to perform structure-activity relationship analysis. It has been shown that these techniques can achieve better results than those derived from neural networks [29].

Genetic algorithms and neural networks are not the only methods of artificial intelligence that are being used for drug designing. Genetic programming is also being applied to predict the bioavailability of a molecule from its chemical structure [30]. Moreover, genetic programming has been used to find molecules functioning as bites and simultaneously minimizing the negative side effects.

The applications of computational techniques for drug designing are not limited to the above mentioned areas. Various types of clustering have been applied to identify the suitable drug target. The most "traditional" one, named the fingerprint-based clustering [31], is the clustering where the different identified chemical classes are active against the target. This approach produces very often clusters with different structural features. Therefore, many variations combining different clustering strategies have been proposed in order to eliminate this negative effect. The one introduced by Stahl and Mauser [32] is considered as the most important one. Additionally, stochastic nature clustering such as fuzzy clustering can be used for data reduction in order to identify clusters of conformations or the most representative clusters that can be used as potential binds of the ligand [33].

Conclusions

The design of new molecules that inhibit or induct protein interactions has emerged as a really promising field in drug designing. In this report, we gave a small review of the protein interactions, their online depositories and two very characteristic examples of the impact that protein interactions have on drug designing. Besides this, we presented different computational methods used to select the most appropriate molecule as a "good" target for dug discovery. It is almost certain that in the future new techniques or combinations of them will be applied in the drug designing problem.

References

1. Ryan, D.P. and J.M. Matthews, Protein-protein interactions in human disease. Curr Opin Struct Biol, 2005. 15(4): p. 441-6.
2. Nooren, I.M. and J.M. Thornton, Diversity of protein-protein interactions. EMBO J, 2003. 22(14): p. 3486-92.
3. Luscombe, N.M., et al., An overview of the structures of protein-DNA complexes. Genome Biol, 2000. 1(1): p. REVIEWS001.

4. Jones, S., et al., Protein-DNA interactions: A structural analysis. J Mol Biol, 1999. 287(5): p. 877-96.

5. Stelzl, U., et al., A human protein-protein interaction network: a resource for annotating the proteome. Cell, 2005. 122(6): p. 957-68.

6. Lehner, B. and A.G. Fraser, A first-draft human protein-interaction map. Genome Biol, 2004. 5(9): p. R63.

7. Jonsson, P.F. and P.A. Bates, Global topological features of cancer proteins in the human interactome. Bioinformatics, 2006. 22(18): p. 2291-7.

8. Peri, S., et al., Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res, 2003. 13(10): p. 2363-71.

9. Alfarano, C., et al., The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res, 2005. 33(Database issue): p. D418-24.

10. Zanzoni, A., et al., MINT: a Molecular INTeraction database. FEBS Lett, 2002. 513(1): p. 135-40.

11. Hermjakob, H., et al., IntAct: an open source molecular interaction database. Nucleic Acids Res, 2004. 32(Database issue): p. D452-5.

12. Brown, K.R. and I. Jurisica, Online predicted human interaction database. Bioinformatics, 2005. 21(9): p. 2076-82.

13. von Mering, C., et al., STRING 7--recent developments in the integration and prediction of protein interactions. Nucleic Acids Res, 2007. 35(Database issue): p. D358-62.

14. Xenarios, I., et al., DIP: the database of interacting proteins. Nucleic Acids Res, 2000. 28(1): p. 289-91.

15. Hamosh, A., et al., Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res, 2005. 33(Database issue): p. D514-7.

16. Berman, H.M., et al., The Protein Data Bank. Nucleic Acids Res, 2000. 28(1): p. 235-42.

17. Pagel, P., et al., The MIPS mammalian protein-protein interaction database. Bioinformatics, 2005. 21(6): p. 832-4.

18. Beuming, T., et al., PDZBase: a protein-protein interaction database for PDZ-domains. Bioinformatics, 2005. 21(6): p. 827-8.

19. Joshi-Tope, G., et al., Reactome: a knowledgebase of biological pathways. Nucleic Acids Res, 2005. 33(Database issue): p. D428-32.

20. Peri, S., et al., Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res, 2004. 32(Database issue): p. D497-501.

21. Berman, H., K. Henrick, and H. Nakamura, Announcing the worldwide Protein Data Bank. Nature Structural Biology, 2003. 10(12): p. 980-980.

22. Vousden, K.H. and X. Lu, Live or let die: the cell's response to p53. Nat Rev Cancer, 2002. 2(8): p. 594-604.

23. Soussi, T., K. Dehouche, and C. Beroud, p53 website and analysis of p53 gene mutations in human cancer: forging a link between epidemiology and carcinogenesis. Hum Mutat, 2000. 15(1): p. 105-13.

24. Grasberger, B.L., et al., Discovery and cocrystal structure of benzodiazepinedione HDM2 antagonists that activate p53 in cells. J Med Chem, 2005. 48(4): p. 909-12.

25. Fry, D.C., et al., NMR structure of a complex between MDM2 and a small molecule inhibitor. J Biomol NMR, 2004. 30(2): p. 163-73.

26. Brenner, S.E., Target selection for structural genomics. Nat Struct Biol, 2000. 7 Suppl: p. 967-9.

27. Lameijer, E.-W., et al., Evolutionary Algorithms in Drug Design. Natural Computing, 2005. 4: p. 177-243.

28. Anzali, S., et al., The use of self-organizing neural networks in drug design. Perspectives in Drug Discovery and Design, 1998. 9-11: p. 273-299.

29. Kim, H. and H. Park, Incremental and Decremental Least Squares Support Vector Machine and Its Application to Drug Design, in the 2004 IEEE Computational Systems Bioinformatics Conference (CSB'04). 2004 IEEE Computer Society press: Stanford, CA. p. 656-657.

