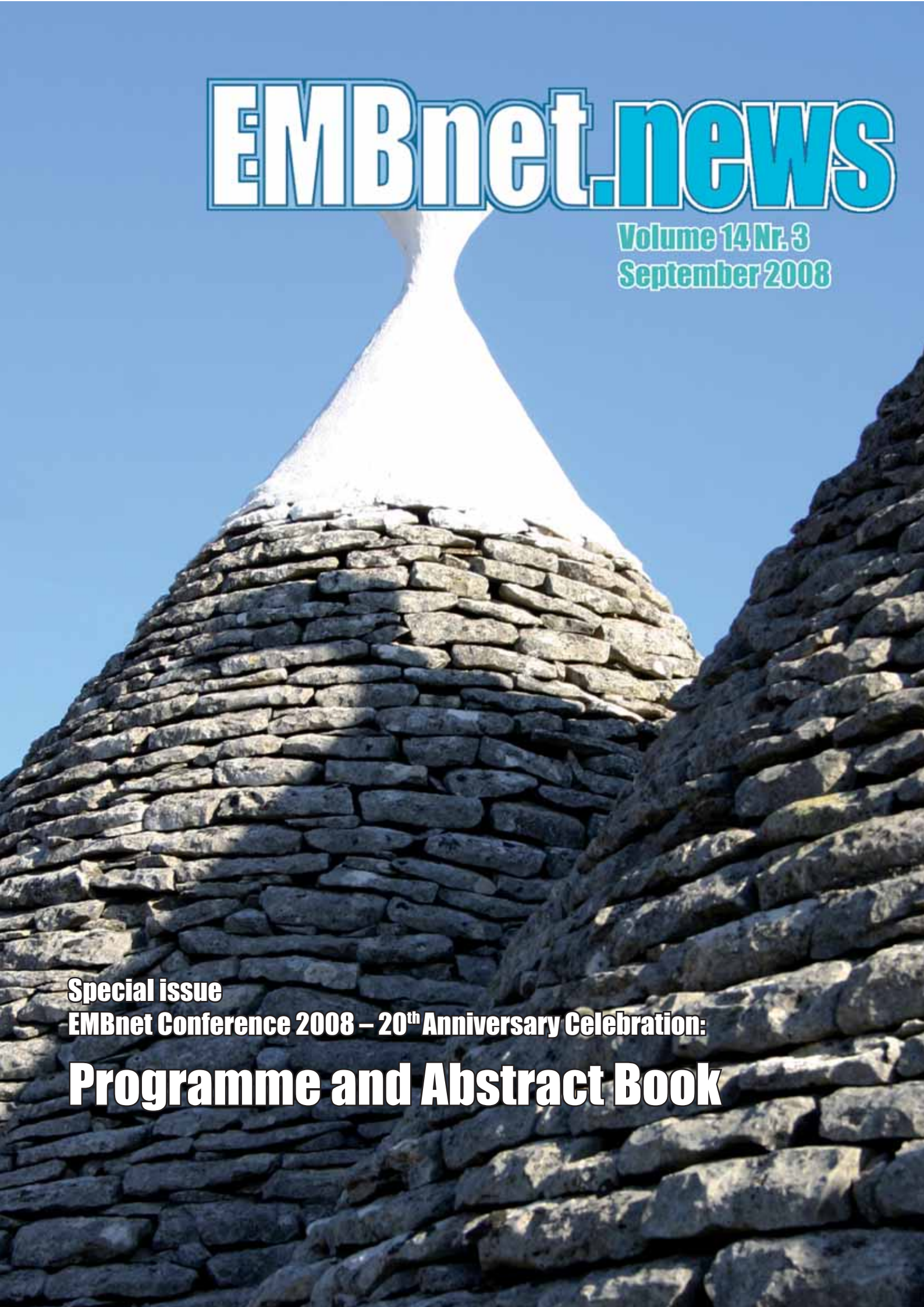
EMBnet.news
Volume 14 Nr. 3
September 2008



Special issue
EMBnet Conference 2008 – 20th Anniversary Celebration:

Programme and Abstract Book

Editorial

The present volume is a special number for several reasons. Our 20th aniversary may be the most important one. But, for our readers it represents a chance of getting the abstracts, the program and the list of participants in the 2008 EMBnet conference, on top of our usual set of articles. In this issue we tried to obtain texts about EMBnet's early steps and its role for so many Bioinformatics users and professionals, with its presence, its spirit as a support provider, and its outreach, mainly obtained with EMBnet.news. An article on EMBnet's e-learning portal brings you up to date with what our community is doing in this area. As usual feel free to comment on EMBnet.news via the portal and send us your contributions. You may like to know that EMBnet.news is very widely spread as an electronic newsletter. The number of downloads is huge and increasing. We hope that you enjoy the reading.

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

Contents

Cover picture: *Trulli,* Martina Franca (Puglia), Italy, 2008 [© Domenica D'Elia]

Nils-Einar Eriksson
Chairman EMBnet technical manager programme committee (TMPC), Swedish national EMBnet node, Biomedical Centre (BMC), Computing Department, Uppsala

Domenica D'Elia
Node manager, Italian national node; Institute for Biomedical Technologies, Bari, Italy

Erik Bongcam-Rudloff
Chairman EMBnet, Swedish national EMBnet node, Department of Animal Breeding and Genetics, SLU, Uppsala

EMBnet, a brief history and a current picture

The 20th anniversary of the European Molecular Biology network, EMBnet, will be celebrated at its Annual General Meeting in Martina Franca, Italy in Sept 18-20, 2008.
During these years a new science, bioinformatics, has evolved and EMBnet has played an important role as a provider of bioinformatics services for a large community of researchers.

The 20th century brought us an enormous growth of knowledge in the field of life sciences. Research on nature's methods to code, store and translate genetic information into biological function resulted in huge amounts of data.

Figure 1. 1993, EMBnet AGM in Basel (CH)

The new data had to be maintained, organised and made accessible to the biomedical research community for analysis. In Europe, this was taken care of by the Data Library group (Graham Cameron, Peter Stoehr and coworkers) of the European Molecular Biology Laboratory (EMBL) at Heidelberg in Germany. EMBL, (www.embl.-org), the flagship of European molecular biology, was established in 1974 and is supported by at present 20 member states and one associate member.

During the 1980-ies it was clear that an almost exponential growth of DNA-sequence data could be anticipated. Plans were evolving for the Human Genome Project (HUGO), i.e. the complete sequencing of the human genome. This project alone, was expected to produce what was, at that time, considered an enormous amount of data requiring tens of gigabytes of data storage.

Organizing, annotating and storing the data were tasks that the Data Library could manage. But when facing the problem of making the data available for analysis to biological scientists in all of Europe, aspects not under control of the biologists had to be considered.

At that time, high-speed data communication across Europe was in its infancy. Scientists wanted to use client workstations with graphical user interfaces that demanded interaction without delays with the server that was providing the data or performing the analysis.
Accessing a remote computer using an ordinary command-line oriented terminal wasn't enough. Obviously, the solution to eliminate communication delays would be to use local computers with local copies of the data. The sheer amount of compute resources needed for European research in this field also pointed to a distributed solu-

Figure 2. 1997, EMBnet AGM in Bari (IT)

tion. Computer cluster technology got widespread acceptance much later.

Thus, an organized way of distributing the data from EMBL to all its member states had to be established. The first practical steps were taken by EMBL in the spring of 1988 to get feedback from scientists around Europe. The concept of a network of national nodes serving each country with up-to-date biological databases and also providing compute resources for data analysis was formulated. It was given the name The European Molecular Biology network, EMBnet.

Some highlights:
In July 1988 the first EMBnet Workshop was organized at EMBL involving participants from EMBL, Daresbury (UK), CITI2 (France), CAOS/CAMM Centre (the Netherlands) and Hoffmann-La Roche.
An early focus was on network protocols for the distribution of data from the EMBL Data Library. At first DECNET was intended as the data carrier but it was soon to be replaced by TCP/IP. A set of client-server data transfer programs, xNDT, was later developed at the Swedish node. Another important issue on the agenda was the preparation of an approach to the European Community to apply for a grant for a pilot EMBnet project.

In November 1988 a letter was sent from the EMBL Director General to all EMBL Council members asking them to stimulate processes in their regions to identify regional EMBnet nodes.

In May 1989 the second EMBnet Workshop was organized at EMBL with representation for all 14 EMBL member states. Established national nodes now included France, Sweden, the UK, the Netherlands, Spain, Israel, Norway, Italy and Denmark. Switzerland, West Germany, Austria, Greece and Finland were gearing up.
In July 1990 the third EMBnet workshop, the first to be organized outside Heidelberg, was held at Uppsala Biomedical Centre (BMC), Sweden. Since then almost all European countries have hosted these annual EMBnet workshops. Quite early the meetings were divided into two parts, an ordinary scientific workshop and a more formal administrative part, the Annual General Meeting (AGM).

In 1991, EMBnet received a grant from the European Community within its framework BRIDGE (Biotechnology Research for Innovation, Development and Growth in Europe 1990-1994). The major objective of the project was essentially the promotion of EMBnet as a European computer network for bioinformatics. The main topics for the development of the network were essentially three: the setting up of a bulletin board, the study and development of the technological tools for data distribution and planning of specialised courses and workshop. A Steering Committee (SC) was nominated during the business meeting held in Nijmegen (NL) in July 1992 whose role was to promote new projects and to stimulate inter-node cooperation.
This and subsequent grants have been important for the success of EMBnet. The initially intended purpose of EMBnet was fulfilled. During its first 6-8 years, the national nodes were the centres where researchers in each European country could access bioinformatics data that were kept in perfect synchrony with

Figure 3. 1999, EMBnet AGM in Brugge (BE)

Figure 4. 2006, EMBnet AGM in Scandinavian countries (Uppsala - Sweden and Helsinki - Finland)

the central data repositories at EMBL and its corresponding agencies NCBI at NIH in the USA and at DDBJ in Japan.
Data communications improved. The Data Library function at EMBL was transferred to its outstation, the European Bioinformatics Institute (EBI) at Hinxton (UK), that was established in 1993.

Some years later, much of the access to bioinformatics data and compute resources didn't necessarily involve the national EMBnet nodes but relied instead often directly on the resources at EBI or on the corresponding US resources at NIH (National Institutes of Health). However, with the advent of new sequencing technologies and massive production of data as well as the development of evermore complex computational and bioinformatics tools to manage, store and analyse biological data, EMBnet's role in teaching, and as a provider of knowledge of tools, solutions and on how to set up and maintain public and local bioinformatics databases and tools, is again in big demand.
Indeed, although old problems seem to be solved, new ones, even more complex, are facing biologists, computer scientists and bioinformaticians in their daily work, increasing the demand for knowledge and sharing of specialist skills.

Since its creation in 1988, EMBnet has evolved from a network of European national nodes in charge of maintaining local biological databases into a worldwide organization bringing bioinformatics professionals to work together to serve the expanding fields of genetics and molecular biology.

Currently, EMBnet offers a panel of experts available to give specialized courses at its nodes. National nodes provide local training and support programmes in local languages.
In order to extend its training offers, EMBnet has recently added to its list of services also a web-based e-learning system (http://edu.embnet.org). The new system is based on the Moodle (http://moodle.org) software with a few plugin extensions and provides facilities to support on-site training as well as a rich remote training facility for on line learning. The e-Learning server is offered as a community service providing training material and experience for end-users, such as bioinformaticians, teachers and life science researchers in general, to use. To facilitate sharing/exchange of teaching material, the e-learning web site also provides an exchange service where the community may share documents and experience in bioinformatics training.

EMBnet nodes also provide their national scientific communities with access to high performance computing resources, specialized databanks and up-to-date software. Some nodes act as redistribution centres to national research institutes. Moreover, collaborative technical expertise within EMBnet provides support for sustaining the biocomputing facilities of the member nodes.

EMBnet produces information and guidance in biocomputing by publishing support

Figure 5. 2007, EMBnet AGM in Torremolinos (ES). Join Meeting EMBnet - RIBIO "Bioinformatics 2007: Workshop on Collaborative Bioinformatics".

material for the end users (Quick guides). The quarterly newsletter "EMBnet.news" presents the latest achievements inside the organization, together with papers on new developments regarding biocomputing. The EMBnet community was involved in the creation of the peer reviewed journal Briefings in Bioinformatics (BiB). BiB was also supported by an educational grant from EMBnet.
EMBnet nodes employ over 150 professionals. Computing resources add up to around 230 hosts ranging from workstation-level computers to supercomputers, and terabytes of storage capacity.

The commitment of EMBnet is to bring the latest software algorithms to the user, free of charge and for this it continues to develop state of the art public software (e.g. EMBOSS/Jemboss, wEMBOSS, Utopia, eBioX, WebLab and many more). Staff from several EMBnet nodes collaborate in developing new biocomputing tools. Many EMBnet nodes were involved in the multimedia educational resource project EMBER and are currently involved in National and European GRID projects such as MyGRID, NorduGRID, LIBI, SWEGRID, SWISSGRID, UPPMAX and EGEE, DEISA, BIOINFOGRID, EELA. EMBnet is a member of HealthGRID.
The recognition of the value of this cooperation is indicated by the fact that new members have been joining from all over the world. A fruitful cooperation with the Iberoamerican (RIBIO http://rib.cecalc.ula.ve/) and the Asia Pacific (APBioNet http://www.apbionet.org/) bioinformatics networks as well as with the US based International Society for Computational Biology (ISCB, http://www.iscb.org) has begun. Close contacts have been established with the African Society for Bioinformatics and Computational Biology (ASBCB, http://www.asbcb.org).
During its 20 years of existence EMBnet thus has grown from initially being a purely European organisation with 14 nodes in 1989, into having 34 national nodes, many of them outside Europe. Almost all continents are represented. There are also eight specialist nodes providing special databases and services. Thus, EMBnet can now be regarded as a truly global organisation.

Here we include only few of hundreds of pictures we took at some old and also at some recent EMBnet meetings, only the most serious pictures obviously.... but please, if you would be interested..... ask an EMBnet member to show you what the collaborative soul of EMBnet is and he/she will not hesitate to tell you about a wonderful adventure and show you an exhilarating picture gallery!!!!

For a list of members and services provided by them, please visit the website at www.embnet.org.

Acknowledgements: The authors are grateful for information provided by Peter Stoehr, EBI, about the early events in the history of EMBnet.

The spirit of EMBnet and its human dimension

On its 20th anniversary, EMBnet sits in my memory, not only as a useful cooperative organization, but also as the center of a rewarding human experience. I have interacted with EMBnet for 18 years in a row. Many changes in the Bioinformatics user community and in Bioinformatics itself were not enough to change its basic spirit. At a time when, once again, its role suffers a necessary shift, its focus and hopefully its spirit doesn't, ad that is what I intend to show you.

In late 1990, one of my personal objectives was to set-up a national Bioinformatics service in Portugal. At that time it seemed quite obvious that doing it on my own, detached from similar efforts that were springing-up everywhere, was not a good idea, to start with.

The European Molecular Biology Network, EMBnet, had recently been formed (in 1988) and coordination meetings were being organized.

Pedro Fernandes
Portuguese EMBnet node manager (since 1991)
Chairman of EMBnet's Publicity and Public Relations Committee
Instituto Gulbenkian de Ciência (IGC)
Oeiras, Portugal

Establishing effective connections to this group was a natural choice.

I have had a chance to get acquainted with EMBnet by attending one of their first formal meetings in Neigmejen, NL. By doing that, I stepped into an operation that would lead me to establish the service that I aimed at, already as a part of an international effort. With this group it was quite easy to learn the do's and don'ts, to make technical options and, in a very effective way, to acquire a good pace with incredibly less effort. In practice this meant that in less than a year the Portuguese node of EMBnet was up.

Technical difficulties, in those times, are difficult to visualize in the present days. Let me just drive you to imagine a world of really slow networking. The Internet was relatively insipient, and not really at reach for interactive operations. Networking was clearly the way to go for all of us so we were using it to step forward. We were using a product from within our small group, Peter Gad's xNDT, to distribute and update databases, an incredible step forward from shipping floppy disks or tapes with the EMBL database using snail mail. Solving this problem was of capital importance. We were all aware of the growth of biological databases and able to anticipate the consequences of some of the foreseeable data deluges (not all, of course!). Much better networking was needed and it was happening.

Distributing data was, as you can imagine, a challenge in itself. As daily updates became possible we could all aim much further. Service provision was still strongly based on the Wisconsin Package (alias GCG) to a very big extent. Topped with EGCG - also springing-up from within EMBnet, it was the bread and butter of every service provision node within the Network.

Aside from these technical issues, the group was assembled around a major collective aim: to establish user communities in each country, around reference points of service provision. This activity led to a remarkable step forward in terms of creating scientific awareness about Bioinformatics itself. For the years to come, the undeniable influence of EMBnet nodes was a result of that spirit. On the other hand, providing similar services in several countries, implicitly eased-up promoted the mobility of scientists with similar needs. This fact has rightfully been used as a live example of what the EC wanted to promote in that field.

Another major difficulty already felt in those days was an educational one. Bioinformatics education was simply not available in most places. Bioinformatics was almost silently taught as a minor theme in only a few Biochemistry courses. Graduate and undergraduate programmes were starting to offer courses, in many cases as part of MSc programmes. EMBnet nodes represented, in many cases, the obvious way out, as they provided flexible training, in a fairly unrestrained way. Many people became aware of the possibilities that were open by Bioinformatics, taken, in general terms, as the in-silico way of solving biological problems using biological information. For many wet Lab scientists, training in Bioinformatics became the way of quickly getting acquainted with the new way of supporting their findings, planning their work and organizing knowledge.

In all these aspects, most of the EMBnet's role has become silently buried in people's memories. The user community has matured and that is what matters most.

The 1990 scenario sounds distant now. We have experienced different technical challenges in the last 20 years. Daily updated databases are now commonplace. So is the availability of Bioinformatics services on the WWW. In the meantime, many other important steps could be given. EMBOSS sprang-up

Figure 1. Cooperation on the edge. Participants in the EMBnet AGM 2002 end a day of meetings at the most western tip of continental Europe, Cabo da Roca, Portugal.

from the EMBnet community and became the “de facto” cornerstone of public domain software for sequence analysis. EMBnet expanded to a worldwide coverage of 50 members, reaching geographical positions that are far from central Europe, to such an extent that it became difficult to count the people that use EMBnet-based services, both in technology and in human presence. Recent estimates quote as over 30,000 the number of users that directly benefit from the activity of our network.

Looking back into our existence it is difficult not to sound nostalgic. However this has never been EMBnet's spirit. EMBnet has adjusted to the different scientific, technological and human scenarios. Young students are still the heavy weight Bioinformatics users all over the world. Their major needs are still difficult to fulfil, as proper education in the area is still not provided to full extent in undergraduate courses. The skills needed to develop proprietary resources are still not easily provided.

EMBnet never lost its target audience. Its focus is and has always been on Bioinformatics users. Its major resource is shared and consists of people and experience. Its major achievement throughout 20 years of operation is user communities, freestanding whenever possible.

The fabulous destiny of EMBnet.news

ISSN1023-4144

By Laurent Falquet (CH)

The little newsletter was born on the first of July 1994, when the world-wide web was just a funny tool used by a few scientists and some geeks around the world. The primary goal of embnet.news is still valid nowadays, it was defined to be a platform of information and reports, hopefully both useful and interesting for researchers using computers in biology.

This first issue was made possible by a charismatic editorial board of well-known node managers, namely Robert Herzog (BE), Rodrigo Lopez (NO), Reinhard Doelz (CH) and last but not least Alan Bleasby (UK). This team was joined soon after the first issue by Andrew Lloyd (IE) and after the issue 3.2 by Peter Rice (UK). The mighty six kept the flame burning until 2002.
Robert managed the edition using his PC on Windows 3.11 and Aldus PageMaker (the ancestor of InDesign was acquired by Adobe in 1994, see http://en.wikipedia.org/wiki/Adobe_PageMaker). He produced HTML and PostScript files that were printed later on by each node or sent by mail to various subscribers. The files were available from various web and ftp sites of the EMBnet community. This was a very smart move that proved helpful a few years later to recollect lost past issues.
Various successful series of articles were appreciated by the readers, e.g., BITS (BioInformatics Theory Section) and INTERviewNET, that contributed a lot to the popularity of the newsletter.
However the good days were gone, in 1999 the crisis began together with EMBnet funding issues. Until 2002 only one issue per volume was published and even a complete absence of any issue in 2001. Their last issue 8.1 in September 2002 was coincidentally the first in color.

Laurent Falquet
Secretary of EMBnet Executive Board and Swiss EMBnet Node Manager
Swiss Institute of Bioinformatics (SIB)
Génopode-UNIL
Lausanne
Switzerland

FIRST HEADER

SECOND HEADER

TIRTH HEADER

Figure 1. The figure reports the 3 historical headers of the EMBnet.news

In October 2002 when I was elected as a new member of the P&PR PC I decided to concentrate my efforts to revive the glorious newsletter. Seconded by the chairman Erik Bongcam-Rudloff (SE) and the members of the committee Pedro Fernandes (PT), Goncalo Guimaraes (BR) and Oscar Grau (AR).
I bought Adobe InDesign CS and used my brave old PowerMac G4 (867Mhz, 256Mb RAM) to generate my first issue 9.1 in April 2003.
The start was difficult, since we had to reinvent the design and get people to write articles... We decided that only PDF versions would be produced, because keeping the design in HTML was too much work and at that time reading PDF was already very common.
We also decided to collaborate with the Swiss-Prot team and publish jointly their "Protein Spotlight" (ISSN1424-4721 http://www.proteins potlight.org/) monthly review of proteins within our newsletter. This is still a remarkable added value to our publication. Thank you Swiss-Prot!

After 3 years of "Swiss timing" publication every 3 months that were downloaded several thousand times from our web site, I handed over to the current P&PR PC committee at the beginning of 2006.

The current editorial board is led by the enthusiastic Domenica D'Elia (IT) and seconded by Lubos Klucar (SK), Pedro Fernandes (PT) and Kimmo Mattila (FI). Domenica collects the articles and pushes the authors to respect deadlines (very efficiently!), while Lubos edits the journal with InDesign CS2 on his WinXP PC (2.8Ghz, 1Gb RAM). The management and commenting of the documents is carried out using the forum facilities of the www.embnet.org server. After a short adaptation period at the beginning, they maintained the publication alive, improved it with excellent articles, proposed a new way to read the newsletters via a free Flash application (http://www.issuu .com) and even produced a beautiful special printed issue in 2007 reiterating the experience with this special issue for the 20th Anniversary of EMBnet. They are doing a wonderful job to perpetuate the tradition of our newsletter.

How to continue? Being more professional? Being referred by PubMed? Of course a full-time editor and peer-reviewed article would be great, we would need to find funding, for example via advertisement or via a grant. But keeping the spirit is also essential! How about publishing your article in the next issue? Please send it to the editorial board... now! ;-)

In my name and in the name of the EMBnet community, I would like to warmly thank all the present and past members of the editorial board for the huge amount of excellent articles, reports, pictures, tutorials etc., as well as, all the authors of the EMBnet nodes, and outside of it, who contributed to make this newsletter a "must read" journal in the community of bioinformatics users and developers!

The EMBnet e-learning server

By José R. Valverde
On behalf of the European Molecular Biology Network, EMBnet.

Abstract
EMBnet has added to its list of public services a general purpose, web-based, e-learning system to promote training in Bioinformatics. This system builds on and expands previous work on Bioinformatics e-learning (EMBER) and is based on Moodle. We provide facilities to support on-site training as well as a rich remote training environment for on line learning. Besides its use in EMBnet training activities, the server is offered as a service to the broad Bioinformatics Community: it can host third party courses and provides a neutral, public repository for the exchange of quality, unencumbered training materials, methods and experiences that can be freely used by other teachers in building their own courses and materials.

José R. Valverde
Spanish national EMBnet node/Centro Nacional de Biotecnología (CSIC)
Campus Univ. Autónoma
Madrid, Spain

Introduction

EMBnet mission is to promote the development and adoption of Bioinformatics through a network of cooperating nodes distributed worldwide. To accomplish our mission, we must ensure the technology is widely adopted and well understood, and o this end we conduct a large number of courses at the various nodes (see EMBnet web site [1]) for announcements).

Training in Bioinformatics is an evolving task, just like the underlying discipline demands. This means that there is a great deal of work involved in preparing and updating training materials, organizing courses and doing the formation itself. In our case the tasks are further complicated by the fact that we need to reach a wide audience, often providing training in local languages - not just English- and in many cases must address students in developing countries, where traveling costs are expensive and economic issues are a serious concern.

In order to simplify course management we have decided to implement a facilitating technology to aid us perform all the tasks involved. Building on our previous experience doing on site and on line training (EMBER project [2][3] and many courses) we already had some expectations about what we wanted to achieve. Most significantly were aware of the difficulties to find good reference material free of encumbrances that can be freely used to compose new courses and accessed by students.

As a result of our previous analysis we started out to seek a tool that could satisfy our needs, providing a rich, efficient environment for e-learning that could be shared and spread to many sites. We also wanted to provide a reference resource for other professionals in the field to use as a starting point for their own teaching activities, and thus needed that our solution be open, compatible and accessible enough for anybody to replicate and use. We have achieved our goal by conducting a detailed analysis of existing tools, selecting the one we consider most appropriate for our ends. This resulted in our installation of a Moodle [4] server that has been in production for well over two years now. In this time we have had time to prove the technology and start exploiting it to provide on site and on line courses, and to test and develop novel approaches to teaching. We have also set up an exchange course repository and started liaisons with major Bioinformatics networks to cooperate on the initiative.

Materials and Methods

Choice of server software

Before deciding on a base software to use for the site, we went on a research expedition to explore the different choices available, including both commercial and free software e-learning environments.

We started by reviewing the documentation and user comments available for most existing products, and then selecting those systems holding more promise to fulfill our perceived needs. We then proceeded to test as many systems as possible on institutions with existing production installations and demo sites (among them some public sites like NCSA webCT based training on HPC [5] or MIT OpenCourseWare site [6] and some private sites to which we had access) and to gather feedback on their experience using them. From this screening we selected a smaller number of systems for further analysis.

After the above screenings we decided to test first hand a selected number of choices, among them .LRN [7][8], Atutor [9][10], Bazaar [11], BSCW [12], Claroline [13][14], Dokeos [15], Fle3 [16][17], ILIAS [18] and Moodle [19][20], installing them at the EMBnet project incubator site [21], and comparing them in detail for a number of features:

- ease of installation
- ease of use
- intuitiveness
- features for cursor authoring
- features for student training
- features for course management
- portability
- popularity
- support of standards
- price/quality ratio

Special attention was payed to openness and shareability: to ensure the widest impact we needed a completely free, easy to install, infrastructure from the bottom-up, so that it would be easily replicable by even at the most reduced, budget constrained sites (specially so since they will be major beneficiaries of this initiative).

Support for on-site training

Our initial work using the server has been directed to date to supporting on-site training as this is the main training activity within EMBnet to date.

We have started by simply converting existing materials to a basic e-learning format, often just importing them directly into the site. This is an easy task that is well documented elsewhere [19][22].

Support for on-line training

As EMBnet has a widely geographically distributed infrastructure it is natural to provide on-line learning to save students travel and tuition costs.

The EMBnet site provides a

rich set of features for on-line training, intrinsic to the software selected. Getting acquainted with these tools has a steeper learning curve but pays well off at various levels. It also requires major changes in the way to present and use the materials as has been widely discussed.
Our server provides for a wealth of activities to monitor and promote user advances in their learning experience: starting from a calendar to remind students of deadlines, one may add forums for public discussions, workshops and seminars, assignments, quizzes journals, wiki, etc..

EMBnet eLearning is a public community service
EMBnet is a non-profit organization whose aim is to promote and support Bioinformatics in the Scientific community. Training is just one of its means, but it would be insensate to ignore that we cannot reach the whole community alone. It does indeed make much more sense to rely as much as possible on cooperation with other institutions to further our goal.
We have provided a public repository for the exchange of training materials within the Bioinformatics community. Our goal is to provide a site where it is easy to discover what is already available and under which usage conditions, so that any trainer can locate useful materials and use them freely to enhance their courses. It also allows trainers to publish their works so others can user them under the same terms. Ideally we would like to see this site grow to become a central reference for training in Bioinformatics.

RESULTS:
EMBnet e-learning server is based on Moodle.
Many systems could be discarded after reviewing their documentation as they provided limited functionality, reduced support for standrds or missed some relevant feature. Then we tested the most relevant options at installed sites and under close control on local installations. We won't go into much detail here as there are far better reviews already available on the Net (see e. g. the Edu-CMS comparison table [23]).
Some of the systems soon distanced themselves ahead like Atutor or Moodle: in the end we have selected Moodle for a number of reasons that will be discussed later, and we feel happy that our selection has been backed by its wide adoption in the academic community as well.

On site training
After the decision was made, we started to populate a test server with training materials and to use it for driving test courses, and once we felt confident enough on our choice moved it to a production server (http://edu.embnet.org/) on January 2007. We have been running the main server for two years and using it to assist our training and gain further experience. In the meantime we have also populated the site with existing materials gathered from EMBnet nodes and other sites on the Net. In general, user satisfaction has been high whenever students were exposed to the technology, even if it was only as an additional extension to traditional on site courses.
In most cases we have just uploaded new or existing presentation materials and exercise texts to the server. Often only slide presentations would be made available, but in our experience it has proved more than enough for on-site training.
As our experience grows we are expanding contents with textual materials and automated quizzes. It is worth noting that once an initial set of materials is available, it is easy to work on it and expand it to enrich the students' experience: the simplest and yet most useful additions are
- using the built-in calendar to make announcements to students
- using a forum to answer students questions in a public way
- adding an electronic submission system to handle course assignments
- automating satisfaction surveys with predefined standard electronic surveys
- adding additional materials or links to other web pages

All the above additions require little effort and can change a classical on-site course into a more rewarding and efficient experience, providing significant changes in managing and supporting students' work. This point is the most interesting indeed: it remarks that one can start by a plain conversion of a traditional course (simply uploading existing materials) and that this may be more than enough for a start, but it also implies that once these starting materials are in place, adding more features is trivial and rewarding.
The next logical step has been extending support for on-site courses until the assignments due date. This allows us to provide more complex assignments to students and give them longer times to prepare and submit their works, a valuable asset when most of the attendants to courses (as is often the case within EMBnet) come from other institutions and must travel to attend the course. Simply leaving the course open and a discussion forum allows us to continue solving students' doubts after they return home and while they work on their assignments.

Support for on-line training

The EMBnet e-learning site is an open Bioinformatics training facility. While we are still in our early stages, mainly using it to support on site courses, we have already started to conduct some fully on line courses which to date have been open to any scientist or student in the world. In the future we will consider opening more specific courses, perhaps oriented to specific technical, regional or linguistical communities.

Joining the site is totally free and gives access to all the materials stored in the site. Members can participate in forums, use the existing materials (according to their licenses), and join discussions or contribute materials to the exchange facilities.

In general we have received good and encouraging feedback on the use of Moodle for training. Our main concerns are now to better exploit its facilities to provide an improved learning environment.

On line-training requires additional work: it is no longer enough to provide slide presentations, as their content is necessarily too succinct and terse, instead we must complement the course with complete materials substituting for traditional oral lectures.

Most similar to traditional lectures is to record actual lectures as streaming videos and provide them over the web. This had already been done at the Finnish EMBnet node [24], but general experience has shown that attention levels decay quickly when students follow an oral presentation in this way.

This led us to other approaches. The simplest one is to define a reference book for reading. A better approach is to include the materials in the course, possibly tailored to the course goals. Adding reading materials can be as simple as uploading text documents, images or web pages; this helps students reduce costs, but requires some control of student progress by the way of deadlines, activities and assignments.

In our experience, this is a lengthy process requiring time and work from teachers. Both, lecture preparation and student tutoring are common tasks of on line and on site courses, however, preparing the contents of an on line course takes a substantially greater amount of time. Most of the extra work deals with preparing materials (text, illustrations, examples, etc...) to substitute common lecture activities (like drawing schemes on a blackboard or explanations). On the bonus side, once done, it can be reused in all instances of the course (while live lectures must be repeated).

The preparative work can be reduced substantially if one can find existing materials that cover the subject topics of the course, possibly resorting to existing repositories [6][25][26]. Still, there is time involved in finding the materials, verifying their suitability, checking copyright restrictions and possibly adapting them to one's needs. In several cases we have found it easier to modify the course to fit existing materials than to adapt materials to our initially intended content. This saves time and is not so different from traditional textbook based teaching. Then, once the course has been delivered one can use the preexisting materials to progressively adapt them for the next instance of the course.

Course exchange

Our experience using existing materials has proved this to be a most useful approach, but there is a general lack of available materials that can safely be used in on line courses (mostly due to the fact that most of them - though possibly public- lack any copyright notice and contacting the original author is often a difficult task).

Given our experience in preparing courses and EMBnet mission, we believe that we can better serve the community by providing quality materials for learning that can be used by trainers at any institution. In the same sense we do greatly benefit from materials developed by others for their own purposes and made available to the community. In doing so, we have noticed a need for a repository site where teachers could locate quality training materials that can be freely used to enhance their courses, or where they could publish their works for others to locate and use.

Consequently, we have decided to make all contents in our server available *by default* under a Creative Commons [27] license. Authors may decide to use more or less restrictive licenses for their works if they so wish, but this provides a safe default reference backed by policy recommendations to authors [22]. Most of EMBnet courses and their contents can be freely downloaded and reused by other sites.

While we are offering EMBnet materials and courses publicly, we are positive that our feeble forces alone can not provide for all the training needs of modern Bioinformatics. For this reason we have created a special section on the server devoted to the public exchange of course and training materials and to the discussion of training experiences in e-Learning [28]. This section is open to anybody interested

and offers the possibility of uploading information and contents, stating license terms, and discussing experiences on a public forum. We intend that with time it will become a reference hub for training expertise in Bioinformatics.

The EMBnet site is designed to accommodate training in most languages. It has multilingual support for user interaction as well as providing sections for training in different languages so students and teachers can find more easily materials in their language of choice.

In order to further disseminate the initiative, we have contacted other networks to launch collaborative training initiatives: the Iberoamerican Bioinformatics Network (RIB) [29], the Asia Pacific Bioinformatics Network (APBionet) [30] and various National Bioinformatics Networks. While we are still in early cooperation stages, we expect these initiatives to finally result in enhanced training facilities world wide.

DISCUSSION

Software selection

After applying our selection procedures we were left mainly with Atutor and Moodle: both are good solutions for a free, open e-learning system (actually NCSA has switched from its previous use of webCT to Atutor [31] lately).

We finally have adopted Moodle as the basis for our server: we consider that although Atutor has better support for web standards and accessibility, Moodle has a richer feature set, is more easy to install and use, has better support for standards like SCORM and other popular teaching technologies (like Flash, Hot Potatoes, etc..), provides more plugins and has useful extensions for Bioinformatics training (e.g. Jmol plugins [32], Java molecule editor [33]). Moodle's only requirements are PHP and SQL, and its XML format and support for standards greatly favor exchange of materials. It is noteworthy that Moodle has been adopted as their solution of choice by a large number of academic institutions; Moodle also has a huge community of users and sites worldwide, possibly the largest in existence. As a matter of fact, ACM's eLearn magazine ranking of the top 100 tools for eLearning in 2007, ranks Moodle 12 on the list (after generic tools like Firefox, Google or Skype), stating "The popularity of this open-source course management system is impressive, and it is miles ahead of its nearest academic competitors-and there's not a commercial learning management system on the list." [34].

Supporting E-Learning in Bioinformatics

Extending support for students after the on-site lectures is already a great step forward when most students come from external institutions. It may make a small difference for centers teaching only local students, but it allowed us to offer visiting students a more natural and lasting experience, and to accommodate to their different learning paces.

Preparing full on line courses is more demanding. Use of reference books for e-learning is tempting as it gives some assurance students will get guaranteed quality contents (provided the books are well chosen and students can get them). Book accessibility is a major issue for us: while it is feasible today to get almost any book using online shops like Amazon [35], the sad truth is that for most population on Earth it is close to impossible to access them due to economic costs. Even though there are free e-book services for students (e.g. [36][37]) they are little known and have minimal support for Bioinformatics topics. Since EMBnet spans many developing countries, we favor inclusion of unencumbered on line materials in the course themselves as a better approach.

There is an additional reason to prefer on line materials: no matter how good books may be, they are solidly "cast in stone" under the tight control of the book authors or editors. One may not tailor them easily to a course, needing to ignore some parts or recourse to additional materials for specific topics. On the other hand, an electronic resource can be modified to evolve as needs change (e.g. [38][30], [31]); it is easy to add on, mix and match from different sources, delete sections, modify, etc... Truly, there are still copyright restrictions that need to be addressed, but provided they are resolved, the plasticity of electronic materials can not be matched by conventional books.

When building a course on the EMBnet site, one may use a plethora of resources. Probably the most difficult task is exercising discretion on which features are or not worth using given the wide potential audience. Most relevant, people in poorly communicated areas may have serious trouble accessing rich multimedia contents. Our advice in general is that main content be made primarily available as plain text as the less common denominator, and then enriched with extra contents as desired. This ensures that at the very least all students will be able to access the basic material, while allowing better communicated students to benefit

of all the features of the system.

There already exist various repositories of course materials: one of the most well known is the MIT OCW site [6], containing Creative Commons [27] licensed materials on a wide range of disciplines. The MIT OCW site is limited to MIT courses (as a guarantee of their quality) and SCORM formatted content, and its multidisciplinary scope means it has a reduced amount Bioinformatics related contents.

While preparing a course on Biostatistics recently [39], we became aware of another huge repository, Supercourse [25]. Supercourse is devoted to Epidemiology and while it is open to contributions by anybody and has a good coverage of Biostatistics and epidemiology related practices, its falls short for Bioinformatics. It also relies on a specific format: Hypertext Comic Book [40], which, although an interesting extension to traditional presentations is to our taste too limited for modern rich environments. One nice bonus of Supercourse is its multilingual support, providing translations of contents to various languages.

In contrast, the EMBnet exchange repository is open to all interested parties, it is not tied to any specific format (we welcome indeed any kind of educative material), has a well defined scope (training in Bioinformatics) and provides support for open discussions, exchange of experiences and multilingual support.

We have started cooperations with RIB and APBionet to collaborate on the development of common training infrastructures and look forward to fruitful results. Of course the initiative is -as already stated- open to everybody interested and we will certainly welcome cooperation and contributions from any parties sharing our common interest.