30. Archetti, F., et al., Genetic programming for human oral bioavailability of drugs in the 8th annual conference on Genetic and evolutionary computation A. Press, Editor. 2006, ACM: Seattle, Washington, USA. p. 255-262

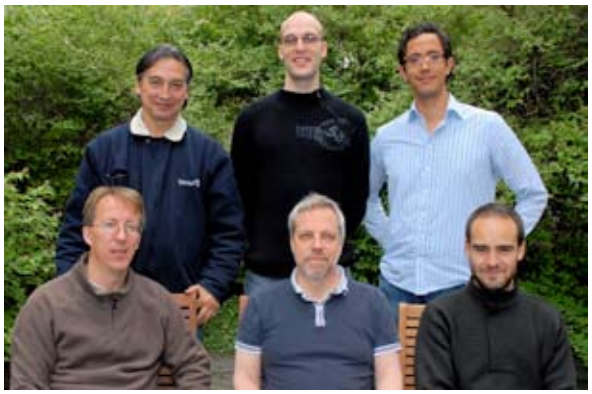
31. Schnecke, V. and J. Bostrom, Computational chemistry-driven decision making in lead generation. Drug Discov Today, 2006. 11(1-2): p. 43-50.

32. Stahl, M. and H. Mauser, Database clustering with a combination of fingerprint and maximum common substructure methods. J Chem Inf Model, 2005. 45(3): p. 542-8.

33. Banerjee, A., et al. Fuzzy clustering in drug design: Application to cocaine abuse. in Fuzzy Information, 2004. Processing NAFIPS '04. IEEE Annual Meeting of the. 2004.

A guide to EVALLER (2.0) web server:

A new tool for in silico testing of protein allergenicity



Erik Bongcam-Rudloff[1,2], Daniel Edsgärd[4*], Alvaro Martinez Barrio[1,2], Daniel Soeria-Atmadja[4,5], Mats G. Gustafsson[3,4] and Ulf Hammerling[5]

[1.] Linnaeus Centre for Bioinformatics, Uppsala Biomedical Centre (BMC), Uppsala University, Uppsala, Sweden
[2.] Dept of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden
[3.] Dept of Engineering Sciences, Uppsala University, Uppsala, Sweden
[4.] Dept of Medical Sciences, Academic Hospital, Uppsala, Sweden
[5.] Division of Toxicology, National Food Administration, Uppsala, Sweden
* Present address: BioCentrum-DTU, Technical University of Denmark, Lyngby, Denmark

Introduction

Allergic reactions represent an inappropriate immune response to foreign antigens that are otherwise relatively harmless. This can occur in atopic individuals, who have a genetic predisposition to develop an allergic immune reaction. The immunoglobulin E (IgE)-mediated allergy, also known as immediate-type hypersensitivity, is a common variant among this category of ailment and is typically associated with reactions against weed-, tree- or grass pollen, insect venom, animal dander and several food commodities. The immunological response in a sensitized individual, upon renewed contact with the allergen, may ultimately precipitate into a variety of disease symptoms, the anaphylactic shock being the most serious outcome [1-3]. Moreover, reactions can occur in atopic individuals as a result of exposure to proteins other than those that have elicited the immunological sensitization. This phenomenon is generally referred to as cross-reactivity and proteins extensively involved in such responses are designated pan-allergens [4, 5].

A number of immunochemical and immunological methods have been developed for research purpose and/or for implementation in allergy diagnostic settings. Some among them are, though, also used as tools for risk assessment of protein allergenic potential, in the context of an IgE mediated disease mechanism. Such methods include IgE immunosorbent assays using patient sera, human skin prick tests and basophil histamine release determination. The double-blind placebo-controlled food challenge (DBPCFC) is still a gold standard to assessing immediate-type allergic reactions to ingested components [6-9]. The utilization of molecular genetic technology within the crop breeding business - leading to the advent of genetically modified (GM) food crops and analogously engineered microorganisms, both expressing transgene-encoded products intended to enter the food chain - have, however, prompted for speedy means to scan for protein allergenicity. Hence, bioinformatics-type inspection for this property has attracted an appreciable interest among experts in this field. Actually, a testing scheme for GM-crop safety assessment, including an introductory amino acid sequence comparison step, was suggested already in 1996 [10]. Since that time dedicated bioinformatics methods, largely aiming at the specific identification of allergens through comparing amino acid sequences of allergens and proteins unlikely to elicit such response, have gradually gained increased recognition. This is because such interrogations, apart from their expedient nature, can be accomplished inexpensively and, due to a gradual increase in performance, with reasonably high accuracy [11-13]. It is, however, important to keep in mind that bioinformatics-type inspection for IgE-type allergenic potential should

be used as a scanning practice and as part of an integrated allergology risk assessment procedure.

We and other research groups have developed algorithms devoted to the specific detection of amino acid sequences of allergens, some of them being available as web servers [12-14]. A comprehensive description of EVALLER appears in the *Nucl Acids Res,* Web Server Issue of July 2007 [15].Our new server, designated EVALLER 2.0 and being a substantial revision of the 1.0 counterpart (which still is available, though), is built to provide high-accuracy testing, as accomplished by a unique algorithmic concept and by an enhanced graphical and textual output, which altogether support allergology risk assessment.

Co-recognition and cross-reactivity

Strictly, co-recognition refers to a situation involving either sensitized-mediated responses to several different allergens (multi-sensitation) or reactions due to IgE-binding to antigens with a structural resemblance to such proteins, whereas cross-reactivity refers to the latter sort of response. Typically, homologous allergens from phylogenetically related species are involved in cross-reactions, but promiscuous IgE epitope recognition can also occur between distantly related organisms. Notably, cross-reactive proteins are generally members of the same protein fold family, but not all such members cross-react. This is because there is a complex relationship between protein primary structure and allergenic potential. Commonly 70 %, and at least 50 %, identity at the amino acid sequence is seen between IgE cross-reactive antigens [16]. This rule of thumb has, however, many exceptions because other determinants, apart from the amino acid sequence, can contribute to protein structure [17, 18].