Citations

1. http://www.embnet.org
2. Mabey, J. E. and Attwood, T. K., EMBER: a European Multimedia Bioinformatics Educational Resource. CAL-laborate, volume 6, June 2001
3. Attwood, TK, Selimas, I, Buis, R, Altenburg, R, Herzog, R, Ledent, V, Ghita, V, Fernandes, P, Marques, I and Brugman, M, Report on EMBER project – A European Multimedia Bioinformatics Educational Resource, BEE-j, vol. 6, November 2005.
4. http://moodle.org
5. http://webct.ncsa.uiuc.edu:8900/
6. http://web.mit.edu/ocw/
7. http://dotlrn.org
8. Computers and Education E-learning, from Theory to Practice: E-learning, from Theory to Practice. By Baltasar Fernández Manjón, Inc NetLibrary, SpringerLink (Online service Published by Springer, 2007. ISBN 1402049145, 9781402049149
9. http://www.atutor.ca
10. Gay, G. Atutor: Adaptive learning online. IEEE Learning Technol. Newslett, 2002
11. http://bazaar.athabascau.ca
12. http://www.bscw.de
13. http://www.claroline.net
14. Claroline, une plate-forme d'enseignement/apprentissage sur Internet. Pour propulser la pédagogie active et l'innovation ? Docq, F., Lebrun, M., Smidts, D. In Frenay, M., Raucent, B. & Wouters. P. (Eds.), Actes du quatrième colloque "Questions de pédagogies dans l'enseignement supérieur" (pp. 99-109). Louvain-la-Neuve : Presses universitaires de Louvain. 2007
15. http://www.dokeos.com
16. http://fle3.uiah.fi/
17. Teemu Leinonen, Giedre Kligyte, Tarmo Toikkanen, Janne Pietarila, Philip Dean (2003). Learning with Collaborative Software - A guide to Fle3. Helsinki, Taideteollinen korkeakoulu 2003. ISBN: 951-558-127-3.
18. http://www.ilias.deapbio
19. http://moodle.org
20. Martin Dougiamas, Peter C. Taylor, "Moodle: Using Learning Communities to Create an Open Source Course Management System", ED-MEDIA 2003: World Conference on Educational Multimedia Hypermedia & Telecommunications , Honolulu Hawaii USA 2003, http://dougiamas.com/writing/edmedia2003/ .
21. http://bioportal.cnb.uam.es/embnet/
22. Valverde, J. R. (2007) Some key issues of electronic publishing in e-Learning platform. embnet.news, vol. 13, no. 1, 23-30
23. http://www.edutools.info/item_list.jsp?pj=8
24. http://www.csc.fi/english/research/sciences/bioscience/Courses_and_events/recorded
25. http://www.pitt.edu/~super1/
26. Laporte RE, Omenn GS, Serageldin I, Cerf VG, Linkov F. A Scientific Supercourse Science. 2006 Apr 28;312(5773):526.
27. http://www.creativecommons.org
28. http://elearning.embnet.org/course/view.php?id=40
29. http://rib.cecalc.ula.ve
30. http://www.apbionet.org
31. http://www.ncsa.uiuc.edu/UserInfo/Training/CI-Tutor/index.html
32. http://moodle.org/mod/data/view.php?d=13&rid=88
33. http://moodle.org/mod/data/view.php?d=13&rid=296
34. http://www.elearnmag.org/subpage.cfm?section=articles&article=56-1]
35. www.amazon.com
36. http://www.computer-books.us/
37. http://programmingebooks.tk/
38. http://www.embracegrid.info/
39. Valverde, J. R. (2007) IBS-ES-07 course: The making of. embnet.news, vol 13, no. 4. 11-17
40. Global Health Network Contributors. The Reincarnation of Biomedical Journals as Hypertext Comic Books. Nature Medicine - Vol. 4 (1998)

CONFERENCE PROGRAMME

EMBnet Conference 2008
20th Anniversary Celebration
Leading Applications and Technologies in Bioinformatics

September 18-20, 2008
Martina Franca (Taranto), Italy
Park Hotel San Michele

Conference Chair

Domenica D'Elia

Conference Co-Chair

Andreas Gisel

Organising Committee

Cesar	Bonavides-Martinez	Mexico
Erik	Bongcam-Rudloff	Sweden
Domenica	D'Elia,	Italy
Nils	Einar-Eriksson	Sweden
Laurent	Falquet	Switzerland
Pedro	Fernandes	Portugal
Andreas	Gisel	Italy
Jack A.M.	Leunissen	the Netherlands
Sandor	Pongor	Italy
Federico	Ruggieri	Italy

Scientific Committee

Teresa	Attwood	United Kingdom
Emiliano	Barreto	Colombia
Endre	Barta	Hungary
Erik	Bongcam-Rudloff	Sweden
Ricardo	Bringas Perez	Cuba
Shahid	Chohan	Pakistan
Domenica	D'Elia	Italy
Laurent	Falquet	Switzerland
Pedro	Fernandes	Portugal
Andreas	Gisel	Italy
Vassilios	Ioannidis	Switzerland
Sophia	Kossida	Greece
Jingchu	Luo	China
George	Magklaras	Norway
Allan	Orozco	Costa Rica
Guy	Perriere	France
Graziano	Pesole	Italy
Sandor	Pongor	Italy
Federico	Ruggeri	Italy
Cecilia	Saccone	Italy
Piotr	Zielenkiwicz	Poland

Organising Secretariat

Francesca Mariani
EEM Congressi & Eventi
Via E. Lampridio Cerva, 167 - 00143 Rome (Italy)
f.mariani@eemservices.com

Dear Participants,

on behalf of the EMBnet and of the Scientific and Organizing Committees, we have the pleasure to welcome you to the "EMBnet Conference 2008: Leading Application and Technologies in Bioinformatics" celebrating the 20th anniversary of EMBnet.

The conference brings together bioinformaticians and biologists from all over the world to present and discuss general themes such as biodiversity, 'omics', advanced technologies and e-learning in the field of bioinformatics.

Research work presented by top international scientists will give this conference the right character to celebrate such an important event. In particular the conference will cover topics such as genomics, proteomics, transcriptomics, systems biology, metagenomics, and new technologies for high-throughput sequencing including debates on how to manage the data flow produced by such technologies as well as to find new ways to organize and analyze large sequence data sets.

Other technological fields covered will be data- and text-mining, ontologies and Grid technologies since these are technologies bioinformatics will be more and more dependent on. To demonstrate the potentiality, but also the possibilities that the Grid technology can provide to bioinformatics, we offer a tutorial on "Grid Computing" in collaboration with partners of the Italian project LIBI.

We wish you a very intense, interesting and productive stay in Puglia and hope you will enjoy Italian cuisine and will be able to gather also some cultural and panoramic snapshots of Puglia.

Domenica D'Elia and Andreas Gisel

This Conference is under the auspices
of the President of the Regional Council of Puglia (IT)
and the University of Bari.

Conference Speakers:

Alina	Agramonte
Marcella	Attimonelli
Erik	Bongcam-Rudloff
Vincent	Breton
Marcello	Castellano
Ana	Conesa
Domenica	D'Elia
Patrice	Duroux
Alexandru	Floares
Andreas	Gisel
Mahnaz	Habibi
Mehrdad	Hajibabaei
Pascal	Hingamp
Richard	Kamuzinzi
Eija	Korpelainen
Alexander E.	Kel
Erik	Lagercrantz
Giuseppe	La Rocca
Jerome	Lane
Olivier	Lespinet
Alvaro	Martinez Barrio
Vincent	Miele
Magali	Naville
Goran	Neshich
Guy	Perriere
Steve	Pettifer
Viviana	Piccolo
Fabio	Polticelli
Teresa	Regina
Guillaume	Rizk
Kristian	Rother
Maria	Roubelakis
Cecilia	Saccone
J. Cristian	Salgano
Chris	Sander
Monica	Santamaria
Indra Neil	Sarkar
Jamie	Shiers
Helena	Strombergsson
Tan	Tin Wee

Tutorial Speakers:

Giulia	De Sario
Giacinto	Donvito
Francesco	Falciano
Sandro	Fiore
Giuseppe	La Rocca
Maria	Mirto
Graziano	Pappadà
Gaetano	Scioscia
Angelica	Tulipano
Josè R.	Valverde

CONFERENCE PROGRAMME

Wednesday, September 17, 2008

10:00 -18:00	Registration
10:00 -18:00	EMBnet Annual General Meeting
	Tutorial on Grid Computing
09:30-13:00	Morning session
09:30-10:00	The Grid in practice *Josè R. Valverde (Centro Nacional de Biotecnología (CSIC) - Madrid - Spain)*
10:00-10:30	Authenticated Grid access with robot certificates *Giuseppe La Rocca (INFN Catania)*
10:30-11:00	The Grid Problem Solving Environment for Bioinformatics: the LIBI experience *Maria Mirto (Università del Salento - Lecce) , Italo Epicoco (Università del Salento - Lecce)*
11:00-11:30	*Coffee break*
11:30-12:00	GRelC an easy way to manage Grid database *Sandro Fiore (Università del Salento - Lecce)*
12:00-12:30	Federated database *Gaetano Scioscia (IBM Italy S.p.A - Innovation Center - Bari)*
12:30-13:00	Federated database: data retrieval tool *Graziano Pappadà (Exhicon s.r.l. - Bari)*
13:00-14:30	*Lunch*
14:30-17:50	Afternoon session
14:30-15:00	The GENIUS Grid Portal, an INFN portal to access the EGEE grid infrastructure (application examples) *Giuseppe La Rocca (INFN Catania)*
15:00-15:30	The AntiHunter application *Francesco Falciano (CINECA - Bologna)*
15:30-15:50	The Job Submission tool (JST) *Giacinto Donvito (INFN - Bari)*
15:50-16:20	The CST Miner application *Giacinto Donvito (INFN-BARI)*

16:20-16:50 *Coffee break*

16:50-17:20 The Gene Analogue Finder application
Giulia De Sario (ITB-Bari)

17:20-17:50 Microarray Clusters validation with resampling techniques
Angelica Tulipano (ITB-Bari)

Thursday, September 18, 2008

10.00-18.00 Registration

10.00-11.00 Opening Cerimony
EMBnet 20th Anniversary Celebration
Opening remarks: EMBnet yesterday, today, tomorrow
Domenica D'Elia, Erik Bongcam-Rudloff, Cecilia Saccone, Chris Sander

11:00-11.30 *Coffee Break*

SESSION 1: Bioinformatics for Biodiversity
Chair: Cecilia Saccone

Keynote lecture
11:30-12:10 Biodiversity Informatics: Enabling a Macroscopic View of Biology
Indra Neil Sarkar, MBL, Massachussetts (USA)

Keynote lecture
12:10-12:50 The Barcode of Life: Bringing Genomics to Biodiversity
Mehrdad Hajibabaei, Canadian Centre for DNA Barcoding, University of Guelph (CA)

12:50-14:30 *Lunch*

14:30-14:50 Towards barcode markers in Fungi: an intron map of Ascomycota mitochondria
Monica Santamaria, ITB - CNR (IT)

14:50-15:10 Gains and losses of lineage-specific group II introns in mitochondria of Gymnosperms: Molecular Evolutionary And Phylogenetic Implications
Teresa M.R. Regina, Università della Calabria (IT)

SESSION 2: Training and E-Learning
Chair: Josè R. Valverde

Keynote lecture

15:10-15:50 Policies, Network, Resources, Materials and Curricula for advancing bioinformatics education: 10 years of APBioNet
Tin Wee Tan, YLL School of Medicine, National University of Singapore (SG)

15:50-16:10 Metagenome annotation: an opportunity for undergraduate bioinformatics teaching
Pascal Hingamp, Mediterranean University (FR)

16:0-16:30 Grid-based business-to-academia collaborations
Richard Kamunzinzi, Université Libre de Bruxelles (BE)

16:30-17:00 *Coffee break*

17:00-17:20 Sprints at genesilico - software engineering techniques in a bioinformatics lab
Kristian Rother, International Institute of Molecular Cell Biology (PL)

17:20-17:40 Towards semantic interoperability of bioinformatics tools and biological databases
Steve Pettifer, Manchester University (UK)

Sponsor talk

17:40-18:10 IMGT®, an ontology and a system that bridge the gap between sequences and 3D structures
Patrice Duroux, IMGT®, Institut de Génétique Humaine, CNRS (FR)

18:10-20:30 POSTER SESSION

20:30 *Welcome Party*

Friday, September 19, 2008

SESSION 3: "Omics", comparative studies and evolution
Chair: Graziano Pesole

Keynote lecture

09:30-10:10 Evolution of gene regulatory code
Alexander E. Kel, BIOBASE GmbH (DE)

10:10-10:30 A greater diversity of riboswitches identified through the presence of alternative structures and other constraints
Magali Naville, Université Paris-Sud (FR)

10:30-10:50 IMGT/LIGMOTIF: a tool for immunoglobulin and T cell receptor gene identification and description in large genomic sequences
Jérome Lane, Université Montpellier 2 (FR)

10.50-11:20 *Coffee Break &* POSTER SESSION

11:20-11:40 Functional assessment of time course microarray data
Ana Conesa, Centro de Investigacion Principe Felipe (ES)

11:40-12:00 CHIPSTER - user friendly analysis software for DNA microarray data
Eija Korpelainen, CSC (FI)

12:00-12:20 Characterization and analysis of the expression pattern of microRNAs in the grapevine Vitis vinifera
Viviana Piccolo, University of Milan (IT)

12:20-12:40 Improving the prediction of protein behaviour in hydrophobic interaction chromatography and aqueous two-phase systems with clustering methods
Cristian J. Salgado, University of Chile (CL)

12:40-13:00 Contact coordination patterns and electrostatic potential at alpha carbon atoms: a dossier of protein secondary structure elements
Goran *Neshich, Embrapa Informatica Agropecuaria (BR)*

13:00-14:30 *Lunch*

Keynote lecture

14:30-15:10 Systems biology of cancer pathways
Chris Sander, Memorial Sloan-Kettering Cancer Center, New York (USA)

15:10-15:30 Deciphering the connectivity structure of biological networks using MIXNET
Vincent Miele, CNRS, Lyon (FR)

15:30-15:50 Automatic inferring drug gene regulatory networks using computational intelligences tools
Alexandru Floares, SAIA (RO)

15:50-16:10 A chemogenomics view of protein-ligand spaces
Helena Strömbergsson, Uppsala University (SE)

16:10-16:40 *Coffee Break &* POSTER SESSION

16:40-17:00 In silico prediction of escape mutants of the HIV-1 protease
Alina Agramonte, University of Informatic Sciences (CU)

17:00-17:20 The RHNUMTS compilation
Marcella Attimonelli, University of Bari (IT)

17:20-17:40 Homologous gene families databases for comparative genomics
Guy Perriere, University of Lyon (FR)

SESSION 4: Advanced Bioinformatics Technologies and Applications
Chair: Gisel Andreas

Keynote lecture
17:40-18:20 Grids for Life Sciences: status and perspectives
Vincent Breton, CNRS-IN2P3, Université Blaise Pascal (FR)

Keynote lecture
18:20-19:00 Data Challenges in the Worldwide LHC Computing Grid (WLCG)
Jamie Shiers, CERN Grid Deployment Group, IT Department, Geneva, Swiss

19:00-20.00 POSTER SESSION

21:00 *Gala Dinner*

Saturday, September 20, 2008

SESSION 4: Advanced Bioinformatics Technologies and Applications
Chair: George Magklaras

09:00-09:20 Massive non natural proteins structure prediction using Grid technologies
Fabio Polticelli, University Roma Tre (IT)

09:20-09:40 The Genius Grid portal and the robot certificates: a new tool for e-science
Giuseppe La Rocca, INFN, Catania (IT)

09:40-10:00 GPU accelerated RNA-RNA interaction algorithm
Guillaume Rizk, IRISA-Symbiose (FR)

10:00-10:10 The EMBRACE Project
Gisel Andreas ITB-Bari, CNR (IT)

10:10-10:40 *Coffee Break &* POSTER SESSION

10:40-11:00 The interpretation of protein structures based on graph theory
Mahnaz Habibi, Shahid Beheshti University, Tehran, Iran

11:00-11:20 ENGINEDB: A repository of functional analogues
Andreas Gisel, Istituto di Tecnologie Biomediche, CNR, Bari, (IT)

11:20-11:40 GOMIR: A stand alone application for human microRNA target analysis and gene ontology clustering
Maria Roubelakis, Academy of Athens, Biomedical Research Foundation, Athens (GR)

11:40-12:00 When data integration leads to a new concept: The orphan enzymes
Olivier Lespinet, Institut de Génétique et Microbiologie, Université Paris-Sud, Orsay (FR)

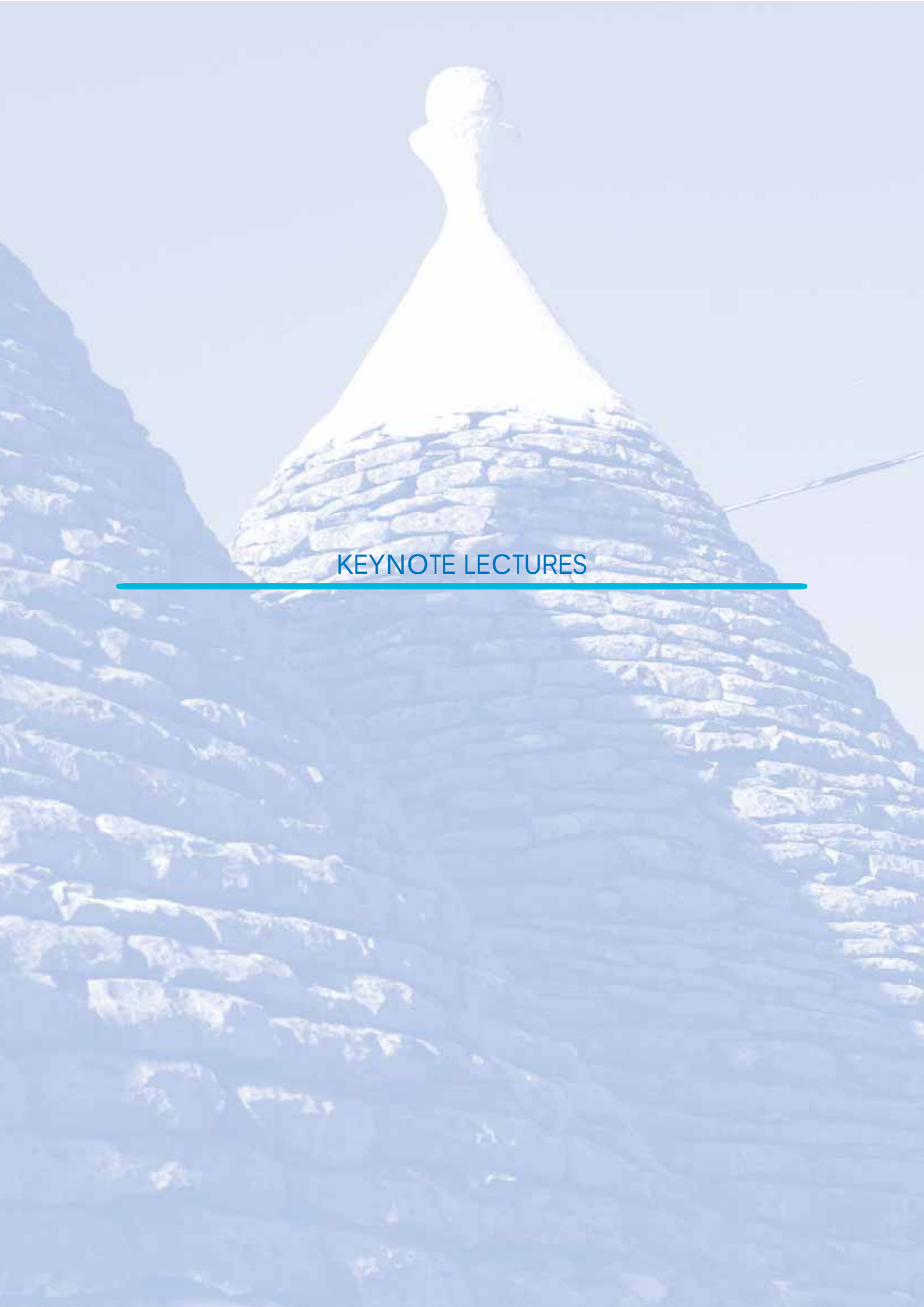
12:00-12:20 Integrating ERV sequence and structural features with DAS and EBIOX
Alvaro Martinez Barrio and Erik Lagercrantz, The Linnaeus Centre for Bioinformatics, Uppsala University (SE)

12:20-12:40 Computational annotation of UTR cis-regulatory modules through frequent pattern mining
D'Elia Domenica ITB-Bari, CNR (IT)

12:40-13:00 A bioinformatics knowledge discovery application for Grid computing
Marcello Castellano, Politecnico di Bari (IT)

13:00-13:15 Concluding remarks

13:15-15:00 *Lunch*



KEYNOTE LECTURES

- BIODIVERSITY INFORMATICS: ENABLING A MACROSCOPIC VIEW OF BIOLOGY -

Indra Neil Sarkar

MBLWHOI Library, Marine Biological Laboratory, Woods Hole, MA USA

There are an estimated 1.8 million organisms that are known to us on Earth. Still, much of biological and biomedical studies remain focused on a limited number of "model" organisms. These model organisms are crucial in the understanding of topics such as basic genetics, genotype-phenotype correlations, and disease inquiries. At the same time, such studies can be complemented with data from non-model organisms. Biodiversity informatics is focused on the application and development of techniques to link data spanning the full spectrum of life.

This presentation will explore the various aspects of biodiversity informatics, ranging from molecular data to population data. First we will explore the types of structured data that are available from existing resources (e.g., molecular sequence data and occurrence data). We will particularly emphasize DNA barcode data and its analysis, and how it can lead to the development of areas where there can be immediate synergy between biomedical inquiry and biodiversity knowledge. Next, techniques for extracting biodiversity data from unstructured sources (e.g., literature) will be discussed. In particular, we will focus on the linking of information across data sources using natural language processing techniques that are honed to identify biodiversity entities (with a particular emphasis on organism name identification).

Throughout the presentation, we will ground ourselves in the discussion of how biodiversity data can complement biomedical studies. To this end, the discussion of biodiversity informatics will be done within two contexts: (1) Infectious Diseases and (2) the Biology of Aging. In so doing, we will demonstrate how a "macroscopic" view of life on Earth can be used to facilitate biomedical studies.