For many allergen sources a rather elaborated image of cross-reactive patterns has emerged in recent years, e.g. patients allergic to major pollen allergens in birch, mugworth or ragweed may also react with a variety of fruits and/or vegetables due to shared epitopes. Some common examples thereof, referred to as the pollen-fruit syndrome, involve the birch/apple/cherry, mugwort/cele¬ry/spice and ragweed/water-melon/banana groups [19, 20]. Several other reactivity patterns, typified by the latex/banana/avocado and mite/snail/shrimp, are also well characterised [21, 22]. Moreover, an exhaustive multivariate data analysis of IgE reactivity within a large data set has recently been reported. Many established reactivity patterns as well as several not formerly documented associations were delineated [23].

Bioinformatics in allergology risk assessment

No single predictive test is able to securely identify the allergenic potential. In response to this situation two flow-charts schemes, each depicting a tiered set of testing methods, were proposed by international regulatory/advisory bodies [10, 24]. Both schemes strongly favoured a bioinformatics interrogation approach, although being simplistic at that time, which actually was recommended as an initial screening step entailed to major consequences if alarming for allergenicity. Subsequently, the international organisation providing frames for food law - *Codex Alimentarius* – adopted a selected menu of the aforementioned suggestions [25]. The generally simplistic approaches of that time (alignment over 80 amino acid segments and scanning for identical matches of either 6 or 8 residues) could, however, give rather inaccurate outputs [26]. The alignment-based part of this scheme has, however, proved considerable more useful for the purpose of protein allergen detection than the exact match search procedure. It is founded on a widely known sort of algorithm, typically appearing as FASTA or BLAST, which has proven highly useful in a wide range of bioinformatic applications. It is nonetheless connected with constraints in the allergenicity detection context. For example, allergens tend to cluster in relatively few protein families; algorithms dedicated to protein family identification may accordingly return relatively satisfactory scores in a general-oriented performance test [27, 28]. Most members of such families, however, appear innocuous in an allergenicity context and more sophisticated measures, relative to those relying on straightforward alignment alone, are needed to tackle this and other hurdles. Moreover, the FAO/WHO report prescribes the search to be conducted over a sliding window of 80 amino acid residues,

whereas this constraint is not set for the default FASTA alignment procedure. A recent report indicates that the conventional FASTA procedure returned fewer false positive results, compared with that of FAO/WHO [29].

The last several years, though, have witnessed a methodological advancement in bioinformatics-based risk assessment within the allergology field. Such refinements include usage of allergen-derived motifs, known IgE epitopes, alignment-based feature extraction and a variety of amino acid sequence coding protocols as well as the introduction of statistical learning algorithms [12, 13]. A fast development of dedicated and publicly accessible catalogues/knowledge repositories on allergens, typified by Allergome, SDAP and AllergenOnline, has undoubtedly been a major prerequisite for algorithmic development in this field [11, 13]. Despite significant advancements in the area, the specific detection of amino acid sequences of potential allergens remains a very challenging undertaking. Notably, hitherto devised bioinformatics-type protein-allergenicity inspection systems probably have their highest level of aptitude within recognition of cross-reactivity of the B-cell sort, i.e. detection of promiscuous IgE-binding to structurally similar proteins. On the assumption that allergens represent a relatively small subset of proteins *in silico* detection procedures should also be reasonably appropriate for the identification of *de novo* protein allergens. There are, though, additional mechanisms and factors involved in the establishment and maintenance of an IgE-type allergic response, such as T-cell cross-recognition, being less characterised than that involving immunoglobulins, as well as route of exposure.

Major architectural characteristics of EVALLER and its SVM core

The core application of EVALLER is built on a trained supervised classification system, which we – in a provisional form – reported already in 2005, but an enhanced version appeared about one year later [14, 30]. Briefly, amino acid sequences of allergens – cut in overlapping segments – are compared to a data set largely composed of the human proteome to ultimately generate an assembly of peptides that presumably would represent an enriched set, with respect to allergenic properties. This set, referred to as the FLAP set, serves as a reference catalogue to generate features (alignment score values) in the ensuing training of a Support Vector Machine (SVM) algorithm, using both allergens and (presumable) non-allergens. Interrogation of a query amino acid sequence, with respect to allergenic potential, involves its presentation to the educated SVM algorithm. Performance tests of the accordingly constructed SVM have revealed good overall accuracy, but also a unique ability to correctly assign non-allergens in protein families known to hold many allergens. Moreover, a low proportion of the SwissProt repository was assigned as allergens, indicating low false-alarm rate [14].

EVALLER 2.0 features

EVALLER 2.0 represents a significant enhancement over the 1.0 counterpart, due to improvements within three main categories: i) Our in-house compiled repository of allergens has been replaced by that of AllergenOnline holding an appreciably larger set of amino acid sequences, ii) the filter data set (used to create FLAPs through an alignment comparison procedure involving segmented allergens and proteins with low probability of being allergenic) is compiled from the human proteome only, and iii) the server is fully integrated into the National Food Administration's web interface. Moreover, the significant revision of data is also accompanied by algorithmic re-training.

Figure 1. Opening view of EVALLER 2.0. This page, as well as that of sequence pasting/uploading and all other views, are fully integrated into the National Food Administration's international web-site. Instructions for use and selected facts on allergenicity testing are readily available in left and right panes.