- THE BAROCODE OF LIFE: BRINGING GENOMICS TO BIODIVERSITY -

Hajibabaei Mehrdad

- Canadian Centre for BNA Barcoding, Biodiversity Institute of Ontario, University of Guelph, Guelph, Canada

In the past 5 years DNA barcoding has evolved from a concept to an international movement. It has generated interest (and controversy) among scientists and has been widely broadcasted in scientific literature and popular media.
DNA barcoding promotes biodiversity analysis through the use of standard genomic sequences in a minimalist way. The Barcode of Life Data Systems (BOLD)-- a web-based global data storage, management and analysis platform for DNA barcodes-- so far contains ~0.5M DNA barcodes from 50K species. Barcode libraries have been used in a wide range of applications from basic research to industrial and economic sectors.
Research is now focused on gathering barcode data directly from environmental samples by using next generation massively-parallelized sequencing platforms. This environmental barcoding approach will dramatically broaden the application of DNA barcoding in biodiversity analysis and environmental biomonitoring. In addition, recent advances in microfluidics and genomics technologies can transform DNA barcoding workflow from a lab operation to a single portable device. This device, once built, will allow anyone at any time to read biodiversity.

- POLICIES, NETWORK, RESOURCES, MATERIALS AND CURRICULA FOR ADVANCING BIOINFORMATICS EDUCATION: 10 YEARS OF APBIONET -

Tin Wee Tan

Founding Secretariat, APBioNet - Tan Tin Wee, currently ISCB board director in his second term, is an associate professor and deputy head at the Dept of Biochemistry, YLL School of Medicine, National University of Singapore, among other jobs and responsibilities he has to shoulder

In ten years, the landscape of bioinformatics in the world has changed dramatically. Our efforts to bring bioinformatics to Asia has become much harder because of shifting targets. Without proper network infrastructure, bioinformatics is unthinkable, and since 1996, we strove to build advanced internet networks amongst our countries, from the Asia Pacific Advanced Network (APAN apan.net) to USA to the TransEurasia Information Network TEIN2 to Europe. Within our own nations, we built our own broadband high speed networks, Singaren (1997), TANET to ThaiSARN, MYREN, PERN and so on and established our grid computing infrastructure; needless to say, some of the earliest and biggest users came from the bioinformatics and life science community. As biological databases grew exponentially, the need for rapid dissemination of datasets was met by the Bio-Mirrors project in the late 90s and subsequently by the P2P initiative for developing country institutions with bandwidth constraints, even as we contemplate hard disk delivery today for the remaining data-have-nots. In addressing the lack of software tools that can be easily set up as resources for end users, we explored the development of APBioBox for a comprehensive set of grid-enabled bioinformatics software with Sun Microsystems BioCluster Grid. This led to complete OS of precompiled biosoftware such as APBioKnoppix LiveCD in 2004 and more recently, the modular BioSlax LiveDVD (2007), bootable USB stick and the BioSlax VMplayer version (2008). A server version of BioSlax with a 1 terabyte external hard disk containing Bio-Mirrors can transform any higher end desktop into a reasonable bioserver for a class of 30 within half an hour, offering blast, wemboss, clustalw web and ftp services, etc. This will supplement online offerings of bioinformatics e-learning which was started in 2001 with the S* Life Science Informatics Alliance's introductory bioinformatics course on the IVLE and later the Moodle platform. There, we also explored skype conferencing for bioinformatics PBL tutorials with volunteer teaching assistants as facilitators, as well as CENTRA and Microsoft's NetMeeting.

To provide greater awareness in bioinformatics with an outlet for poster and oral presentations amongst our budding bioinformatics research community, we started the International Conference on Bioinformatics (InCoB) in 2002, now in its 7th year, and held them close to communities of life scientists. This led to the publication of the best papers in BMC Bioinformatics and Bioinformation journals.

In order to continue this steady growth, we have been discussing curriculum development and pro-bioinformatics policies in the East Asia Bioinformation Network meetings, now in its 3rd session and within the intergovernmental ASEAN COST. The need for outreach training courses saw APBioNet partnering FAOBMB, IUBMB, ISCB and S* to hold events from Bogor to Hanoi, from Phnom Penh and Manila to Riyadh and Lahore over the past ten years. Because we do not have systematic government or inter-government funding such as is available within the European Community, financial resources for the bioinformatics acheivements were mainly ad hoc and through the kind sponsorship of industry and forward looking organisations including IDRC's PAN Asia Networking grants. The rest were ad hoc institutional funding and personal voluntary work pro bono. Today we have good evidence of steady research work published out of our region, especially from China, India, Japan, Korea, Taiwan, Singapore and increasingly from developing countries in south-east Asia. However, they are mainly from computing labs.

We are continuing our efforts to upgrade the quality of education and make bioinformatics education basic to Asian life sciences and an essential component of an Asian biologist's and biotechnologist's knowledge and skill set. For this, we are embarking on advocacy work to influence governmental and institutional policy, concurrently with our never-ending effort to keep up with the needs of bandwidth, database, computing resource, educational materials, courses and pedagogy such as the CanalAVIST initiative and in LAMS technology for improved e-learning. One day, we might just close up the enormous gap between the intellectual haves and have-nots for bioinformatics and computational biology.

- EVOLUTION OF GENE REGULATORY CODE -

Alexander E. Kel

BIOBASE GmbH, Halchtersche Str. 33, D-38304 Wolfenbüttel, Germany

Regulation of gene expression is accomplished through binding of transcription factors (TFs) to distinct regions of DNA (TF binding sites, TFBS), and, after anchoring at these sites, transmission of the regulatory signal to the basal transcription complex. Gene regulation code which would enable a "translation" of DNA regulatory sequences into regulatory function they possess are still unclear. Some of TFs are specific for a particular tissue, a definite stage of development, or a given extracellular signal, but most transcription factors are involved in gene regulation under a rather wide spectrum of cellular conditions. It is clear by now that combinations of transcription factors rather than single factors drive gene transcription and define its specificity. Dynamic function-specific complexes of many different transcription factors, so called enhanceosomes are formed at gene promoters and enhancers controlling gene expression in a specific manner. At the level of DNA, the blueprints for assembling such variable TF complexes on promoter regions may be observed as specific combinations of TFBS located in close proximity to each other. We call such structures "Composite Modules (CMs)".

The multiplicity of cellular conditions in which eukaryotic genes should be expressed is the cause of polyfunctionality of the structure of their transcription regulatory regions. We believe that this polyfunctionality is governed through alternative CMs. In the lecture I will present a "fuzzy puzzle" model of the gene regulatory code, which based on the principle of encoding multiple regulatory messages in the same DNA sequence. The structure of regulatory sequences on one hand and the specific features of transcription factors on the other provide a possibilities to encode several regulatory programs within one regulatory region. Such structure allows receiving and integration of multiple regulatory signals through reuse of the same DNA sequence of gene promoters and enhancers.

We developed a novel tool, Composite Module Analyst (CMA) (Kel et al., 2006), that applies a novel approach for defining promoter models based on composition of single transcription factor binding sites as well as their pairs located inside local regulatory domains (corresponding to enchancer/silencer subregions). We use the genetic algorithm technique and utilize a multicomponent fitness function for defining the function specific composite modules in promoters of genes co-regulated in specific conditions (tissue, organ, stage of development, cell cycle, response to particular extracellular signal).

We applied the CMA tool to analyze promoters of tissue/organ specific genes (based on the organ ontology Cytomer® and define Composite Modules characteristic for promoters of genes specifically (or abundantly) expressed in those tissues, which might be considered as the components of the gene regulation code.

We think that "fuzzy puzzle" principle of the gene regulatory code is the result of genome evolution of multicellular organisms that shall overcome evolutionary bottlenecks caused by the requirement of multiple ontogenetic programs to be encoded in a single genome.

1. Kel,A, Konovalova,T, Waleev,T, Cheremushkin,E, Kel-Margoulis,O, Wingender,E. (2006) Composite Module Analyst: a fitness-based tool for identification of transcription factor binding site combinations. Bioinformatics. 22, 1190-1197 (2006).

- GRIDS FOR LIFE SCIENCES: STATUS AND PERSPECTIVES -

Vincent Breton

Laboratoire de Physique Corpusculaire de Clermont-Ferrand, CNRS-IN2P3 Université Blaise Pascal, France

Life sciences are facing an exponential growth of the volume of experimental data. We will discuss how grids can help handle this evolution as well as provide novel approaches to improve data integration and interoperability.
We will also report on the progresses achieved in the development and deployment of life science applications in grid environments.

- DATA CHALLENGES IN THE WORLDWIDE LHC COMPUTING GRID (WLCG) -

Jamie Shiers

Laboratoire de Physique Corpusculaire de Clermont-Ferrand, CNRS-IN2P3 Université Blaise Pascal, Grid Deployment Group, IT Department, CERN, Geneva, Switzerland

The world's largest - and, operating at 1.9oK, also the coolest - scientific machine, the Large Hadron Collider (LHC) at CERN, Geneva, is undergoing final preparations for data taking from proton-proton collisions that should commence later in 2008 (data taking from interactions of cosmic rays with the detectors has already started). Four massive detectors will take data at rates from hundreds to thousands of MB/s, leading to a total data sample that will grow by approximately 15PB per year. The data from these “experiments” - collaborations of several thousand physicists from all around the globe - will be processed by a federation of production Grid infrastructures. CERN - the Tier0 - is complemented by roughly 10 Tier1 sites and 100 Tier2s. To first approximation, the sum of resources at each 'tier' is constant. The Tier0 site is responsible for data taking and first-pass processing, the Tier1s for all subsequent reprocessing and the Tier2s for analysis and Monte Carlo simulation. In such an environment, the need for de facto standards is clear, where interfaces, rather than implementations are defined. Furthermore, data taking can be expected to take place for some 10-15 years, with analysis continuing a few years longer - over which time, significant changes in storage and computing can be expected.

This talk describes the principle challenges involved in offering a data-intensive production service that crosses both time-zones (easy?) and management domains (highly complex). Simple yet proven techniques for delivering robust and resilient services that optimally use the key features of grid computing are discussed, together with the potential benefits that these solutions could bring to widely different disciplines. We also look back briefly to the lessons learned from the previous generation of experiments - including those from the LEP collider that was housed in the same 27km tunnel some 100m below the surface of the earth. Given the lifetime of the LHC experiments, the need to perform massive data migration(s) should not be overlooked and we review experience where significant data migrations (hundreds of TB to PB) were required - some years prior to LHC data taking!

A recent analysis of the most critical services in the WLCG environment confirms that those related to data management and to (distributed) database services are of top priority and that problems with these services have the most significant impact on the experiments' production. This is translated into powerful - but yet widely understood - service techniques, which we describe in detail, together with targets for expert intervention in case of problems, as well as problem resolution. The possibly surprising conclusion is that reliable services are less manpower intensive than less reliable ones - provided the appropriate care and attention is taken from the design stage. (Many of these benefits can also be realized later - but at significantly higher cost.)

ORAL PRESENTATIONS

Abstract 72

- TOWARDS BARCODE MARKERS IN FUNGI: AN INTRON MAP OF ASCOMYCOTA MITOCHONDRIA -

Santamaria Monica*[1], Vicario Saverio[1], Domenica D'Elia[1], Pappadà Graziano[2], Scioscia Gaetano[3], Vicario Saverio[1], Scazzocchio Claudio[4], Saccone Cecilia[5]

- [1] CNR - Istituto di Tecnologie Biomediche, Sede di Bari ~ Bari ~ Italy - [2] Exhicon srl ~ Bari ~ Italy - IBM Italy S.p.A. - [3] IBM Innovation Lab ~ Bari ~ - [4] Institut de Gènètique et Microbiologie, UMR 8621 CNRS, Universitè Paris-Sud (XI) ~ Orsay cedex ~ France - [5] Dipartimento di Biochimica e Biologia Molecolare, Università di Bari ~ Bari ~ Italy

Motivation: A rapid, standardized and cost-effective identification system is now essential for Fungi owing to their wide involvement in human health and life quality. Currently the molecular identification of species in Fungi is based primarily on nuclear DNA, but the potential use of mitochondrial markers has also been considered, due to their peculiar and favourable features, among which, above all, their high copy number, the possibility of an easier and cheaper recovering and the paucity of repetitive regions. Unfortunately, a serious difficulty in the PCR and bioinformatic surveys is due to the presence of mobile introns in almost all the fungal mitochondrial genes. The aim of the present work is to verify the incidence of this phenomenon in
Ascomycota and to identify one or more mitochondrial gene regions where introns are missing so as to propose them as species markers (barcodes).
Methods: The general trend towards a large occurrence of introns in the mitochondrial genome of Fungi has been confirmed in Ascomycota, except for some specific regions, by an extensive bioinformatic analysis, performed on 7234 records of 11 mitochondrial protein coding genes and 2 mitochondrial rRNA coding genes belonging to this phylum, available in Genbank. A new query approach, developed within a databases federation system designed to manage, integrate and enhance connections among the information possibly hosted in heterogeneous data sources, has been applied to retrieve, in an effective manner, relevant information usually present in the entries of a biological database, but hardly selectable through the classical query systems. This approach has allowed to avoid a series of alignment and retrieval stages based on the similarity calculation, which inevitably produce false positives and negatives in the final results.
Results: Despite the large pervasiveness of introns in Ascomycota mitochondrial genes, the results of the present work have shown that specific regions from at least three alternative genes, namely ND2, ND4 and ND6, seem intron-poor and large enough to be considered barcode candidates for Ascomycota. This finding could be the first step towards a mitochondrial barcoding strategy similar to the standard approach routinely employed in metazoa, and its use would prevent other efforts to look for alternative, less efficient or more expensive, strategies to bypass the introns problem.

Abstract 46

- GAINS AND LOSSES OF LINEAGE-SPECIFIC GROUP II INTRON IN MITOCHONDRIA OF GYMNOSPERMS: MOLECULAR EVOLUTIONARY AND PHYLOGENETIC IMPLICATIONS -

Regina Teresa M.R.[1], Quagliariello Carla*[1]

-[1]*Dipartimento di Biologia Cellulare, Università degli Studi della Calabria ~ Arcavacata di Rende (CS) ~ Italy*

1F) Molecular biodiversity, DNA Barcode and metagenomics

Motivation: The mitochondrial rps3 gene harbours a single group II intron (rps3i1) at a well conserved insertion site from algae up to the angiosperms analyzed so far, with the exception of Beta and Marchantia. Interestingly, in gymnosperms the rps3 reading frame is split by two group II intron, rps3i1 and rps3i2 [Regina et al. J. Mol. Evol. 2005; 60, 196-206]. In this study we surveyed a wide range of representatives of all the extant gymnosperms to get insights into allocation and conservation of group II introns and further test the performance of the novel rps3 intron gains and losses as informative character in phylogenetic inferences among the four living gymnosperm orders (Burleigh et al. Am. J. Bot. 2004; 91, 1599-1613).
Methods: Total genomic DNA was isolated by standard CTAB method (Doyle and Doyle Focus 1990; 12, 13-15) or provided directly by the DNA bank at the Kew Royal Botanic Gardens (UK). The mitochondrial rps3 introns were amplified by PCR using specific primers and directly sequenced. Nuclear (18S), plastidial (rbcL) and mitochondrial (cox1, atpA, rps3) sequences were retrieved from GenBank. Structural alignments of (i.) concatenated mitochondrial sequences and (ii.) concatenated mitochondrial-plastid-nuclear sequences were conducted with ClustalX (Thompson et al. Nucleic Acids Res. 1997; 24, 4876-4882) and used to form a multigene and a multigenome matrix, respectively. Maximum parsimony (MP) and maximum likelihood (ML) analyses were, thus, performed using PAUP* V. 4.0b10 program (Swofford 2003 Sinauer, Sunderland, MA).
Results: We report the shared presence of both rps3i1 and rps3i2 in most of the surveyed gymnosperms but unveil several remarkable exceptions among closely related species. Therefore, we show that the distribution pattern of the rps3 introns is able to discriminate among divergent lineages of living gymnosperms. Furthermore, our multigene and/or multigenome MP and ML analyses demonstrate the mitochondrial rps3i2 as a proper informative character to highlight new mitochondrial genomic endeavours and diverse innovations characterizing the plant molecular biodiversity as well as to reinterpret the phylogenetic inter- and intrafamilial relationships among the extant lineages of gymnosperms.

Abstract 10

- METAGENOME ANNOTATION: AN OPPORTUNITY FOR UNDER-GRADUATE BIOINFORMATICS TEACHING -

Hingamp Pascal*[1], Brochier Céline[2], Talla Emmanuel[1], Gautheret Daniel[3], Thieffry Denis[1], Herrmann Carl[1]

- [1]Biology Department, Mediterranean University ~ Marseille ~ France - [2]Biology Department, Provence University ~ Marseille ~ France - [3]Université Paris Sud ~ Paris ~ France

2B) Education and Training: Instruments, cooperation and collaboration

Motivation: The bottleneck in genomics is shifting from sequencing to annotating, increasing the demand for expert annotators. It is in the interest of research and future job seekers that graduate training anticipates this trend by introducing students to the art of raw sequence annotation.

Methods: We have taken advantage of the increasing amount of metagenomic data publicly available to develop a teaching environment in which undergraduate students are given the opportunity to "turn data into knowledge". This internet teaching platform, called "Annotathon", fosters team work and guides apprentice annotators through each step of in-silico analyses, from ORF identification to functional and phylogenetic classification. Generating raw results is an integral part of the exercise, but emphasis is put on their interpretation and critical assessment.

The online format is ideally suited for student involvement outside class whilst allowing instructors to provide annotators with continuous feedback. Communication relies on classical internal forums and chats, but more importantly on an iterative evaluation cycle which allows students to respond to constructive criticism and produce enhanced versions of their annotations.

Results: The 720 students that have taken part in the Annotathon over the past three years have analyzed a total of 23 Mb of ocean microbial DNA, representing 9500 hours of cumulative annotation. The following aspects of the approach appear to significantly contribute to its success:

a) learning by doing: bioinformatics is best introduced by first hand experience; theoretical considerations are easier to grasp once truly familiar with the tools.
b) learning by repetition: repeating the analyses on several sequences gives the students the opportunity to experience a wide range of situations, e.g. BLAST report for widespread proteins versus ORFans etc.
c) learning by excitement: according to students, exploring yet unannotated sequences is a major incentive.
d) learning from constructive criticism: giving students the opportunity to correct themselves results in noticeable progression over time.

The Annotathon environment is available as an open source software, but teams are also welcome to join us on our public server http://biologie.univ-mrs.fr/annotathon/. Ideally as more teams join in, this could lead to an educative distributed annotation jamboree with room for modest scientific contribution.

Abstract 42

- GRID-BASED BUSINESS-TO-ACADEMIA COLLABORATIONS -

Kamuzinzi Richard*[1], Bottu Guy[2], Colet Marc[1]

- [1]Bioinformatics Unit, Université Libre de Bruxelles ~ Gosselies ~ Belgium - [2]Belgian EMBnet Node - Bioinformatics Unit, Université Libre de Bruxelles ~ Bruxelles ~ Belgium

1l) Grid technologies and Web Services

Motivation: Modern organizations active in the Life Sciences can no longer envisage research and development (R&D) without the involvement of multiple, distributed and independent partners. Actually, the R&D processes are often complex tasks where resources such as data and analysis services have to be shared and integrated among internal and external collaborative entities. At the same time organizations, especially those from the industrial sector, expect to reach the same or even better research objectives while reducing supporting costs and risks. Thus, in a particular research project, participating partners create a virtual organization (VO) and bring together different scientific disciplines and computing resources to address the underlying goal of the research project. Information confidential to a party must be properly managed to ensure the preservation of intellectual property rights.
Results: In this presentation, we present some of the last outcomes of SIMDAT, a research project funded by the European commission. Within this project, we successfully addressed this need of research virtualisation by developing advanced prototypes to demonstrate the feasibility of online collaborations. Additionally, the prototypes showed that both academic and commercial organizations can build reliable and dynamic partnerships to support collaborative R&D processes. In other words, along with the common business-to-business (B2B) collaboration scheme, organizations in e-science can really also adopt the business-to-academia (B2A) collaboration scheme. The approach adopted to design these prototypes is based on a service oriented GRID infrastructure, namely GRIA, which complies with security constraints imposed by the industrial sector. Moreover, the KDE workflow platform from InforSense is used to access the services infrastructure in order to combine applications and services exposed by different parties. Finally, the web interface wEMBOSS is used to provide the end user (the analyst) with a simple environment from which she/he could interact with the VO without having to worry about the technological level issues.

To conclude, we believe that the EMBnet organization could benefit from SIMDAT experience and developments to build a GRID-based network of industry strength services where different resources could be serviced by different nodes and the whole deployed and viewed as a single organization.

Abstract 11

- SPRINTS AT GENESILICO - SOFTWARE ENGINEERING TECHNIQUES IN A BIOINFORMATICS LAB -

Rother Kristian*[1], Papaj Grzegorz[1], Feder Marcin[1], Kosinski Jan[1], Koslowski Lukasz[1], Potrzebowski Wojciech[1], Kaminski Andrzej[1], Pawlowski Marcin[1], Kogut Jan[1], Fijalkowski Maciek[1], Gajda Michal[1], Jarzynka Tomasz[1], Tkalinska Ewa[1], Orlowski Jerzy[1], Tkaczuk Karolina[1], Puton Tomasz[2], Musielak Magdalena[2], Koscinski Lukasz[2], Czwojdrak Joanna[2], Milanowska Kaja[2], Kaminska Katarzyna H[2], Osinski Tomasz[2], Domagalski Marcin[2], Kaczynski Jan[1], Figiel Malgorzata[1], Tuszynska Irina[1], Smit Sandra[3], Knight Rob[4], Huttley Gavin A[5], Bujnicki Janusz M[1]

- [1]International Institute of Molecular and Cell Biology ~ Warsaw ~ Poland - [2]Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University ~ Poznan ~ Poland - [3]Centre for Integrative Bioinformatics VU (IBIVU), Vrije Universiteit Amsterdam ~ Amsterdam ~ Netherlands - [4]Department of Chemistry and Biochemistry, University of Colorado ~ Boulder ~ United States - [5]Computational Genomics Laboratory, John Curtin School of Medical Research, The Australian National University ~ Canberra ~ Australia

1K) Biological data integration

Motivation: In our everyday research a multitude of programs were written to handle protein sequence analyses. Despite Python as a common programming language, little code was shared among people, and the result was usually a mess. In late 2007, we decided make a concerted effort to create a pipeline for protein family analysis that is documented, tested, and easy to maintain. To implement the project, a series of Sprints - focused two-day programming sessions - was organized.
Methods: In total, 26 people participated: 18 coders, 4 users, 3 technical staff, and 1 correspondent for internal news. Because many undergraduate students attended, the coding was done in pairs of one junior and one experienced programmer. Each Sprint was followed by a two-week cleanup period, where three experienced programmers added Unit Tests and documentation. One Sprint was devoted to bug fixing. Using the Trac ticket system, 101 bug reports were collected, 77 of which could be fixed within the first week. During bug fixing, additional test code was written to make sure the same bugs cannot reoccur. Our software intensively uses PyCogent [1], a Python library that supports many biological applications. During a smaller, three-continent Sprint with the PyCogent developers, the usage of the library was optimized, and cookbook-style documentation for PyCogent could be developed.
Results: This approach benefits from programming in a focused and communicative environment. Experienced programmers write better code because they know it will be read, and students are trained 'on the job'. The code quality benefits from using Unit Tests, ticket systems, and a code repository. As potential users, all participants became familiar with the software long before it was finished, and therefore provided many important suggestions. The outcome is a software pipeline supporting many steps from BLAST/PSI-BLAST queries, creating, updating and filtering alignments, to writing phylogenetic and other reports. The software and source code is available on www.genesilico.pl/python_sprint.

[1] Knight R, Maxwell P, Birmingham A, Carnes J, Caporaso JG, Easton BC, Eaton M, Hamady M, Lindsay H, Liu Z, Lozupone C, McDonald D, Robeson M, Sammut R, Smit S, Wakefield MJ, Widmann J, Wikman S, Wilson S, Ying H, Huttley GA. 2007. PyCogent: a toolkit for making sense from sequence. Genome Biol. 8(8):R171.

Abstract 22

- GOMIR: A STAND ALONE APPLICATION FOR HUMAN MICRORNA TARGET ANALYSIS AND GENE ONTOLOGY CLUSTERING -

Zotos Pantelis[1], Papachristoudis George[2], Michalopoulos Ioannis[1], Roubelakis Maria*[1], Pappa Kalliopi[1], Anagnou Nikolaos[1], Kossida Sophia[1]

- [1]Academy of Athens, Biomedical Research Foundation ~ Athens ~ Greece - [2]MIT ~ Cambridge, MA ~ United States

Motivation: MicroRNAs are single-stranded RNA molecules of about 20-23 nucleotides length found in a wide variety of organisms. MicroRNAs regulate gene expression, by interacting with target mRNAs at specific sites in order to induce cleavage of the message or inhibit translation. Predicting or verifying mRNA targets of specific microRNAs is a difficult process of great importance.
Methods: GOmir is a novel stand-alone application consisting of two separate tools: JTarget and TAGGO.
JTarget integrates microRNA target prediction and functional analysis by combining the predicted target genes from TargetScan, miRanda, RNAhybrid and PicTar computational tools and also providing a full gene description and functional analysis for each target gene. On the other hand, TAGGO application is designed to automatically group gene ontology annotations, taking advantage of the Gene Ontology (GO), in order to extract the main attributes of sets of proteins.
Results: GOmir represents a new tool incorporating two separate Java applications integrated into one stand-alone Java application. GOmir (by using up to four different databases) introduces, for the first time, miRNA predicted targets accompanied by (a) full gene description, (b) functional analysis and (c) detailed gene ontology clustering. Additionally a reverse search initiated by a potential target can also be conducted. GOmir can freely be downloaded from http://bioacademy.gr/bioinformatics/projects/GOmir .

Abstract 30

- A GREATER DIVERSITY OF RIBOSWITCHES IDENTIFIED THROUGH THE PRESENCE OF ALTERNATIVE STRUCTURES AND OTHER CONSTRAINTS -

Naville Magali*[1], Marchais Antonin[1], Gautheret Daniel[1]

-[1]*Institut de Génétique et Microbiologie, Université Paris-Sud 11 ~ Orsay ~ France*

1A) Genomics

Motivation: Riboswitches are non-coding RNA elements located in 5' untranslated region of genes that control gene expression in response to specific ligands. Currently, efficient methods for riboswitch computational prediction mostly rely on sequence and/or structure conservation. As a likely consequence of this kind of approach, riboswitch families present a marked uniformity in terms of structure, if not sequence, conservation, even between distant species. Here we propose a new and different protocol for the detection of more evolutionary isolated systems, based on their mechanism of action. This approach, which is not limited by conservation criteria, allows the prediction of novel riboswitches that escaped established screening methods.

Methods: Our detection strategy identifies regulatory systems based on a terminator/anti-terminator model. This includes the majority of riboswitches in certain bacterial lignages like Firmicutes, as well as T-boxes or simple attenuators. For now, the prediction was applied to 7 species including Bacillus subtilis, in which the relatively abundant annotation allowed a statistical validation of the method. First, all 5' non-coding regions of genes are extracted and screened for rho-independent terminators. A region encompassing the direct strand of each detected terminator is used as a probe for RNAhybrid, a program that looks for a possible hybridization target in the remaining upstream sequence. The probe/target couple corresponds to a putative anti-terminator structure. Predictions are ranked using a combination of criteria including putative anti-terminator free energy or flanking gene distance, orientation and function.

Results: In Bacillus subtilis, our initial screen led to the detection of 718 5'-terminators, among which 38 correspond to known riboswitch/attenuator systems. The subsequent screens based on free energy and flanking gene information retained 82 candidates among which 32 known riboswitches/attenuators. By increasing significantly the specificity without altering much the sensitivity, our screening procedure thus appear particularly relevant. Many novel candidates are found upstream of transporters or secondary metabolite processing genes. This promising approach should considerably diversify our collection of riboswitches and attenuator systems, notably in rho-independant terminator-rich species.

Abstract 36

- IMGT/LIGMOTIF: A TOOL FOR IMMUNOGLOBULIN AND T CELL RECEPTOR GENE IDENTIFICATION AND DESCRIPTION IN LARGE GENOMIC SEQUENCES -

Lane Jérôme*[1], Lefranc Marie-Paule[1], Duroux Patrice[1]

- [1]IMGT®, the international ImMunoGeneTics information system®, Université Montpellier 2 ~ Montpellier ~ France

1A) Genomics

Motivation: The immunoglobulins (IG) and T cell receptors (TR) are the major molecular components of the adaptive immune response of vertebrates. IG and TR loci consist of variable (V), diversity (D) and joining (J) genes organized in multigene groups. Owing to the unusual structure of IG and TR genes, conventional bioinformatic software are not adapted to their identification and description in large genomic sequences. A tool, IMGT/LIGMotif, has been developed for IG and TR gene prediction and annotation. It is based on IMGT-ONTOLOGY, the first ontology in immunogenetics, at IMGT®, the international ImMunoGeneTics information system® (http://imgt.cines.fr).

Methods: The software model is based on V, D and J gene prototypes. These prototypes are described by 45 standardized IMGT® labels organized on structural biological criteria. Sixteen labels refer to short specific motifs: splicing sites, heptamers, nonamers and conserved amino acids. They are defined by their conserved sequences and their positions in prototypes. Fifteen longer labels refer to core coding regions, recombination signals and gene units. These long motifs are stored in referential sequence databases. The remaining labels are required for a complete annotation.

The algorithm comprises several steps. (1) The analysed genomic sequence is aligned using BLAST with the referential sequence databases. (2) The obtained alignments (or HSP for High Scoring Pairs) are selected on their E-value, score, identity and length. (3) As the HSPs are defined with labels and as the localization of the labels in genes is defined by prototypes, the next step consists in grouping HSP in genes. (4) Short motifs are searched within and in the neighbouring of labels identified in the previous step. (5) In the last step, short motifs are used as anchors to localized labels precisely and to complete the annotation. At this step IMGT/V-QUEST software is also used for the V genes.

Results: IMGT/LIGMotif was evaluated for gene description and identification in human and mouse IG and TR large genomic sequences (7 megabases). In term of prediction, all known human V, D and J genes were identified by IMGT/LIGMotif. More pseudogenes were found than expected. These sequences are currently analysed to check their biological significance. In term of gene description, the V, D and J gene labels were all correctly assigned.

Abstract 37

- FUNCTIONAL ASSESSMENT OF TIME COURSE MICROARRAY DATA -

Conesa Ana*[1], Nueda Maria José[2], García-García Francisco[1], Sebastian Patricia[1], Dopazo Joaquín[1], Ferrer Alberto[3]

- [1]Bioinformatics and Genomics Department, Centro de Investigación Principe Felipe ~ Valencia ~ Spain - [2]Department of Statistics and Operative Research, University of Alicante ~ Alicante ~ Spain - [3]Department of Statistics and Operative Research, Poliechnical University of Valencia ~ Valencia ~ Spain

1B) Transcriptomics

Motivation: Time-course microarray study the evolution of gene expression along time across one or several experimental conditions (series). Most developed analysis methods are focussed on the clustering or differential expression analysis of the genes in the dataset and do not integrate functional information. In this work we propose two methods for the functional assessment in time course microarray data that directly exploits the dynamics of expression of the functional categories genes are annotated to.

Methods: We have adapted two methods previously developed for the analysis of time course data, to integrate gene functional information. maSigFun derives from the maSigPro methodology, a regression strategy that models expression patterns and identify genes with patterns differences across experimental series. maSigFun fits a regression model for groups of genes labelled by a functional class and selects those categories which has significant model. ASCA-functional is an extension of the ASCA-genes method. ASCA uses ANOVA and PCA to identify principal components associated to time expression signals. The ASCA-functional strategy uses these leverage values to rank genes which are then analyzed by GSEA (Gene Set Enrichment Analysis). Significant classes are those which are enriched within the genes of best resemblance to the major patterns of time gene expression evolution. We used simulated and experimental datasets. Results were compared with the more traditional approach represented by a linear method followed by enrichment analysis of significant genes. As functional scheme, the Gene Ontology was used.

Results: Simulation studies indicated that classes were positively identified when more than 50% of the annotated genes were correlated. Small size clasess were better detected by maSigFun than ASCA-functional. A cutoff value of 0.4 was optimal for the R2 parameter in the maSigFun. Both maSigFun and ASCA-functional were able to identify more and semantically distinctive functional classes than the comparing method. maSigFun selected specific classes with a good level of internal coherence. Furthermore, the time expression co-expression pattern was of was directly depicted by this methology. ASCA-genes tend to identify classes with a larger number of genes and appeared to be a good adaptation of the Gene Set methods to experiments where more than two conditions are involved.

Abstract 24

- CHIPSTER - USER FRIENDLY ANALYSIS SOFTWARE FOR DNA MICROARRAY DATA -

Kallio Aleksi[1], Tuimala Jarno[1], Hupponen Taavi[1], Klemelä Petri[1], Korpelainen Eija*[1]

- [1]CSC -the Finnish IT center for science ~ Espoo ~ Finland

1B) Transcriptomics

Motivation: DNA microarray data analysis is a fast developing field and most of the new methods are published in the international Bioconductor project (http://www.bioconductor.org/). These methods are freely available, but their use requires knowledge of the R programming language. This is limiting, because microarray researchers typically have a life science background without programming experience. In order to bridge this gap we have created Chipster (http://chipster.csc.fi/), a user friendly analysis software which brings a comprehensive collection of up-to-date analysis methods within the reach of bioscientists via its graphical user interface.

Methods: Chipster supports Affymetrix, Illumina, Agilent and cDNA arrays and, being a Java program, it runs on Windows, Linux and Mac OS X. The usual analysis features such as preprocessing, statistical tests, clustering, and annotation are complemented with e.g. linear (mixed) models, bootstrapping hierarchical clustering results, and promoter analysis tools. Chipster currently contains almost 100 analysis and visualization tools, and adding new tools is easy. Users can combine and automate frequently used tools into workflows, or use Chipster's ready made wizards. Chipster keeps track of performed analyses and the user can save the analysis history. The analysis scripts can also be viewed at the source code level.

Chipster's graphical client program runs on the user's own computer and the actual analyses are performed on central computing servers. It is also possible to connect external Web Services to the system. The client software utilizes Java Web Start to make installation and version updates as easy as possible. Chipster is available for local installations and it is open source.

As part of the EMBRACE project (http://www.embracegrid.info/), we are currently developing also programmatic access to Chipster. The pipeline offered as a Web service finds differentially expressed genes in Affymetrix data and runs clustering, annotation, and GO and KEGG enrichment analysis for them.

Results: Taken together, Chipster enables more researchers to benefit from the method development in the R/Bioconductor project by offering an intuitive graphical user interface to the analysis tools. The system allows users to save and share workflows, and the installation and version updates are taken care of centrally.

Abstract 15

- CHARACTERIZATION AND ANALYSIS OF THE EXPRESSION PATTERN OF MICRORNAS IN THE GRAPEVINE VITIS VINIFERA -

Piccolo Viviana*[1], Mica Erica[1], Pè Enrico[2], Pesole Graziano[3], Horner David[1]

- [1]Department of Biomolecular Sciences and Biotechnology, University of Milan ~ Milan ~ Italy - [2]Dip. Settore Agraria, Scuola Sup. di Studi Univ. e Perfezionamento S.Anna ~ Pisa ~ Italy - [3]Istituto Tecnologie Biomediche, Consiglio Nazionale delle Ricerche ~ Bari ~ Italy

1B) Transcriptomics

Motivation: MicroRNAs are small (19-24 nt) noncoding RNAs that play an important role in the regulation of multiple cell events, inhibiting gene expression at the posttranscriptional level by binding target mRNAs that are subsequently degraded or sequestered from translation. Plant microRNA genes are typically transcribed by Pol II to yield polyadenylated primary miRNAs (pri-miRNA). These undergo nuclear cleavage to produce to a stem loop intermediate (pre-miRNA) with specific thermodinamic features. Further processing yields a miRNA:miRNA* duplex with 2 nt 3' overhangs that enters a cytoplasmic ribonucleprotein complex which mediates interaction with target mRNAs.
Systematic analyses of micro RNAs and their expression patterns have been performed in only a few plant model species. The availability of the complete genome sequence of the grapevine (Vitis vinifera), has already permitted genome-wide predictions of microRNAs by purely computational methods. Here we present a comprehensive analysis of expression of both mature microRNAs and their primary transcripts in the grapevine using oligonucleotide arrays and next generation sequencing technologies.
Methods: We integrate tanscriptome information derived from high-throughput Illumina SOLEXA and ABI SOLiD sequence tags derived from both polyA+ transcripts and isolated small RNAs with oligonucleotide array data. We are thus able to detect both mature microRNAs and to establish whether genomic loci corresponding to the pre-miRNA are expressed in various tissues.
RESULTS: Using “next generation” sequencing technologies and oligonucleotide arrays, we are able to demonstrate tissue specificity of expression of many microRNA genes and their precursor sequences. In many cases, the unambiguous alignment of sequence tags derived from polyA+ RNA to the genomic sequence allow provisional mapping of primary microRNA transcripts. It is hoped that the approach outlined here will ultimately provide insights into the regulation of processing of primary microRNAs and precursor microRNAs as well as facilitating identification of sequence elements involved in the regulation of transcription of microRNA genes.

Abstract 59

- IMPROVING THE PREDICTION OF PROTEIN BEHAVIOR IN HYDROPHOBIC INTERACTION CHROMATOGRAPHY AND AQUEOUS TWO-PHASE SYSTEMS WITH CLUSTERING METHODS -

Ugarte Jorge E.[1], Andrews Barbara A.[1], Salgado J. Cristian*[1]

- [1]*Centre for Biochemical Engineering and Biotechnology, Department of Chemical Engineering and Biotechnology, University of Chile ~ Santiago ~ Chile*

1D) Molecular structure prediction, modelling and dynamic

Motivation: The aim of this study is the improvement of mathematical models used to predict the behavior of proteins in hydrophobic interaction chromatography (HIC) and aqueous two-phase systems (ATPS) based on their amino acid composition. This problem was tackled by carrying out clustering analysis over a large database of amino acid properties (APV): Self Organizing Maps, k-means, Simulated Annealing, Growing Neuronal Gas, Growing Grid, and hierarchical Clustering were used. This analysis allows us to generate new APVs from those found in literature, which were used to improve prediction models. Three of these models require only the amino acid composition of proteins and different assumptions regarding the tendency of the amino acids to be exposed to the solvent; the other requires the three dimensional structure of the proteins. These models were adjusted using the new APVs and were evaluated in a set of 12 proteins with known experimental retention time in HIC and 11 with known partition coefficient in ATPS. We found that the best APVs were generated by the Growing Neuronal Gas algorithm. In fact, two vectors that significantly improve the performance of the prediction models were found. Using these vectors the prediction performance of the model based on the 3D structure and the best model based on amino acid composition were improved by 38% and 31%, respectively.

ACKNOWLEDGEMENTS: FONDECYT PostDoctoral Research Project 3070031 and Millennium Scientific Initiative ICM P05-001F.

KEYWORDS: Clustering, prediction models, hydrophobic interaction chromatography and aqueous two-phase systems.

Abstract 67

- CONTACT COORDINATION PATTERNS AND ELECTROSTATIC POTENTIAL AT ALPHA CARBON ATOMS: A DOSSIER OF PROTEIN SECONDARY STRUCTURE ELEMENTS -

Borro Luiz[1], Mazoni Ivan[1], Alvarenga Daniel[1], Cecilio Pablo[1], Grassi Jose[1], Jardine Gilberto[1], Mancini Adauto[1], Neshich Goran*[1]

- [1]*Embrapa Informatica Agropecuaria ~ Campinas, SP ~ Brazil*

1D) Molecular structure prediction, modelling and dynamic

Motivation: The process of protein folding might be investigated by analyzing the secondary structure elements (SSE) in light of the physical chemical characteristics of amino acids. Molecular structure prediction is also dependent fundamentally on how much we know about the SSE and their interplay within the 3D constellation. Our major motivating factor to study in detail the relationship of selected physical chemical parameters and the capacity of determined sequences to build specific SSE came from the fact that we succeeded to build the most extensive database of such parameters - the STING_DB. This allows us to make "signal enhancement" of patterns hidden within diversity of sequences capable of generating the very same SSE.

Methods: We present here analysis of pre-calculated values for the electrostatic potential at the alpha carbons and for the cross-links, previously stored in the STING_RDB. Our procedure was to first separate the proteins from the PDB according to their classes: all alpha, all beta, alpha + beta and alpha/beta. All SSE were grouped and aligned with respect to their length. All aligned SSE, were then analyzed in terms of 47 sequence/structure descriptors (grouped in 32 major classes) such as: electrostatic potential, sequence conservation, hydrophobicity, accessibility, dihedral angles, internal contacts etc.