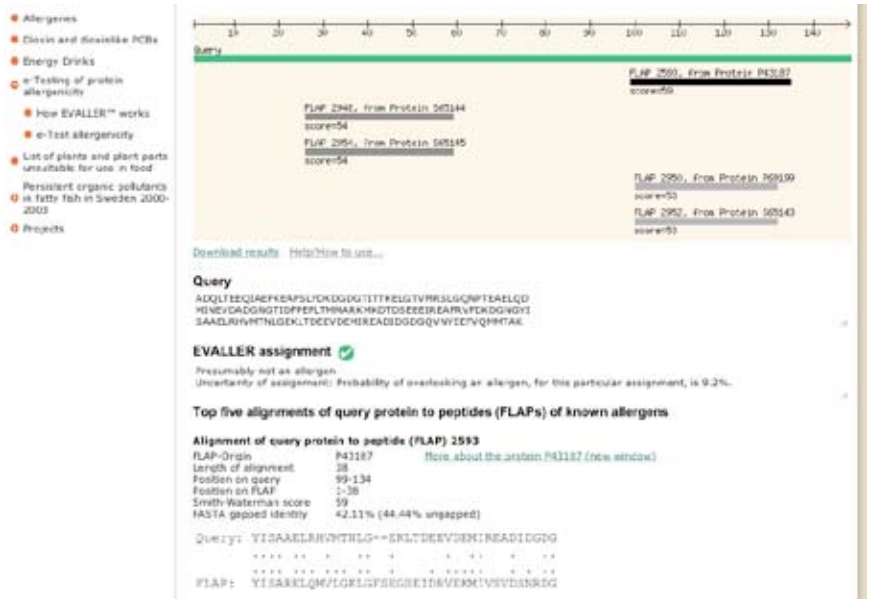


Figure 2. Example result output of a query, concluded on by EVALLER 2.0 as not likely being an immediate-type allergen. The image holds three categories of information: i) a graphical display showing a bar representation of the query amino acid sequence and, below, several best-matching FLAPs in descending gray intensity, in proportion to degree of similarity to the query sequence, ii) the EVALLER assignment, including an uncertainty estimation and iii) FASTA alignments, including their respective associated specifications, to all FLAPs occurring as bars. In addition to the assignment "Presumably not an allergen" the test result is indicated by the appearance of an adjacent green disc-shaped mark (with an OK sign) and by appearance of the query sequence bar representation in green colour. The image is truncated in its bottom part.

The server is altogether designed to enable straightforward practice by users not appreciably trained in bioinformatics. A front page (`http://www.slv.se/templates/SLV_Page.aspx?id=19259&epslanguage=EN-GB`) (Figure 1) holds brief introductory information on EVALLER 2.0, including its purpose, but the major guidance documentation is accessible via several pop-up links on either side of front-page centre. The links "Introduction", "Essentials of EVALLER" and "How EVALLER works" all summarises key parts of the server function. Users who have an interest to acquire deeper understanding of the EVALLER algorithmic design can consult the "Documentation" link.

A new inquiry session is invoked by either a mouse-click on the link "e-Test allergenicity" (located in left part of the panel) or simply on the centrally positioned EVALLER logotype. The accordingly activated query page holds two panels for data input (pasting and uploading) as well as a pane for optional specification of a desired number of matching peptides (FLAPs), to become visible in the results display. The default setting (5 matches), will be sufficient in most cases. Sequences shorter than 40 amino acid residues are considered inappropriate to testing with EVALLER and will therefore be dismissed. The output page holds three major sorts of information being directly connected with a specific interrogation (Figures 2 and 3):

1. A top panel featuring a bar representation of the query protein and one or several analogously displayed matching FLAP(s). Red and green colours of the query sequences representation indicate conclusions on likely allergenicity and presumed non-allergenicity, respectively. The FLAP representation bars - occurring below - are shown in descending greyscale, signifying decreasing degree of similarity. Essential in this context is that peptide matches, as well as the associated output (see below), are shown regardless of EVALLER assignment category, i.e. without respect to whether the query protein is assessed as being allergen or non-allergen. This feature, particularly in ambiguous cases, will help the user to proceed with investigation independently of output.

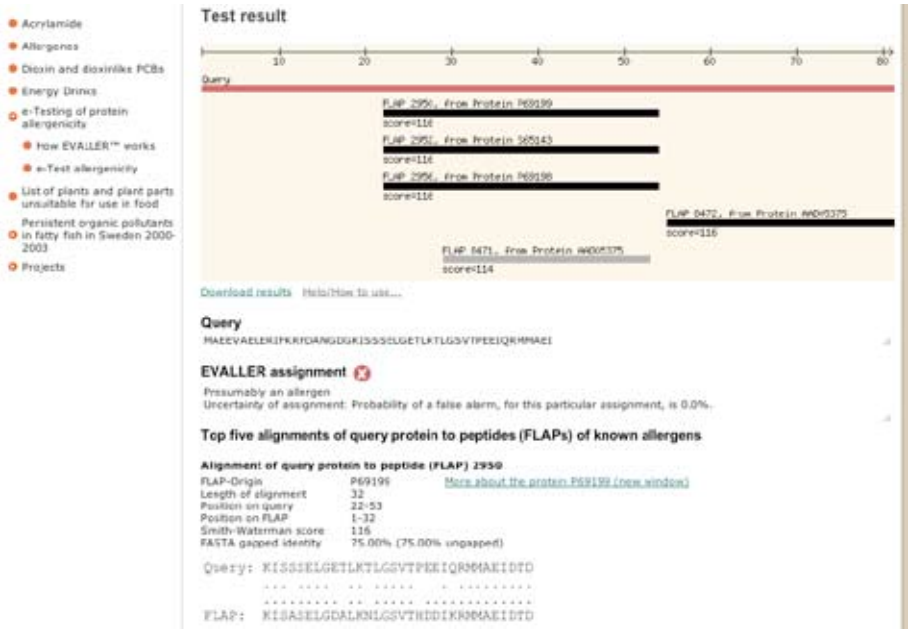


Figure 3. Example result output of a query, as judged by EVALLER 2.0 as being a likely immediate-type allergen. The display is equivalent to that of Figure 3, but shows the assignment "Presumably an allergen", and the associated indicators of an accordingly made assessment, i.e. a red bar representation of the amino acid sequence (uppermost part of image) and a red-coloured disc-mark (with an X-sign) in the central part of the view. The image is truncated in its bottom part.

2. An assignment, being either *Presumably an allergen* or *Presumably not an allergen*, including a figure on the computed uncertainty. This part is defined specifically for each of these two output categories. For query proteins judged to be likely allergens, the uncertainty value reflects the probability of EVALLER reporting a false alarm for this assignment. In the event of a query being assigned as a presumable non-allergen, the uncertainty value corresponds to the probability of EVALLER overlooking an allergen, in the context of the particular assignment.