The Cross Links are defined as contacts (any type from possible 5 classes: Hydrophobic, Hydrogen Bonding, Aromatic Stacking, Salt bridging and Cystein-bridging) established among residues that are far apart in the protein primary sequence, but are close in its 3D fold. The order of cross link is identified as a number of such cross-links established among independent stretches of sequence (the size of which was fixed to 30 Amino Acids). The higher the order, the more important that residue must be for the protein folding/stability.

Results: We found a clear tendency for EP at alpha carbon atoms for alpha helices and beta strands, having negative and positive values, respectively. We also found a clear and opposing tendency for the value of cross links for alpha helices and beta strands, showing less and more, respectively, cross links in comparison to the other parts of proteins.

Abstract 33

- DECIPHERING THE CONNECTIVITY STRUCTURE OF BIOLOGICAL NETWORKS USING MIXNET -

Picard Franck[1], Miele Vincent*[1], Daudin Jean-Jacques[2], Cottret Ludovic[3], Robin Stephane[2]

-[1]CNRS ~ Lyon ~ France - [2]AgroParistech ~ Paris ~ France - [3]Universite Lyon 1 ~ Lyon ~ France

1E) System Biology

Motivation: Understanding the structure of complex networks has become a challenging task which is tackled using clustering techniques. The principle is to gather the nodes of the network into subsets easier to interpret. Several strategies have been proposed, hubs and modules constituting the two most commonly searched substructures, and for those methods, nodes degree often constitutes the building information. Two major criticisms can be made: hubs and modules constitute only two examples of substructures that can be found in networks, and degree only gives a crude view of the network connectivity.

Methods: We present MixNet, a method dedicated to the analysis of networks connectivity structure. It is based on a clustering procedure that uses mixture models to find groups of nodes sharing similar connectivity patterns without any a priori on the characteristics of the groups. Consequently Mixnet can find modules and hubs, but also other structures such as stars, cliques and product connectivity within the same network, these structures being learned directly from the data. Another advantage of MixNet is that it offers a real opportunity to find structures without defining multicriteria strategies that are not robust.

Results: MixNet is applied to various biological networks. MixNet can be used to summarize and understand the information flow that structures the cortex network. We discuss the finding of hubs in the Cortex macaque network that were previously investigated, show how the method can identify core structure like peripherical and central hubs based on the model only. Then MixNet is used to summarize the regulation diagram of the E. Coli transcriptional network. Interestingly, we show that the connectivity structure revealed by the model reflects the building blocks of the network. We define meta motifs at the group level, such as the feed forward loop which imply global regulators, and discuss the identification of important regulatory nodes from the connectivity point of view. We also show the potential of the method on metabolic and food web networks. Interestingly we find that the summary network which is provided by MixNet reflects the core connectivity structures that build the network, which makes the method a valuable tool to understand the functioning of complex network in a reliable manner. The software package as well as examples are available at http://pbil.univ-lyon1.fr/software/mixnet/.

Abstract 16

- AUTOMATIC INFERRING DRUG GENE REGULATORY NETWORKS USING COMPUTATIONAL INTELLIGENCES TOOLS -

Floares Alexandru*[1]

- [1]*SAIA - Solutions of Artificial Intelligence Applications ~ Cluj-Napoca ~ Romania*

1E) System Biology

Motivation: Various drugs and their dosage regimens. The ordinary differential equations approach is probably the most sensible. Unfortunately, this is also the most difficult, tedious, expensive, and time-consuming approach. There is a need for algorithms to automatically infer such models from high-throughput temporal series data. Computational intelligence techniques seem to be better suited to this challenging task than conventional modeling approaches.

Methods: We developed a reverse engineering algorithm - RODES, from Reversing Ordinary Differential Equations Systems (see e.g., Floares, Neural Networks 2008; 21, 379-386) - for drug gene regulating networks. These are gene networks were the regulation is exerted by transcriptions factors and also by drugs. RODES is based on two computational intelligence techniques: genetic programming and neural networks feedback linearization. RODES takes as inputs high-throughput (e.g., microarray) time series data and automatically infer an ordinary differential equations model, discovering the network's structure, and estimating its parameters. The model can be used to identify the molecular mechanisms involved. The algorithm can deal with missing information - some temporal series of the transcription factors, drugs or drug related compounds are missing from the data. For example, an extreme situation is when one wants to model a drug gene regulating and have only microarray temporal series data at his disposal.

Results: RODES algorithm produces systems of ordinary differential equations from experimental or simulated high-throughput time series data, e.g. microarray data. On simulated data, the accuracy and the CPU time were very good - R2 was 0.99 or 1.00 in most experiments, 1 being the maximal R2. In particular, the RODES CPU time is orders of magnitude smaller than the CPU time of other algorithms proposed in the literature.

This is due to reducing the reversing of an ordinary differential equations system to that of individual algebraic equations, and to the possibility of incorporating common a priori knowledge. To our knowledge, this is the first realistic reverse engineering algorithm, based on genetic programming and neural networks, applicable to large drug gene networks.

Abstract 32

- A CHEMOGENOMICS VIEW OF PROTEIN-LIGAND SPACES -

Helena Strömbergsson*[1], Gerard Kleywegt[1]

- [1]*Uppsala University ~ Uppsala ~ Sweden*

1C) Proteomics

Motivation: Chemogenomics is an emerging inter-disciplinary approach to drug discovery that can be defined as the systematic study of the biological effect of a wide array of small molecular-weight ligands on a wide array of macromolecular targets. The field merges traditional ligand-based approaches with biological information on drug targets and lies at the interface of chemistry, biology and informatics. The ultimate goal in chemogenomics is to understand molecular recognition between all possible ligands and all possible drug targets. However, the size of the protein-ligand space makes any systematic experimental characterization of that space impossible. Protein and ligand space have previously been studied as separate entities, but chemogenomic studies deal with large datasets that cover parts of the protein-ligand space. Since chemogenomics deals not only with ligands but also with the macromolecules the ligands interact with it is of interest to find means to explore, compare and visualize protein-ligand spaces as single entities.

Methods: Two chemogenomic protein-ligand interaction datasets were generated for this study. The first dataset represents the structural protein-ligand space, and includes all non-redundant protein-ligand interactions found in the worldwide Protein Data Bank. The second dataset contains all approved drugs and drug targets stored in the DrugBank database, and represents the approved drug-drug target space. To capture biological and physicochemical features of the chemogenomics datasets, descriptors were computed from the primary sequences of the proteins and the three-dimensional structures of the ligands. Principal component analysis was used to analyze the multidimensional data and to create global models of protein-ligand space.

Results: In this study, we present an approach to visualize protein-ligand spaces from a chemogenomics perspective, where both ligand and protein features are taken into account. The method can be applied to any protein-ligand interaction dataset. Here, the approach is applied to analyze the structural protein-ligand space and the protein-ligand space of all approved drugs. We show that this approach can be used to visualize and compare chemogenomics datasets, and to identify close neighbours in the protein-ligand space.

Abstract 3

- IN SILICO PREDICTION OF ESCAPE MUTANTS OF THE HIV-1 PROTEASE -

Agramonte Alina*[1], Pajón Rolando[2], Carrasco Ramón[3], Padrón Juan Alexander[4]

- [1]Bioinformatic Group, University of Informatic Sciences ~ Havana ~ Cuba - [2]Centre for Genetic Engineering and Biotechnology ~ Havana ~ Cuba - [3]Centre for Pharmaceutical Chemistry ~ Havana ~ Cuba - [4]Laboratory for Theoretical and Computational Chemistry, Chemistry Faculty, Havana University ~ Havana ~ Cuba

1C) Proteomics

Motivation: The “reverse vaccinology” is a new approach that allows the in silico study of a vaccine candidate. This method reduces the time needed for the identification of these vaccines candidates and increases the success rate in those that conventional ways seems impossible. Its principal weakness consists of the experimental appearance of escape mutants at short-medium time. In this work, a first approach to predict the emergence of drug-resistant mutants under positive selection in the HIV-1 protease, correlating structural variables with the occurrence of viable mutants at population level, is proposed. Single-point mutants in a protein evolving under positive selection pressure don't randomly occur. They depends on the structural capability of the protein to accept changes without compromising function; that's why it is necessary to predict those sites where it is more likely to appear viable mutations which can be selected as escape mutants after drug treatments.

Methods: To carry out the analysis, several bioinformatics tools were used. These tools are integrated on a distributed calculation platform implemented in our University. All drug-resistant mutants of human immunodeficiency virus type 1 (HIV-1) protease, were retrieved from Los Alamos HIV drug resistance database (http://hiv-web.lanl.gov). Structural models were generated and the energy contribution in the stability of the protein for each single point mutant was evaluated.

Results: Maximum likelihood analysis provides strong evidence of positive selection acting on 19 residues of the HIV-1 protease. This number represents 48% of the total drug-resistant mutants reported in the database, until February 2007. Most of the analyzed drug-resistant mutants favorably contribute to the stability of the protein structure. They show a good correlation between the susceptibility to the occurrence of positive selection on a single point mutation and a favorable contribution of them to the stability of the protein structure. With these results, the computational Grid technology is put in hands of the researchers as an efficient tool to detect viable positive selective mutants. In this sense, it is possible to say that this strategy can be the first step to design a rational automated approach for the search of vaccine candidates to this specific therapeutic target and some others.

Abstract 69

- THE RHNUMTS COMPILATION -

Attimonelli Marcella*[1], Lascaro Daniela[1], Castellana Stefano[1], Gasparre Giuseppe[2], Romeo Giovanni[2], Saccone Cecilia[1]

- [1]Department of Biochemistry and Molecular Biology "E.Quagliariello" ~ Bari ~ Italy - [2]Unit of clinical genetics ~ Bologna ~ Italy

Motivation: Introduction
To a greater or lesser extent, eukaryotic nuclear genomes contain fragments of their mitochondrial genome counterpart, deriving from the random insertion of damaged mtDNA fragments. NumtS (Nuclear mt Sequences) are not equally abundant in all species, and are redundant and polymorphic in terms of number of copies. In population and clinical genetics, it is important to have a complete overview of NumtS quantity and location. Searching PubMed yields hundreds of papers regarding Human NumtS compilation. A comparison of published compilations clearly shows significant discrepancies among data, due both to unwise application of Bioinformatics methods and a not yet correctly assembled nuclear genome. To optimize quantification and localization of NumtS, we have produced a consensus compilation of Human NumtS obtained by applying various bioinformatics approaches.
Methods: Location and quantification of NumtS have been achieved by applying database similarity searching methods: different methods, Blastn, MegaBlast and BLAT, changing both parameters and database have been used. To verify the in silico predicted NumtS we are amplifying and sequencing them from DNA samples of different mitochondrial haplogroups.
Results: The obtained results have been compared, further analysed and checked against the already published compilations thus producing the Reference Human Numt Sequences (RHNumtS) compilation. The resulting NumtS are 190. At present we have sequenced 40 NumtS, those with lower score, thus demonstrating the efficiency of our in silico protocol.
Conclusions
The RHNumtS compilation represents a highly reliable reference basis on which designing a lab protocol to test the reality of each NumtS. We are designing the RHNumtS Compilation database structure for implementation in the HmtDB resource (www.hmtd.uniba.it) .

Abstract 1

- HOMOLOGOUS GENE FAMILIES DATABASES FOR COMPARATIVE GENOMICS -

Penel Simon[1], Duret Laurent[1], Gouy Manolo[1], Perriere Guy*[1]

- [1]*Laboratoire de Biométrie et Biologie Evolutive, University of Lyon ~ Villeurbanne ~ France*

1J) Biobanks (databases and knowledgebases)

Motivation: Since the availability of a huge number of sequences, comparative genomics is a central step in many sequence analysis studies. For instance, it is used to help identify regions of interest in DNA sequences, to study evolution at the molecular level (speciation events, gene duplications, whole genome duplication, etc.), to determine phylogeny of species or to predict the function of a new gene. In that context we developed a set of three homologous gene families databases that can be used in many aspects of comparative genomics. Those databases - HOVERGEN, HOGENOM and HOMOLENS - share the same architecture, and they include protein and nucleotide sequences, alignments and phylogenetic trees.
Methods: The databases are built using protein sequences from different sources. HOVERGEN contains vertebrate sequences taken from UniProt. HOGENOM is devoted to completely sequenced organisms, and its sequences come from various sources (Genome Reviews, JGI, Ensembl, Bacterial species from the NCBI, etc.) HOMOLENS is devoted to the completely sequenced eukaryotes found in Ensembl. For the three systems, protein sequences are clustered into homologous families, and then alignments and trees are built on those families. The large-scale similarity searches required to cluster sequences, as well as the massive alignments and tree computations, are performed on a cluster containing more than 2000 CPUs. Lastly, phylogenetic trees are reconciled with a reference species tree.
Results: The three databases can be fully downloaded from the PBIL (Pôle Bioinformatique Lyonnais) site (http://pbil.univ-lyon1.fr). On-line access is also provided through three different ways: query forms on the PBIL site, a general retrieval system (Query) and a devoted client-server graphical interface (FamFetch). The later can be used to perform tree-patterns based searches allowing, among other uses, to retrieve easily set of orthologous genes thanks to phylogenetic criteria.

Abstract 4

- MASSIVE NON NATURAL PROTEINS STRUCTURE PREDICTION USING GRID TECHNOLOGIES -

Minervini Giovanni[1], La Rocca Giuseppe[2], Evangelista Giuseppe[1], Luisi Pier Luigi[1], Polticelli Fabio*[1]

- [1]*Department of Biology, University Roma Tre ~ Rome ~ Italy* - [2]*INFN-Catania ~ Catania ~ Italy*

1l) Grid technologies and Web Services

Motivation: The number of natural proteins is an infinitesimal fraction of all the theoretically possible protein sequences. In fact, considering random protein sequences of only 100 amino acids it is possible to obtain 100^20 structurally different proteins. Thus, there is an enormous number of proteins never exploited by nature or, in other words, “never born proteins” (NBPs). A fundamental question in this regard is if the ensemble of natural proteins possesses peculiar properties in terms for example of thermodynamic, kinetic or functional properties. A key feature of natural proteins is the ability to form a stable and well defined three-dimensional structure. Thus, the structural study of NBPs can help to understand if natural protein sequences were selected during molecular evolution for their peculiar properties or if they are just the product of contingency. This problem cannot be approached experimentally, as this would require the structural characterization of a huge number of random proteins. Thus we chose to tackle the problem using a computational approach.
Methods: A random protein sequences library (2x10^4 sequences) was generated using the utility RandomBlast which produces random amino acid sequences with no significant similarity to natural proteins. The structural properties of NBPs were studied using the ab initio protein structure prediction software Rosetta (Rohl et al. Methods Enzymol. 2004; 383, 66-93). Given the highly computational demanding problem, the Rosetta software was ported in the EUChinaGRID infrastructure (http://www.euchinagrid.org) and a user friendly job submission environment was developed within the GENIUS Grid Portal (https://genius.ct.infn.it/). Protein structures generated were analysed in terms of secondary structure content, overall topology, surface/volume ratio, hydrophobic core composition, net charge.
Results: Results obtained indicate that the vast majority of NBPs, according to the Rosetta model, are characterized by a compact three-dimensional structure with a high secondary structure content. Structure compactness is comparable to that of natural proteins, suggesting similar stability. Deviations are observed in hydrophobic core composition, as NBPs appear to be richer in aromatic amino acids with respect to natural proteins. The results will be discussed in view of the evolutionary implications of NBPs properties both at the amino acid and nucleotide level.

Abstract 2

- THE GENIUS GRID PORTAL AND THE ROBOT CERTIFICATES: A NEW TOOL FOR E-SCIENCE -

Donvito Giacinto[1], Maggi Giorgio Pietro[1], Barbera Roberto[2], La Rocca Giuseppe*[3], Milanesi Luciano[4], Falzone Alberto[5]

- [1]INFN ~ Bari ~ Italy - [2]INFN and University ~ Catania ~ Italy - [3]INFN ~ Catania ~ Italy - [4]Instituto di Tecnologie Biomediche - CNR ~ Milan ~ Italy - [5]NICE S.r.l. ~ Cortanze (AT) ~ Italy

1l) Grid technologies and Web Services

Motivation: Grid technology, based on opens standards and protocols, is the computing model which allows users to share a wide pletora of distributed computational resources regardless of their geographical and Institutional location. Up to now, the high security policy requested to access the distributed computing resources is a rather big limiting factor to increase the usage of Grids to a wider community of users. Grid security is indeed based on the public key infrastructure of X.509 certificates and the procedure to get and manage those certificates is unfortunately not straightforward.
Methods: A notable step forward to increase the exploitation of this new paradigm, has recently been made with the adoption of robot certificates. These new certificates have been introduced to permit users to easily use Grid infrastructures for their research activity. The robot certificate, associated to the specific application a user wants to share with all the members of a given VO, can be installed on a smart card and used with a portal by everyone interested in running that application in a Grid environment using an user-friendly graphic interface. In this work, the EnginFrame framework of the GENIUS portal has been extended in order to support the new user's authentication based on the use of robot certificates stored on smart cards. When the smart card is inserted in the server where GENIUS is running, the portal will start generating a new user's proxy signed by the robot certificate, otherwise the normal authentication based on a dedicated Java applet will be performed. Once the proxy is generated the user is automatically redirected to the home page of the application associated with the certificate. Any other attempts to access to unauthorized applications will be blocked by the portal. Moreover, in order to enhance the security of the system an User Tracking System has also been introduced to register and monitor the most relevant actions performed by users.
Results: The work carried out and reported in this contribution is particular relevant for all users who are not familiar with personal digital certificates and the internals of the Grid Security Infrastructure. The valuable benefits introduced by robot certificates in e-Science can so be extended to users belonging to several scientific domains, providing an asset in raising Grid awareness to a larger number of potential users.

Abstract 7

- GPU ACCELERATED RNA-RNA INTERACTION ALGORITHM -

Rizk Guillaume*[1], Lavenier Dominique[1]

- [1]IRISA-Symbiose ~ Rennes ~ France

1D) Molecular structure prediction, modelling and dynamic

Motivation: Many bioinformatics studies require the analysis of RNA or DNA structures. Packages like Unafold (Markham, N. R. & Zuker, M. Nucleic Acids Res. 2005; 33, W577-W581) provide many tools to study secondary structures. However, the high computational complexity of these algorithms combined with the rapid increase of genomic data triggers the need of faster methods. Current approaches are (1) designing faster algorithms or (2) parallelizing work on multiprocessor systems. Here, we explore the use of graphics processing unit (GPU) to speed up these kind of computations, which possibly exhibits a higher performance/cost ratio than clusters. It has already been successfully used for the computation of the Smith-Waterman alignment (Svetlin A Manavski, Giorgio Valle BMC Bioinformatics 2008 9-S2). We propose to parallelize on GPU the hybrid function of the Unafold package, which computes the stability of the duplex formed by two RNA sequences.

Methods: For an efficient parallelization, GPU need thousands of independent tasks. Parts taking the most time are found via program profiling and are then re-written in a way to expose parallelism. Our GPU implementation uses both parallelism within a single computation of the algorithm and between several execution of the algorithm across multiple pairs of sequences.

Moreover, to achieve good performance the data needed by the algorithm have to be carefully dispatched in different memory spaces of the GPU, according to their size and their access pattern. Another difficulty comes from the need to reduce to a minimum the if-then-else control instructions of the GPU kernels as the GPU is a SIMD (single instruction multiple data) architecture.

Results: Experiments have been done on an octo-core platform (2*Xeon E5430 2.66GHz, 8 GB RAM) with two NVIDIA Tesla cards. We benchmark our GPU implementation on 26000 pairs of sequences of length 50,50 with one or two cards versus the CPU version of the algorithm from one to eight cores. Total time spent for the complete application are respectively 100, 13.1, 9.8 and 5.3 seconds for 1 core, 8 cores, 1 card and 2 cards. GPU are a competitive alternative : the price of a platform with two Tesla cards is about the same as a platform with 8 processors but with 2.5 times the performance. Similar algorithms are used in a wide array of functions, such as the computation of the secondary structure of a single sequence which might also be parallelizable efficiently.

Abstract 43

- THE INTERPRETATION OF PROTEIN STRUCTURES BASED ON GRAPH THEORY -

Habibi Mahnaz*[1], Eslahchi Changiz[1], Sadeghi Mehdy[1], Pezeshk Hamid[1]

- [1]*Faculty of Mathematics, Shahid Beheshti University, Tehran, Iran ~ Tehran ~ Iran*

1C) Proteomics

Motivation: The analysis of protein structure is a challenging problem in bioinformatic allowing detailed exploration of the biological function. There are several features of protein structure which help to predict the protein function.
The main goal of this paper is to understand notions of various geometric aspects of a protein by considering a protein structure as a graph. This approach enables us to calculate important geometric concepts such as packing density and atom accessible surface by investigating the graph properties.
In the current method, for calculating packing density, "Voronoi polyhedron" of a protein is considered. Furthermore, it is possible to define a closed polyhedron on the surface. Various algorithms are used to cause the probe to visit all possible points of contact with the model. The locus of either the centre of the probe or the tangent point to the model is recorded.
Methods: We introduce two new algorithms. The first algorithm creates the maximum polymer inside the sphere of radius R. Using the packing polymers; we determine the packing density of a molecule. The second is based on the graph theory. We determine the hydrophobic cores of a protein as a subgraph which have a large average degree.
Results: For the calculation of the packing density based on the position of a probe sphere, the radius of probe sphere has to approach to zero. But by decreasing the radius of probe ball the time of algorithm rapidly increases. The time complexity of our proposed graph theatrical approach is O(n2), where n is the number of residues of a protein. Applying this algorithm, packing density can be obtained without the calculation of solvent accessible surface and 3D coordinates.
In addition, we present a new algorithm of order O(n2) (where n is the number of atoms of a protein) to calculate molecular surface area of each atom and amino acids. Using these values, we obtained total molecular surface area of a protein and the amino acids which are located in the surface of protein. We show that the packing density value and total accessible surface of a protein are negatively correlated.

Abstract 49

- ENGINEDB: A REPOSITORY OF FUNCTIONAL ANALOGUES -

De Sario Giulia*[1], Donvito Giacinto[2], Tulipano Angelica[1], Maggi Giorgio[3], Gisel Andreas[1]

- [1]Istituto di Tecnologie Biomediche, Sede Bari, CNR ~ Bari ~ Italy - [2]INFN Bari ~ Bari ~ Italy - [3]Dipartimento Interateneo di Fisica, Università e Politecnico di Bari ~ Bari ~ Italy

1J) Biobanks (databases and knowledgebases)

Motivation: Up to now, more than 4,0 million gene products from more than 150000 different species have been described specifying their functions, the processes they are involved in and their cellular localization using a well defined and structured vocabulary, the Gene Ontology (GO). Finding gene products with similar functions or involved in similar biological processes within the same or between different organisms, not relying on the conventional sequence similarity method, is an approach to find analogous gene products, which have similar functions, but not necessarily similar sequences as homologous gene products. However comparing gene products functionalities according to the GO terminology is a very time consuming process.

Methods: ENGINE (gENe analoGue fINdEr) is a tool that parallelizes the search process and distributes the calculation over the computational GRID, splitting the process into many sub-processes (Tulipano et al., BMC Bioinformatics 2007; 8,329-342). We developed a new, more performing version of engine and a process to select the most significant functional analogous gene products. Further, the search results are stored in a relational database (engineDB) hosting the most important information validating the proposed functional analogy between different gene products. A graphical interface enables the user to visualize the proposed functional analogues for his gene product under investigation ordered by the level of calculated analogy. engineDB visualizes the value of the chi-square test we used for the comparison as a rating for the analogy, the GO terms of both compared gene products and the number of GO terms in common and not in common since those are the terms influencing the analogy calculation and important for the user to understand which functionalities made the gene products in comparison more or less similar.

Results: ENGINE has produced for every gene product stored in the GO database a list of potential functionally analogues within and between species using, in place of the sequence, the GO gene description. Those data are publicly available either through a search tool as a GUI to engineDB or as a database dump of the whole data set. The GUI offers to the end user several external links such as UniProt, ENSEMBL and RefSeq and to download specific data and further information such as sequence similarity and protein domain comparison giving a complete overview about the proposed functional analogues.

Abstract 23

- TOWARDS SEMANTIC INTEROPERABILITY OF BIOINFORMATICS TOOLS AND BIOLOGICAL DATABASES -

Pettifer Steve*[1], Sinnott James[1], Thorne Dave[1], McDermott Phil[1], Marsh James[1], Attwood Teresa[2]

- [1]School of Computer Science, Manchester University ~ Manchester ~ United Kingdom - [2]Faculty of Life Sciences, Manchester University ~ Manchester ~ United Kingdom

1K) Biological data integration

Motivation: In the biological sciences, the need to analyse vast amounts of information has become commonplace. Such large-scale analyses often involve drawing together data from a variety of different databases, held remotely on the Internet or locally on in-house servers. Supporting these tasks are ad hoc collections of data-manipulation tools, scripting languages and visualisation software, which are often combined in arcane ways to create cumbersome systems that have been customised for a particular purpose, and are consequently not readily adaptable to other uses. For many day-to-day bioinformatics tasks, the sizes of current databases, and the scale of the analyses necessary, now demand increasing levels of automation; nevertheless, the unique experience and intuition of human researchers is still required to interpret the end results in any meaningful biological way. Putting humans in the loop requires tools to support real-time interaction with these vast and complex data-sets. Numerous tools do exist for this purpose, but many do not have optimal interfaces, most are effectively isolated from other tools and databases owing to incompatible data formats, and many have limited real-time performance when applied to realistically large data-sets: much of the user's cognitive capacity is therefore focused on controlling the software and manipulating esoteric file formats rather than on performing the research.

Methods: To confront these issues, harnessing expertise in human computer interaction, high-performance rendering and distributed systems, and guided by bioinformaticians and end-user biologists, we are building re-usable software components that, together, create a toolkit that is both architecturally sound from a computing point of view, and addresses both user and developer requirements. Key to the system's usability is its direct exploitation of semantics, which, crucially, gives individual components knowledge of their own functionality and allows them to interoperate seamlessly, removing many of the existing barriers and bottlenecks from standard bioinformatics analyses.

Results: The toolkit, termed UTOPIA, is freely available from http://utopia.cs.man.ac.uk.

Abstract 12

- WHEN DATA INTEGRATION LEADS TO A NEW CONCEPT : THE ORPHAN ENZYMES -

Labedan Bernard[1], Lespinet Olivier*[1]

- [1]*Institut de Génétique et Microbiologie, Université Paris-Sud 11 ~ Orsay ~ France*

1K) Biological data integration

Motivation: Despite the current availability of more than two millions of protein sequences, almost 35% of the enzyme activities (EC numbers) defined by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology are not associated with any amino acid sequence in major public databases.
The presence of so many EC numbers without an associated sequence (orphan enzymes) appears rather surprising at a time where we are inundated by genomic data. Alleviating this problem of orphanity would be very helpful for the difficult task of annotating and/or reannotating genomes. At any rate, there is an urgent need to bridge this unwanted gap between biochemical knowledge and massive identification of coding sequences and we suggest that the whole community could contribute to this task.
Accordingly, we are proposing a dedicated web service to identify the encoding gene for the maximum number of sequence-less enzyme activities.
Methods: We retrieved data from various public databases (UniProtKB, IntEnz, PDB, BRENDA, KEGG) and we have organized them into ORENZA, an efficient relational data warehouse committed to the exploration of the orphan enzyme universe.
Results: To identify the putative sequences associated with orphan enzyme activities, we clearly need the help of a large array of experts. As a result, the ORENZA resource contains a friendly tool allowing people having sound knowledge about specific enzyme activities to make helpful suggestions online. Each suggestion appears as a new item on each EC number's individual files in ORENZA. If several experts agree on the same suggestion, it would be transmitted to the curators of UniProtKB with a high degree of confidence. If experts disagree, their different advices will be published as they have been set.

We hope that ORENZA will help to resolve a few of our startling results: 1, orphan enzymes are present at about the same proportion in every class and subclass of enzyme activities. 2. orphan enzymes are widely distributed in the main functional categories. This is the case, for instance, of a significant number of enzymes that are involved in various metabolic pathways, despite a multitude of groups worldwide that studied them intensively and extensively for many years. 3. Even model organisms contain orphan enzyme activities (e.g. 189 in E. coli, 225 in man).

Abstract 68

- INTEGRATING ERV SEQUENCE AND STRUCTURAL FEATURES WITH DAS AND EBIOX -

Martínez Barrio Álvaro*[1], Lagercrantz Erik[2], Sperber Göran O[3], Blomberg Jonas[4], Bongcam-Rudloff Erik[2]

- [1]The Linnaeus Centre for Bioinformatics, Uppsala University, Biomedical centre, P.O. Box 598, SE-75124 ~ Uppsala ~ Sweden - [2]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Biomedical centre, P.O. Box 597, SE-751 24 ~ Uppsala ~ Sweden - [3]Department of Neuroscience, Physiology, Uppsala University ~ Uppsala ~ Sweden - [4]Section of Virology, Department of Medical Sciences, Uppsala University ~ Uppsala ~ Sweden

1K) Biological data integration

Motivation: The Distributed Annotation System (DAS) is a protocol used to exchange biological information. The network distribution concept of the protocol makes possible the use of different DAS reference and annotation servers to combine biological sequence data with annotations in order to depict an integrated view of the data to the final user.
Methods: Here we present a DAS annotation server devised to provide information about the endogenous retroviruses (ERV) detected and annotated "in silico" by a specialized tool called RetroTectorTM. We describe the procedure to implement the necessary DAS 1.5 protocol commands to construct DAS annotation servers. We use our server to exemplify those steps. The data distribution is separated from visualization which is carried out by eBioX, a general and user-friendly open-source programme suite with multiple bioinformatics utilities.
Results: We apply the server to discuss the advantages of distributing ERV data using the DAS protocol. Some well characterised ERVs are shown for two different organisms. By doing this, we also demonstrate the modularity of a distributed protocol like DAS as a solution for combining annotations belonging to different species. Reference and annotation data servers are then used in combination with eBioX to provide a friendly visualization of ERVs as well.

Abstract 53

- COMPUTATIONAL ANNOTATION OF UTR CIS-REGULATORY MODULES THROUGH FREQUENT PATTERN MINING -

Turi Antonio*[1], Loglisci Corrado[1], Salvemini Eliana[1], Grillo Giorgio[2], Malerba Donato[1], D'Elia Domenica[2]

- [1]*Department of Computer Science, University of Bari ~ Bari ~ Italy* - [2]*Institute for Biomedical Technologies, CNR ~ Bari ~ Italy*

1H) Text and data mining

Motivation: The huge amount of data produced by genome sequencing projects has allowed to highlight information on the genetic content of many organisms in the form of lists of genes they can express. Although necessary, this knowledge is not sufficient to understand the mechanisms regulating many events underlying life (i.e., cell growth, differentiation, development). In this sense, it is crucial to decipher the control mechanisms ruling the expression of genome in time and space. To address this problem we have developed a bioinformatic approach based on the use of data mining techniques to detect frequent association of regulatory motifs in untranslated regions (UTRs) of transcripts in Metazoa. The idea is that of mining frequent combinations of translation regulatory motifs, since their significant co-occurrences could reveal functional relationships important for the post-transcriptional control of genome expression.
Methods: The experimentation has been carried out using as a test case UTRs sequences extracted from the MitoRes database, annotated with information available in UTRef and UTRsite databases and collected in a relational database named UTRminer, which supports the pattern mining procedure. The mining approach is two-stepped: first, patterns of regulatory motifs are extracted and annotated in the form of sequences of motifs with information on their sequence location and mutual distances (spacers), then the mutual distances are discretized and the most frequent sequences of motifs and spacers are discovered by means of an algorithm for sequence pattern mining. Frequent sequences have a support greater than a user-specified threshold and the procedure for the generation of frequent sequences is guaranteed to be complete.
Results: The UTR sequences analysed concern ten different species. The total number of analysed sequences is 3896, among which 1944 5'UTRs and 1952 3'UTRs. Frequent motifs patterns, generated at first step, have a complexity ranging from 2 to 3 (number of distinct motifs detected on the same UTR) in 5'UTRs and from 2 to 5 in 3'UTRs. Preliminary results based on the observations and comparative analysis of discovered sequential pattern add new insights to our knowledge about post-transcriptional regulatory mechanisms controlling genome expression, while demonstrating the effectiveness of the bioinformatics approach presented in supporting discovery of motifs patterns.

Abstract 78

- A BIOINFORMATICS KNOWLEDGE DISCOVERY APPLICATION FOR GRID COMPUTING -

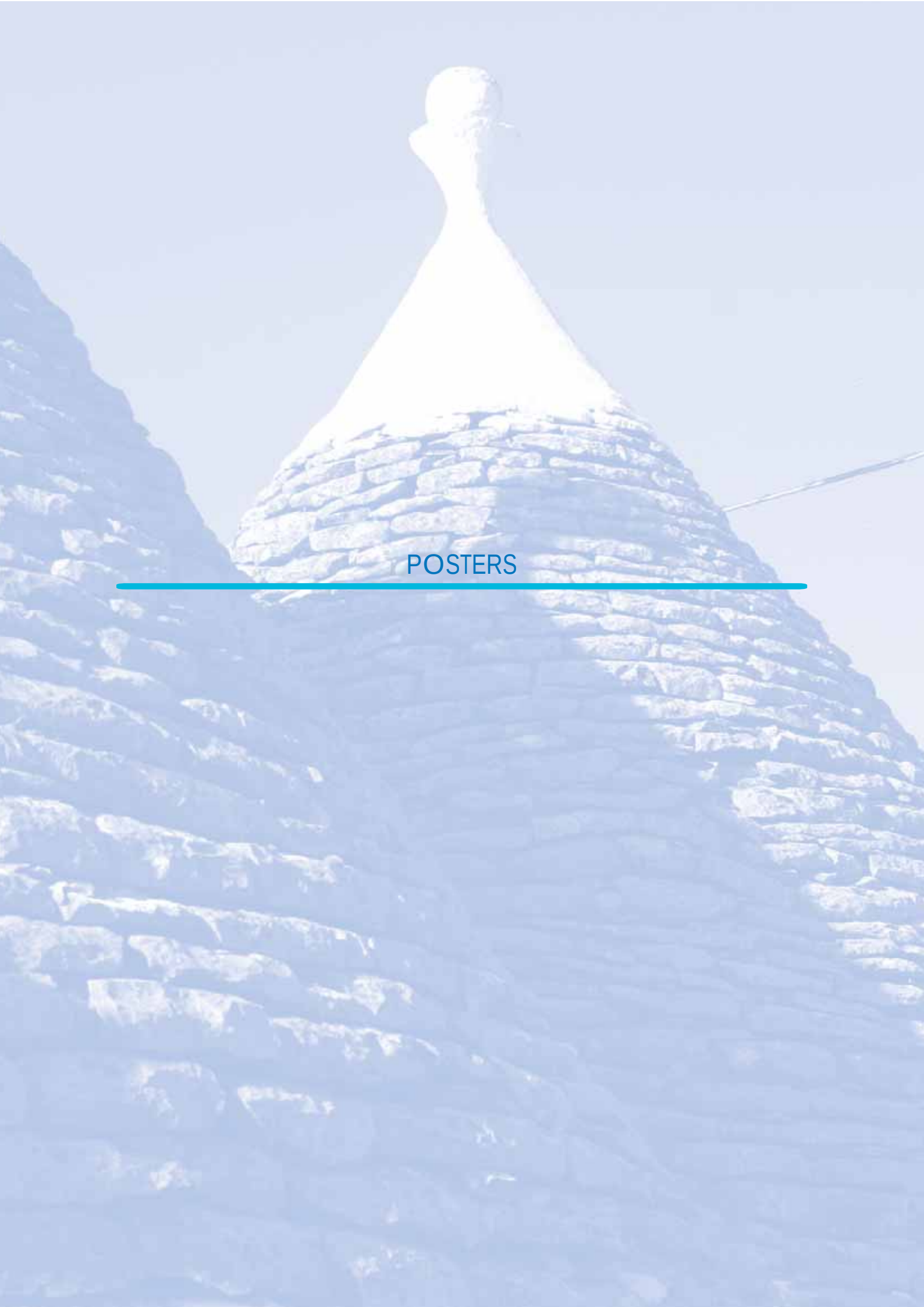
Castellano Marcello*[1], Mastronardi Giuseppe[1], Bellotti Roberto[2], Decataldo Giacinto[1], Pisciotta Luca[1], Tarricone Gianfranco[1]

- [1]*DEE - Dipartimento di Elettrotecnica ed Elettronica Politecnico di Bari, Bari, Italy* - [2]*Istituto Nazionale di Fisica Nucleare Sezione di Bari e Dipartimento Interateneo di Fisica "M. Merlin", Bari, Italy*

1H) Text and data mining

Motivation: : A fundamental activity in biomedical scientific research concerns the Knowledge Discovery process by large amount of biomedical information as documents and data. To have biomedical knowledge baggage more and more updated results a competitive advantage in the scientific progress and a best awareness in biomedical decision support. On the other hand, high performance computational infrastructures, such as Grid technologies, are emerging as available infrastructure to tackle, in principal, the problem of the intensive use of the Information and Communication resources in life science. However, from the application point of view, commodity software solutions are considering to investigate new biomedical information and the use of them is required by the end user. To exploit on a large scale the commodity software for obtain an ICT resources intensive use, a software adapter must be designed.
Methods: The method is based on a layered system software architecture to adapt the no grid based application in a distributed environment. In particular it is a solution to explane how to transform data intensive applications in Single Instruction Multidata stream applications with the aim to enhance bioinformatic application performance. We present a software prototype that interposes between the user's applications and grid middleware to the purpose to enable distribution of workload between grid nodes. The system, written in JAVA on the Globus Toolkit 4 grid middleware in a GNU/Linux computer grid, includes a graphical user interface, in order to access to a node research system, a load balancing system and a transfer optimizer to reduce communication costs. It has a modular structure to allow the end users to integrate and manage their applications. The prototype does not require that applications be written in a specific language or use specific libraries, it can be used with existing applications after a simple structure standardization. The load balancer analyzes the input data set and the selected computational nodes in order to provide a peer distribution of the workload on the grid. Thus, the transfer optimizer makes a compression of the data set and instruction set sending the compressed data to remote nodes. Finally, the grid computation starts.
Results: We present in details our case study, that is to say, a research of new evidences in biomedical field through a textual classification in terms of symptoms and pathologies recognition through a grid-based system on PubMed documents. Our application starts from 5000 medical scientific publications. It extracts all the possible symptoms and pathologies through Text Mining rules performed by GATE

4.0 using the software adapter here presented for computational grid environment. Then, it creates a number of associative rules that tie, in a probabilistic way, a series of symptoms to one or more pathologies using WEKA4WS grid based software. Finally, we evaluate the contribution in terms of time offered by the grid for the our biomedical application. The speedup factor refers to how much a parallel algorithm is faster than a corresponding sequential algorithm using grid. This factor has been determined for a node number from 1 to 30 for various different size documents. By graphs obtained it is noticed that the use of a grid system offers a profit in this application, this profit is more considerable in correspondence to the increase size document.



POSTERS

Abstract 5

- NORINE: A PUBLIC RESOURCE FOR NONRIBOSOMAL PEPTIDES -

Caboche Ségolène*[1], Pupin Maude[1], Leclère Valérie[2], Jacques Philippe[2], Kucherov Gregory[1]

- [1]LIFL (UMR USTL/CNRS 8022) - INRIA ~ Villeneuve d'Ascq ~ France - [2]ProBioGEM (UPRES EA 1026), Lille1 University ~ Villeneuve d'Ascq ~ France

1J) Biobanks (databases and knowledgebases)

Motivation: In micro-organisms, nonribosomal peptide synthesis is an alternative pathway that allows the production of small bioactive peptides from multienzymatic assembly lines called NonRibosomal Peptide Synthetases (NRPSs). The products, called NonRibosomal Peptides (NRPs), show a great diversity in composition, structure and function. They are short (two to about fifty amino acids), but can potentially contain more than 300 different amino acids (instead of twenty amino acids composing regular proteins). The NRP primary structure can be linear like in classical ribosomal peptides, but it is often more complex (totally or partially cyclic, branched or even poly-cyclic). The NRPs harbour a large spectrum of biological activities (e.g. antibiotics, antitumors, immunosuppressors). In spite of a great interest in NRPs due to their particularities and their important bioactivities, few computational resources and dedicated tools are currently available.
Methods: We have developed Norine, a public resource for NRPs. It contains more than 700 peptides and is still growing. Each peptide is annotated with various data collected from scientific publications. Those include the peptide name, its molecular weight, producer organisms, bibliographical references and links to other databases (UniProt and PubChem). The most original information stored in Norine is the NRP structure. We chose to represent the NRP structures at the amino acid level that reflects their biosynthesis, rather than to use the classical chemical representation. Indeed, the NRPSs successively incorporate complete amino acids rather than atoms. A friendly web interface was developed to search for NRPs according to various search criteria. In addition, users can search for a complete structure or a structural pattern (part of a structure possibly with jokers).
Results: Norine is the first resource entirely devoted to NRPs and is available at http://bioinfo.lifl.fr/norine/. We believe that Norine can have various usages in a wide range of related biological studies and can be useful in different applications of NRPs including very important applications in pharmacology. Indeed, we hope that Norine can contribute to biosynthetic engineering efforts to reprogram the NRP assembly lines, in particular because it makes possible systematic studies of the function-structure relationship of NRPs.