3. Information on alignments of query amino acid sequence and the best matching FLAPs is displayed below outputs on assignment and statistics. Each alignment is accompanied by an external NCBI-associated identification link of the allergen corresponding to the best matching FLAP. A click on this link will redirect the user to the NCBI protein database, wherein a more detailed description of the allergen can be found and retrieved.

Although there are several novelties in EVALLER 2.0, notably a revised FLAP set, a re-educated algorithm and an appreciably changed layout, the article on EVALLER 1.0 is yet strongly recommended to consult for a more inclusive outline of system structure and detailed features, including the SVM algorithm [15].

EVALLER perspectives

Major principles, basic operation design as well as key elements of the current layout will remain largely unchanged over the near future. We are, though, considering a few minor amendments of EVALLER 2.0 that might appear in 2008. The foremost function to implement is the alignment-type procedure, but certainly not the exact motif match search protocol, as proposed by FAO/WHO already in 2001[24]. Although an array of significantly more accurate computational detection systems for protein allergens have been devised since that time, the simple alignment-based inspection has yet established itself as a sort of reference method. Because of this and while it is still frequently consulted for the above-mentioned purpose, hopefully with a critical attitude, this interrogation practise is considered to be available conveniently within the EVALLER 2.0 environment.

In a more general context, tertiary (and quaternary) structure would naturally be highly preferred

for *in silico* modelling purposes, but to date only some 50 allergens (less than 5 % of currently documented allergens) are accordingly characterized. Hence, the primary structure will remain a predominant source of information for developers of bioinformatics risk assessment tools in allergology for many years ahead. There should, however, still be room for refinement of algorithms designed to specifically identify proteins associated with potential IgE-mediated allergenicity, based on amino acid sequence information only. Seemingly, appropriate combinations of dedicated pattern recognition algorithms hold some promise for an overall enhanced predictive performance of allergology e-Testing methods.

Acknowledgements

We are grateful for financial support to this project from the Cancer- and Allergy Fund, Stockholm and the European Model for Bioinformatics Research and Community Education (EMBRACE). We are indebted to Christer Andersson, at the National Food Administration, for critically reading the manuscript.

References

[1] Kay AB. New England Journal of Medicine 2001;344(2):109-113.

[2] Vercelli D. J Allergy Clin Immunol 2005;116(1):60-64.

[3] Woodfolk JA. Curr Allergy Asthma Rep 2005;5(3):227-232.

[4] Ferreira F, Hawranek T, Gruber P, Wopfner N, Mari A. Allergy 2004;59(3):243-267.

[5] Wopfner N, Gadermaier G, Egger M, Asero R, Ebner C, Jahn-Schmid B, Ferreira F. Int Arch Allergy Immunol 2005;138(4):337-346.

[6] Bock SA, Sampson HA, Atkins FM, Zeiger RS, Lehrer S, Sachs M, Bush RK, Metcalfe DD. Journal of Allergy and Clinical Immunology 1988;82(6):986-997.

[7] van der Zee JS, de Groot H, van Swieten P, Jansen HM, Aalberse RC. J Allergy Clin Immunol 1988;82(2):270-281.

[8] Kimber I, Dearman RJ, Penninks AH, Knippels LM, Buchanan RB, Hammerberg B, Jackson HA, Helm RM. Environ Health Perspect 2003;111(8):1125-1130.

[9] Hamilton RG, Adkinson NF, Jr. J Allergy Clin Immunol 2004;114:213-225.

[10] Metcalfe DD, Astwood JD, Townsend R, Sampson HA, Taylor SL, Fuchs RL. Crit Rev Food Sci Nutr 1996;36 Suppl:S165-S186.

[11] Brusic V. Inflammation & Allergy - Drug Targets 2006;5(1):35-42.

[12] Kong W, Tan TS, Tham L, Choo KW. In Silico Biol 2007;7:0006.

[13] Schein CH, Ivanciuc O, Braun W. Immunol Allergy Clin North Am 2007;27(1):1-27.

[14] Soeria-Atmadja D, Lundell T, Gustafsson MG, Hammerling U. Nucleic Acids Res 2006;34(13):3779-3793.

[15] Martinez Barrio A, Soeria-Atmadja D, Nister A, Gustafsson MG, Hammerling U, Bongcam-Rudloff E. Nucleic Acids Res 2007;35(Web Server issue): W694-700.

[16] Aalberse RC. J Allergy Clin Immunol 2000;106(2):228-238.

[17] Astwood JD, Silvanovich A, Bannon GA. J Allergy Clin Immunol 2002;110(1):26-27.

[18] Aalberse RC. Chem Immunol Allergy 2006;91:134-146.

[19] Hannuksela M, Lahti A. Contact Dermatitis 1977;3(2):79-84.

[20] Kremser M, Lindemayr W. Wien Klin Wochenschr 1983;95(23):838-843.

[21] M'Raihi L, Charpin D, Pons A, Bongrand P, Vervloet D. J Allergy Clin Immunol 1991;87(1 Pt 1):129-130.

[22] De Maat-Bleeker F, Akkerdaas JH, van Ree R, Aalberse RC. Allergy 1995;50(5):438-440.

[23] Soeria-Atmadja D, Onell A, Kober A, Matsson P, Gustafsson MG, Hammerling U. J Allergy Clin Immunol 2007, e-pub ahead of print.

[24] FAO/WHO. Evaluation of allergenicity of genetically modified foods. Rome, Italy: Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology., 2001.

[25] Codex. Codex Alimentarius Commission 2003 (ALINORM 03/34A). Guideline for the conduct of food safety assessment of foods derived from recombinant DNA plants. Annex on the assessment fo possible allergenicity, Rome, Italy. Yokohama: Codex Alimentarius Commission, 2003.

[26] Poulsen LK. Mol Nutr Food Res 2004;48(6):413-423.

[27] Jenkins JA, Griffiths-Jones S, Shewry PR, Breiteneder H, Mills EN. J Allergy Clin Immunol 2005;115(1):163-170.

[28] Radauer C, Breiteneder H. J Allergy Clin Immunol 2006;117(1):141-147.

[29] Ladics GS, Bannon GA, Silvanovich A, Cressman RF. Mol Nutr Food Res 2007;51(8):985-998.

[30] Bjorklund AK, Soeria-Atmadja D, Zorzet A, Hammerling U, Gustafsson MG. Bioinformatics 2005;21(1):39-50.

MRS version 3

EMBOSS and wrappers4EMBOSS

Guy Bottu

Belgian EMBnet Node (BEN),
ULB Campus de la Plaine, blv. du Triomphe, 1050 Brussels, Belgium

MRS goes version 3

In EMBnet.news 12(2) I had presented an article about the use of MRS as sequence retrieval mechanism for EMBOSS. Since then MRS has progressed from version 2 to version 3 and most MRS sites have upgraded. The architecture of version 3 is different : the Web interface has been rewritten in Java-Struts, is sitting behind a Tomcat server and communicates with the search engine by means of Web Services. An important augmentation of the functionality is the possibility to perform "ranked" queries, that means that the matching entries that are likely to be most relevant are displayed first, much as is the case in search engines like Google and Yahoo. By now EMBnet nodes are increasingly abandoning SRS (for the reasons we know) and some are considering the freeware and Open Source MRS as a viable alternative. So, it is time to update the information on this subject.

A WWW server as sequence databank access method

From release 4.0 on EMBOSS has a databank access method called "MRS", which allows retrieving one or more sequences from an MRS version 2 WWW server by entry name or accession number. Release 5.0 has been enriched with a new method "MRS3". The syntax of the entry to be added in the file `.../share/EMBOSS/emboss.default` of the EMBOSS installation or in the `.embossrc` file in the home directory of the user looks like

```
DB cmbi_sw [ type: P comment: 'SwissProt at
CMBI'
             method: mrs3 dbalias: sprot
format: swiss
             url: 'http://mrs.cmbi.ru.nl/
mrs-3/plain.do'
]
```

Unfortunately, because of the way the current version of MRS handles URL's, this method supports only the retrieval of one sequence at a time (in EMBOSS parlance a "methodentry"). Furthermore it is possible to specify the sequence by entry name only. On the other hand, a fortunate simplification is that MRS 3 supports composite databank indexes, so that with most MRS sites you can e.g. specify a "dbalias" **uniprot** instead of needing to write something like **sprot+trembl**.

A local installation as sequence databank access method

In the previous article it had been explained how, when you have an installation of MRS on your own computer, you can easily write a Perl script for retrieving sequences and declare it in EMBOSS as method of type "app". This piece of information is still valid. Furthermore, a script `mrsseqget.pl` can now be obtained as part of wrapper4EMBOSS version 2.0 (see `http://wemboss.sourceforge.org`). It allows retrieving sequences by any identifier (entry name, accession number or GI number) as well as retrieving several sequences at the same time using wild cards.

A query tool and a scripting tool

The new version of wrappers4EMBOSS includes two other tools that use a local installation of MRS (they however only work with MRS version 3). `mrsindexsearch` can be run interactively in a UNIX terminal or it can (thanks to a small EMBOSS interface) be used under any of the GUI's that were developed for EMBOSS. It allows searching in sequence databanks while providing logical combinations of keywords for the different indexed fields. The result is returned either as a multiple sequence file in fastA format or as an EMBOSS List File[1].

1 mrsindexsearch is actually based on srsindexsearch (also part of wrappers4EMBOSS), which was itself inspired by the old GCG lookup.

Figure 1. An example of the usage of mrsindexsearch under wEMBOSS. The user wants a neat list with all the mouse mRNA sequences. He therefore searches a local copy of the NCBI RefSeq databank. He searches for sequences with in the species name field mus musculus (actually mus and musculus) and in the accession number field either NM _ * or XM _ * (exploiting the fact that in RefSeq the names of the sequences start with a two letter tag indicating the nature of the sequence, NM standing for a mRNA sequence derived from GenBank entries and XM standing for a mRNA from a NCBI automated genome annotation pipeline).

`mrsget.pl` is meant to be used in scripts that keep databanks up-to-date. The output of "`mrsget.pl -help`" makes obvious what it does:

```
usage:
mrsget.pl <databank> <query> <format> [1/0]

supported formats:
id      identifier
entry   complete entry
de      description line
idde    identifier followed by
        description on one line
html    entry in HTML format
seq     raw sequence on one line
fasta   sequence in fastA format
N       (do not return, just count)

1 : append wildcard
0 : do not append wildcard

e.g.  mrsget.pl uniprot 'os:human AND
length>8000' idde 0
```

What about using the Web services?

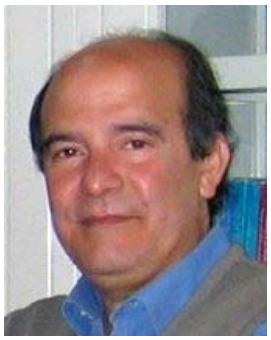
It is possible to use the MRS Web services, either those of a local installation or those of a remote installation that allows outside access. The Web services have now come of age, but information about how to use them has not yet been rendered public. Support for MRS Web services will be included in a future release of wrappers4EMBOSS. In the mean time, for the sake of the curious and the impatient, here follows an example of a script that retrieves from SwissProt the proteases from Papaya latex :

```
#!/usr/bin/perl

use SOAP::Lite;

$ns_url = 'http://mrs.cmbi.ru.nl/mrsws';
$url = 'http://mrs.cmbi.ru.nl/mrsws';
#$url = 'http://localhost:8081/mrs/soap';
$soap = SOAP::Lite->uri($ns_url)-
>proxy($url);

$db = 'sprot';
$query = 'os:carica os:papaya de:
proteinase';
$Nmax = 15;
$queryresult = $soap->call(SOAP::Data-
>name('ns:Find')->
  attr({'xmlns:ns' => $ns_url}) => (
      SOAP::Data->name('ns:db')->type('xsd:
string' => $db),
      SOAP::Data->name('ns:booleanfilter')-
>type('xsd:string' => $query),
      SOAP::Data->name('ns:maxresultcount')-
>type('xsd:int' => $Nmax)
  )
);

@r = $queryresult->paramsall;
$N = $r[0];

for ($i=1;$i<=$N;$i++) {
  $id = $r[$i]{id};
  $getresult = $soap->call(SOAP::Data-
>name('ns:GetEntry')->
    attr({'xmlns:ns' => $ns_url}) => (
      SOAP::Data->name('ns:db')->type('xsd:
string' => $db),
      SOAP::Data->name('ns:id')->type('xsd:
string' => $id),
      SOAP::Data->name('ns:format')-
>type('ns:Format' => 'plain')
    )
  );
  $entry = $getresult->result;
  print $entry;
}
```

The $url variable must be set to the URL pointing to the Web services. Note that instead of format "plain", you can also ask for "title", "html" or "fasta".

Beyond 'The Curse of Babel'[1] in bioinformatics

(is bioinformatics ready for standards?)



Oswaldo Trelles, PhD.

Computer Architecture Department, University of Malaga, Spain

Technological breakthroughs such as high-throughput sequencing and gene-expression monitoring technologies, among others, have nurtured the 'omics' revolution enabling a massive production of data. Although it has become an overused commonplace statement that the volume of data in molecular biology is growing at exponential rates, nonetheless the key feature of biological data is not so much its volume, as its diversity, heterogeneity and dispersion [1].

On the other hand, the accumulation of biological knowledge is fundamental for a more complete view of any biological process. But there is a consensus that only the integrated analysis of this plethora of data will lead to a significant increase in our knowledge, allowing us to address new challenging and complex issues such as system biology, developmental biology, and so on. Unfortunately, this valuable information is disseminated world-wide in the form of biological sequences, structure, expression, pathways, etc. databases, frequently dumped as flat files, image/scheme-based libraries, web-based pages, etc. most of them with proprietary data models and specific services for querying, access and analysis, with no thought of potential external exploitation and integration of such data.

As things stand, this flow of interrelated information is difficult to use and highlights the need of integration of data sources for unified access, independent of possible internal changes, and posing an important technological problem. As a result, there is a snow-balling effect of new ideas emerging and flourishing in the face of these new data sources, compounding the problem.

The shape of the problem has changed over time. Dealing with the exponential growth rates of data in the early era of bioinformatics was a simple problem compared to that posed by diversity, heterogeneity and dispersion of data. Then, the problem was addressed in terms of resources. New computational strategies, based purely on high performance and parallel computing, were enough to provide an illusory and temporary stop-gap solution, perhaps only providing a slot of time to work-out a permanent and more consistent solution.

In the meantime, new paradigms have emerged to deal with the cascade of new requirements. One of the most interesting ones is the workflow concept. Largely used in other different domains such as business, e-commerce, etc., workflows are being regarded as the future way of exploiting data and services in automated and faster mode. Perhaps the key of the success of workflows approach can be identified in the capacity of wiring different services to build-up large and powerful bioinformatics machines on the basis of a common and standardised description of input/output data in the services. A standardised description involves not only syntactic rules but, more importantly, a way of incorporating accurate semantics, so as to extend the knowledge-beyond the specific application domain, thus providing real added value to this data with the capacity of being used even in services that, at present, we are unable to imagine.

However, the gap between data and standards is not an exclusive characteristic of molecular biology and its most prominent tool, bioinformatics. This problem arises every time we try to reconcile the robotic capacity of data production with the human ability to interpret them (e.g., in weather

1 "But God confounded their tongue, so that they did not understand one another's speech, and thus scattered them from that place into all lands, and they ceased to build the city"; Genesis, 11:1-8

forecasting, physics, etc). The solution that works best involves data normalization and standardization giving the data universal utility.

In the light of this, it is surely not difficult for, say, a physician to understand that data will be more used and its potential utility will increase in direct proportion as data are standardised. Even a simple user would clearly benefit from the enrichment of web-pages provided by the use of standard languages such as HTML instead of pure text pages.

Taking another example, we would expect a molecular biologist to be predisposed to use rigorous standards –such as XML- for making their data public. But in fact, it seems that a bio-sequence –i.e. described in FASTA format- is not the same when XML labels are incorporated for strict description of their content. Just at this point, the sequence is no longer a sequence but an instance of an *object model*. But the "XML-ification" of text-formats must only be seen as a way of giving structure to the data whereas what we are advocating will also give semantic meaning, and annotate or label the data with references to external, shared ontologies that can be used for automatic reasoning.

Certainly the solution is straightforward within easy reach: hide the labels and expose only the biological information needed to interpret the data. So what's the problem?

So, the question behind the title of this document still applies: why don't standards works in bioinformatics? It's simply a matter of an appropriated combination of skills. Why isn't that happening in this domain?

Although bioinformatics success has been widely acclaimed -as the name suggests- for the synergy among different branches of science, to this day, in spite of more than 30 years' activity we still continue to work in separate -and even worse, fenced-off, areas of knowledge. There are still biologists with basic computational knowledge and computer specialists with a rudimentary biological base who are in charge of bioinformatics system developments.

When bioinformatics cease to mean the simple addition of skills and becomes a new area of knowledge in its own right, it will be possible to inculcate, from the early stages of training, the need for systematic standardised storage, communication and exploitation of data producing the necessary change in "culture". Reaching this point will be a step towards removing the curse from the tower, no so as to reach the heavens, but to open new windows of knowledge.

[1] Rechenmann, François; (2000); "Editorial: From data to knowledge"; Bioinformatics; vol.16. num. 5 pag.411

an unexpected place

Vivienne Baillie Gerritsen

New Path, Meg Harper
www.megharper.com

On the occasion of Amos Bairoch's 50th Birthday

Life has its ways. We are given opportunities to make choices. We are even given opportunities to nudge life onto a path we wish. And yet, there seems to be an invisible force lurking beneath which leads you to the most unexpected places…an unexpected place which, in time, turns out to be the place where you should be. Call it destiny, perhaps. Today, fifty years after the day he was born, Amos is sitting in an office in Geneva at the head of a project which has travelled around the world and for which many people work. From a cramped attic to a large open space office, Swiss-Prot continues to grow both in work force and in use. Amos has won prizes for it. He has been praised for it. He has put much of his soul and his heart into it. And despite this, far from him was the desire of ever having really wanted it.

The story has been told many a time. Over twenty years ago, while working on his thesis, Amos felt the need to sort and perfect a protein sequence databank which already existed. And he did. Instead of listing the sequences of proteins, he felt it was of greater benefit to the scientific community to offer some information on the protein as well – such as its structure, its function and its possible involvement in illnesses. Having done so, he offered to those concerned what he saw as improvements made to the existing databank. No particular enthusiasm was shown for his effort, so Amos undertook to develop his concept, fully expecting to hand it over to someone else so he could get on with things he was more interested in, such as exobiology.

But that is not the way the wheel turned. Protein sequences kept on pouring in and, though the rate of submission was far slower than it is today, Amos took it on himself to maintain the rhythm. As a result, he continued not only to enter them manually into his embryo-databank but also to annotate them. There was little point in dropping something that was proving to be useful. That was the beginning of the…beginning. It was not long before he

needed help to cope with the profusion of incoming data. And ever since, the number of scientists who strive to keep pace with automated sequencing and a competitive knowledgebase, has never ceased to increase. Today, Amos is surrounded by computer scientists, biologists of all walks of life, secretaries, receptionists, science writers and science communicators. And a lot of us are sitting in a big modern office in the basement of Geneva's academic medical centre where it all started, whilst others are scattered in other parts of the world.

Hundreds of fingers boogie – or mooch – across keyboards every day, for the sake of proteins. Dozens of pairs of eyes scan computer screens every day, for the sake of proteins. Millions of neurons spark – or indeed fail – every day, for the sake of proteins. Numerous cups of coffee are absorbed, thousands of words are exchanged, hordes of paper are printed, hundreds of footsteps are made, hearts beat, lungs breath, sighs are lost and a few aspirin are swallowed – every day, and all for the sake of proteins.

And, in the hands of experts, these singular molecules of life are given a name, brought down to a sequence of letters, sorted into families, sculpted into space and predicted a role. They are fashioned into ribbons, moulded into globes and painted in the colours of a rainbow. They are explained by professors, sought after by researchers and discovered by students. Pharmaceutical companies commercialise them. Journalists write about them. Artists weave them, sculpt them and paint them. All of us, without exception, eat them. And, what is more, we make them.

Without proteins, life would not exist. And neither would any of us. Without proteins, Swiss-Prot would never have seen the light of day. Neither would Amos for that matter. 50 years ago and a little more, proteins were hard at work in that little spermatozoon which wriggled towards its mate. When it reached its other half, dozens of proteins made quite sure that no other would get a chance. A cell was born. The first. The very beginning of a human being. And from there, with the help of billions of proteins, millions and millions of divisions occurred. Proteins supplied energy and ferried chemicals, they triggered off pathways and monitored rates, transcribed DNA and translated RNA, built networks and provided heat. They let a heart beat, ears hear, eyes see, legs walk and they designed the mould from which a mind could grow. Undoubtedly, there is not much we would be without these remarkable molecules. And destiny? Could destiny also be in the hands of proteins? Hardly. But they certainly do have their say.

Cross-references to Swiss-Prot

Amos is behind each one of them.
The first ever entered: Cytochrome c, then: P00001 and now: P99999

References

1. Dark hair turning grey
   A little scant on the top
   A lagging youth
   And his own office in Geneva

National Nodes

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

Australia
RMC Gunn Building B19, University of Sydney, Sydney

Austria
Vienna Bio Center, University of Vienna, Vienna

Belgium
BEN ULB Campus Plaine CP 257, Brussels

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

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

China
Centre of Bioinformatics, Peking University, Beijing

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

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

Cuba
Centro de Ingeniería Genética y Biotecnolgía, La Habana

Finland
CSC, Espoo

France
ReNaBi, French bioinformatics platforms network

Greece
Biomedical Research Foundation of the Academy of Athens, Athens

Hungary
Agricultural Biotechnology Center, Godollo

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

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

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

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

The Netherlands
Dept. of Genome Informatics, Wageningen UR

Norway
The Norwegian EMBnet Node, The Biotechnology Centre of Oslo

Pakistan
COMSATS Institute of Information Technology, Chak Shahzaad, Islamabad

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

Portugal
Instituto Gulbenkian de Ciencia, Unidade de Bioinformatica, Oeiras

Russia
Biocomputing Group, Belozersky Institute, Moscow

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

South Africa
SANBI, University of the Western Cape, Bellville

Spain
EMBnet/CNB, Centro Nacional de Biotecnología, Madrid

Sri Lanka
Institute of Biochemistry, Molecular Biology and Biotechnology, University of Colombo, Colombo

Sweden
Uppsala Biomedical Centre, Computing Department, Uppsala

Switzerland
Swiss Institute of Bioinformatics, Lausanne

Specialist Nodes

EBI
EBI Embl Outstation, Hinxton, Cambridge, UK

ETI
Amsterdam, The Netherlands

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

IHCP
Institute of Health and Consumer Protection, Ispra. Italy

ILRI/BECA
International Livestock Research Institute, Nairobi, Kenya

LION Bioscience
LION Bioscience AG, Heidelberg, Germany

MIPS
Muenchen, Germany

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

for more information visit our Web site

www.embnet.org

EMBnet.news
ISSN 1023-4144

Dear reader,

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

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

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

Submission deadline for the next issue:
February 20, 2008