Abstract 6

- MOLECULAR DINAMIC OF ACTIVE SITE REGION OF MONILIPHOTHORA PERNICIOSA CHITIN SYNTHASE, THE AGENT OF WITCHES' BROOM DISEASE OF COCOA -

Souza Catiane[1], Taranto Alex*[2], Góes-Neto Aristóteles[3], Sandra Assis[4], Avery Mitchell[5]

- [1]Department of Biological Sciences, 2Graduate Program in Biotechnology (PPGBiotec - UEFS/FIOCRUZ-BA) ~ Feira de Santana ~ Brazil - [2]Graduate Program in Biotechnology (PPGBiotec - UEFS/FIOCRUZ-BA), Graduate Program in Vegetal Genetic Resources (RGV), Department of Health Sciences, State University of Feira de Santana ~ Feira de Santana ~ Brazil - [3]Department of Biological Sciences, 2Graduate Program in Biotechnology (PPGBiotec - UEFS/FIOCRUZ-BA) State University of Feira de Santana ~ Feira de Santana ~ Brazil - [4]1Department of Biological Sciences, 2Graduate Program in Biotechnology (PPGBiotec - UEFS/FIOCRUZ-BA), 3Graduate Program in Vegetal Genetic Resources (RGV), 4Department of Health Sciences, 1-4State University of Feira de Santana, Feira de Santana ~ Feira - [5]5Department of Medicinal Chemistry, 6National Center for Natural Products Research, and 7Department of Chemistry and Biochemistry, 5-7The University of Mississippi, MS, USA. ~ Oxford ~ United States

1D) Molecular structure prediction, modelling and dynamic

Motivation: The filamentous fungus Moniliophthora perniciosa (Stahel) Aime & Phillips-Mora is a hemibiotrophic Basidiomycota that causes witches' broom disease of cocoa (Theobroma cacao L.). This disease has resulted in a severe decrease in the Brazilian cocoa production, which changed the position of Brazil in the market from the second largest cocoa exporter to a cocoa importer. Chitin synthases (CHS) converts UDP-N-acetylglycosamine into chitin, the main component of the fungal cell wall. These glycosyltransferases have five different expression levels depending on the fungal life cycle stage. Class III chitin synthases act directly in the formation of the cellular wall and are responsible for most of the chitin synthesis in the cell, and are, therefore, a highly specific molecular target for drugs that could inhibit the growth and development of pathogenic fungi, since CHS is the immediate precursor of chitin and catalyzes an irreversible reaction.

Methods: After obtaining the protein sequence, a model of active site was constructed using Homology Modeling approach. The homologous sequence, with 29% identity, was used as template. The model was constructed by SWISS-MODEL, and refined by a set of Molecular Mechanics (MM) and Molecular Dynamics (MD) calculation, both using ff99 force field and implicit solvent model in Amber 8.0. The quality of resultant model was evaluated by PROCHECK 3.0, ANOLEA, and MD simulations.

Results: Ramachandran plot and MD simulations showed that the model has 98.4% of residues in the most favored regions with thermodynamic stability after 2.0 ns. The complete knowledge about the geometry of active site of CHS can be useful to develop new inhibitors against witches' broom disease

Abstract 8

- CONTRIBUTIONS OF GC ON GENE EXPRESSION: RECOGNIZING THE ROLES OF GC -

Arhondakis Stilianos*[1]

- [1]*Biomedical Research Foundation of Acedemy of Athens ~ Athens ~ Greece*

1B) Transcriptomics

Motivation: The effect of GC on expression levels is a topic of considerable evolutionary importance, and also has several practical implications for technologies that quantify expression levels. Several groups have addressed to study the influence of base composition on transcription levels in mammalian genomes observed via genome-wide technologies (sequencing- and hybridization-based techniques). Despite some variability among the reports, especially where they estimate a magnitude for this influence, a persisting trend has emerged: GC-rich genes tend to be expressed at higher levels than GC-poor genes.

Methods: Using publicly available collections of expression data from sequencing- (i.e., EST, MPSS) and hybridization-based (i.e., cDNA and short-oligo arrays) techniques, representing a wide range of human tissues, the contribution of GC on gene expression was investigated. When correlations were estimated using each of the available technologies, they were thoroughly assessed by checking for possible technology-specific or experimental effects. This was achieved by performing simple compositional analyses of the transcriptomes, and by taking into consideration known technology-specific limitations/deficits of each technique, leading to the detection of several cases of unreliable correlations, mostly negative ones.

Results: The cross-platform comparison presented here not only confirmed the persistence of positive correlations between base composition of human genes and expression level, but also detected several technology specific features that affect results. In addition, this work shows that the GC level and the compositional distributions of transcripts represent a very simple tool to assess biases in different technologies; furthermore the essentially invariant GC3 distribution of human genes can be considered as a reliable reference to assess gene representativity, and could play a useful role in biomedical or cross-platform comparison studies. In conclusion, a first conservative lower compositional border of the human transcriptomes is proposed, with mean GC3 of coding transcripts detected within any tissue typically above 55%.

Abstract 9

- CREATION OF A CULTURE COLLECTION DATABASE OF DIAZOTROPHIC AND PLANT GROWTH PROMOTER BACTERIA OF EMBRAPA SOJA -

Higashi Susan*[1], Hungria Mariangela[1], Barcellos Fernando Gomes[1]

-[1]*Soils Biotechnology, Embrapa Soja ~ Londrina ~ Brazil*

1J) Biobanks (databases and knowledgebases)

Motivation: : Culture collection maintenance is a primordial item if there is necessity of using microbial genetic resources. Therefore, these collections operate as ex-situ conservation centers of genetic resources and they are essential for metabolic and genetic diversity exploration. Culture collection can act as service collection keeping microbial resources and offering opportunities as biological material dispatch to research institutions, universities, etc, and information dispatch facilitating the use of microbial resources.

The support to these collections requires information storage of involved microorganisms. Consequently, databases become extremely important to keep the integrity and organization of these information. To use them, they must be well organized in databases with registries and documentation.

In this sense, a project was developed with the aim of creation of a bacterial culture collection of diazotrophic and plant growth promoter bacteria with agribusiness importance. Hence, the purpose of this work is to create a database to organize the information related to these microbial cultures.

Methods: This database project followed some steps (Conceptual, Logical and Physical Modeling) and specifics techniques. The methodology used proceeded as follows.

At first, the conceptual model was implemented after the study of data related to strains of symbiotic diazotrophic bacteria. The Entity-Relationship approach was selected because it is the most well known conceptual modeling technique.

Second, exhaustive tests of conceptual model was done and then the logical one was constructed.

Third, grounded in the logical model, the physical scheme was implemented with MySQL Database 5.0.16.

Afterwards, the first data were organized to the execution of exhaustive tests of the database. In this sense, the database was approved in all the assays and we could testify that the database structure proposed was really appropriate.

Holding the correct database structure it was possible to organize the database-user interface (web site). The site implementation was done by Renato Camara da Silva from LNCC and it is disposable at www.bmrc.lncc.br.

Results: Analyzing the results we concluded: the database created allowed appropriated information storage and organization, which was essential to supply the necessity of making information available.

Abstract 14

- "HARDY-WEINBERG KERNEL": A NEW SIMILARITY MEASURE FOR THE ANALYSIS OF GENETIC DATA IN COMPLEX PHENOTYPES -

Montesanto Alberto[1], Lagani Vincenzo*[2], Di Cianni Fausta[2], Conforti Domenico[3], Passarino Giuseppe[1]

- [1]Department of Cell Biology, University of Calabria ~ Rende ~ Italy - [2]CEntro di Supercalcolo per l'Ingegneria Computazionale (CESIC) - NEC Italia S.r.l. ~ Rende ~ Italy - [3]Dipartimento di Elettronica, Informatica e Sistemistica, University of Calabria ~ Rende ~ Italy

1H) Text and data mining

Motivation: Recent technological advances have led to the accumulation of a remarkable bulk of data on genetic polymorphisms. However the development of new statistical and informatics tools for the effective processing of these data has not been equally fast.
Machine Learning literature counts only a few examples of works focused on the development and application of data mining methods specifically devised for genetic polymorphisms analysis, although countless data - mining studies are focused on the analysis of other kinds of genetic data (e.g. gene expression data, proteomic sequences, etc.).
Aim of our work is to define a new similarity measure, the "Hardy-Weinberg kernel", specifically conceived for incorporating prior knowledge during the study of genetic datasets of marker genotypes.
The characteristic of "Hardy-Weinberg kernel" is that the similarity between genetic profiles is weighted by the estimates of gene frequencies at Hardy-Weinberg equilibrium in the population.
Methods: In order to compare the effectiveness of our similarity measure with respect to other "well established" kernels (Linear, Polynomial and Gaussian kernel), we applied Support Vector Machine (SVM) classification algorithms to a real-world dataset (Passarino et al., Hum Hered. 2006; 62, 213-220). The dataset had been collected in order to investigate the influence of the genetic variability of candidate genes on survival at old age. Several classification tasks were defined on the data, according to the analyses reported in the above cited paper. For each classification task SVM parameters were optimized through a cross validation procedure, while relevant features were selected via a forward - stepwise algorithm.
Results: Hardy-Weinberg kernel performances always matched or overcame other kernels performances, when used in conjunction with the forward stepwise feature selection algorithm. Experiments performed without feature selection demonstrated a significant decreasing of Hardy-Weinberg kernel performances. Interestingly, these experiments allowed us to discover the conditions under which our similarity measure is appropriate. In particular, Hardy-Weinberg kernel's poor performances may result from the inclusion of irrelevant genetic polymorphisms with rare alleles. A feature selection method based on such observation is currently under study.

Abstract 17

- A STRUCTURE-ACTIVITY STUDY OF CEPHALOSPORINS EMPLOYING SUPPORT VECTOR MACHINES -

Antelo-Collado Aurelio[1], Machin-Gonzalez Andy*[1], Hernandez-Diaz Yaikiel[1], Carrasco-Velar Ramon[2]

- [1]Faculty of Bioinformatic, University of Informatic Sciences ~ La Habana ~ Cuba - [2]Center of Pharmaceutical Chemistry ~ La Habana ~ Cuba

1D) Molecular structure prediction, modelling and dynamic

Motivation: : In 1988, Frere et. al stated that QSAR of ß-lactamic antibiotics were an impossible dream. The cephalosporins pertain to this compounds family and of course, it must be an impossible dream too. In 2003, one of the authors developed a regression and an artificial neural network model of cephalosporins (Carrasco, R., Phd. Thesis, ISBN 978-959-16-0646-4)
Methods: Now, we present a classification model of this compounds type with the same reported sample using Support Vectors Machines. To establish the models, topologic, topographic, quantum chemical and hybrid indices were employed as molecular descriptors of the 100 reported cephalosporins. Both c-svc and ?-svc were evaluated, varying the parameters c, ?, and ?. As kernel, RBF was selected.
Results: The best classification results (100%) were obtained with 11 independent variables and the c-svc machine with different c and ? pair values (1000, 0.1; 10000, 0.5; 1000, 0.5; 100, 0.5; 10, 0.5; 10000, 0.9; 100, 0.9, respectively). The best classification value for six variables was 94% also using c-svc machine. The application is implemented in Java language using the Libsvm library.

Abstract 18

- A NEW TOOL FOR THE PREDICTION OF BIOLOGICAL ACTIVITY USING COMPUTER NETWORK -

Carrasco-Velar Ramon*[1], Antelo-Collado Aurelio[1], Machin-Gonzalez Andy[1], Hernandez-Diaz Yaikiel[1], Prieto-Entenza Julio Omar[1], Rodríguez-León Alexis Rene[1], Pérez-Valdes Yunier Rene[1], Molina-Souto Yania[1], Mejías-César Yuleidys[1], Villaverde-Martínez Julio Antonio[1], Martí-Pérez Ileane[1]

- [1]*Faculty of Bioinformatic, University of Informatic Sciences ~ La Habana ~ Cuba*

1D) Molecular structure prediction, modelling and dynamic

Motivation: The University of Informatics Science has been designed to contribute to the informatization of Cuban society. In this sense, the Faculty of Bioinformatics, in cooperation with the Center of Pharmaceutical Chemistry is working together to developing a platform for the prediction of biological activity. The principal address to reach this objective is the optimal utilization of computational resources of the universities and the research centers.

Methods: The proposed system is implemented in Java for multiplatform use, and includes the following modules:

1. Interface adapted from JMOL visualization software.
2. Molecular editor based in JME applet.
3. Database of organic compounds supported on MySQL.
4. Graphic module for structural search in the database.
5. Module for calculation of topologic and topographic descriptors.
6. Module for quantum chemical calculations.
7. Fuzzy Logic module for data mining and to construct models.
8. Support Vector Machines module for data mining and to construct models.
9. Module to reduce the sample size.

Results: All modules are independent and the inclusion in the platform is done by plug-ins. A classification study in a set of cephalosporins using Support Vector Machines and a distributed quantum chemical structure optimization of 20000 compounds, as examples of the possibilities of the platform, are included.

Abstract 19

- ZEBRAFISH INTERACTOME IN ANALYSIS OF DIOXIN TOXICITY -

Alexeyenko Andrey*[1]

- [1]Stockhiolm Bioinformatics Center , Stockholm University ~ Stockholm ~ Sweden

1E) System Biology

Motivation: An integral analysis of environmental effects at the molecular level requires a global view that dynamically reflects functional changes of individual genes/proteins and their interactions. The latter can be observed in a dedicated well-controlled experiment. However, integrating this data into an interactome is hampered by ubiquitous false positive signals.
Methods: We recently created FunCoup - a public database of gene interaction networks. A deeply optimized technology integrated multiple datasets, such as physical protein interactions, mRNA and protein expression, TF and miRNA gene targeting etc. These multiple pieces of weaker evidence fused into confidently predicted interactions of several types (signaling links, co-membership in a protein complex etc.), and are presented on-line as genome-wide networks of eukaryotic organisms, from A. thaliana to H. sapiens (http://FunCoup.sbc.su.se). Thus, FunCoup augmented the notoriously incomplete interactome landscapes and exposed both regulatory and functional sides of gene networks. With a number of tools for graphical and tabular analysis, interaction components can be aligned and studied both inside and across species.
So far, no interactome have been integrated in fish. We employed FunCoup to compute such a network with data from orthologous proteins in eukaryotic organisms.
Another input was a global set of gene expression profiles in the developing zebrafish (days 1, 2, 3, 4, 5 after fertilization). In the 3-way ANOVA experimental design, dioxin-exposed embryos were compared to control samples. Beyond the traditional co-expression analysis, this enabled calculating functional links that reflect network patterns under the toxic versus normal conditions and tracing them during the embryonic development.
Results: Here we present an integral method of augmenting specific datasets with multi-facetted public information collected across many experiments and species. The interactome gained significant confidence and was employed in an analysis of aquatic toxicity in zebrafish. The network analysis scaled from the global view to individual functional components of the affected sub-networks. Due to the detailed time-course observations, one could specifically see original focal areas of the dioxin poisoning and their further propagation. We could also distinguish collective network components and individual genes that proved to be dioxin-resistant.

Abstract 20

- THE MYCOPLASMA CONJUNCTIVAE GENOME SEQUENCING, ANNOTATION AND ANALYSIS -

Calderon-Copete Sandra P.[1], Falquet Laurent*[1], Wigger Georges[2], Wunderlin Christof[2], Schmidheini Tobias[2], Frey Joachim[3]

- [1]*Swiss Institute of Bioinformatics ~ Lausanne ~ Switzerland* - [2]*Microsynth AG ~ Balgach ~ Switzerland* - [3]*Institute for Veterinary Bacteriology, University of Bern ~ Bern ~ Switzerland*

1A) Genomics

Motivation: The mollicute Mycoplasma conjunctivae is the aetiological agent leading to infectious keratoconjunctivitis (IKC) in domestic sheep and wild caprinae. Although this pathogen is relatively benign for domestic animals treated by antibiotics, it can lead wild animals to blindness and death. This is a major cause of death in the protected species in the Alps (e.g., Capra ibex, Rupicapra rupicapra).
Methods: The genome was sequenced using a combined technique of GS-FLX (454) and Sanger sequencing and annotated by an automatic pipeline that we designed, using several tools interconnected via PERL scripts. The resulting annotations are stored in a MYSQL database. This pipeline is likely to be adaptable to other prokaryotic species.
Results: The annotated sequence is then uploaded into the mollicutes database MolliGen (http://cbi.labri.fr/outils/molligen/) allowing for comparative genomics.
We present the results with several examples of genome comparison and analysis in search for biological targets (e.g., pathogenic proteins).

Abstract 25

- 'BRUKIN2D': A 2D VISUALIZATION AND COMPARISON TOOL FOR LC-MS DATA -

Tsagrasoulis Dimosthenis*[1], Zerefos Panagiotis[2], Loudos George[1], Vlahou Antonia[2], Baumann Marc[3], Kossida Sophia[1]

- [1]Bioinformatics & Medical Informatics Team, Biomedical Research Foundation of the Academy of Athens ~ Athens ~ Greece - [2]Biotechnology Division, Proteomics Unit, Biomedical Research Foundation of the Academy of Athens ~ Athens ~ Greece - [3]Protein Chemistry/Proteomics Laboratory and the Neuroscience Research Program Biomedicum Helsinki ~ Helsinki ~ Finland

1C) Proteomics

Motivation: Liquid Chromatography-Mass Spectrometry (LC-MS) is a commonly used technique to resolve complex protein mixtures. Visualization of large data sets produced from LC-MS, namely the chromatogram and the mass spectra that correspond to its compounds is the focus of this work.
Methods: Specifically, the in-house developed 'Brukin2D' software, built in Matlab 7.4, is presented here. It uses the compound data that is exported from Bruker 'DataAnalysis' program, and depicts the mean mass spectra of all the chromatogram compounds from one LC-MS run, in one 2D contour/density plot. Two contour plots from different chromatograph runs can then be viewed in the same window and automatically compared, in order to find their similarities and their differences.
Results: The results of the comparison can be examined through detailed mass quantification tables, while chromatogram compound statistics are also calculated during the procedure.

Abstract 26

- RETROTECTOR ONLINE, A RATIONAL TOOL FOR ANALYSIS OF RETROVIRAL ELEMENTS IN SMALL AND MEDIUM SIZE VERTEBRATE GENOMIC SEQUENCES -

Sperber Göran[1], Lövgren Anders[2], Eriksson Nils-Einar[2], Blomberg Jonas*[3]

- [1]Physiology Unit, Dept. of Neuroscience, Uppsala University, Uppsala, Sweden - [2]Linnaeus Centre for Bioinformatics, Biomedical Centre, Uppsala University, Uppsala, Sweden - [3]Section of Virology, Dept. of Medical Sciences, Uppsala University, Uppsala, Sweden

Motivation: The rapid accumulation of genomic information in databases necessitates rapid and specific algorithms for extracting biologically meaningful information. More or less complete retroviral sequences constitute 5-50% of vertebrate genomes, also called proviral or endogenous retroviral sequences; ERVs. After infecting the host, these retroviruses have integrated in germ line cells, and have then been carried in progeny genomes for up to several 100 million years. A better understanding of these sequences can have profound biological and medical consequences.

Methods: RetroTector© is a platform-independent JAVA program for identification and characterization of proviral sequences in vertebrate genomes. The full version (Sperber G et al, NAR 2007), requires a local installation with a MySQL database. Although not overly complicated, the installation may take some time. We have now created a “light” version of RetroTector©, (RetroTector online) which does not require specific installation procedures, and which can be accessed via the world wide web.

Results: RetroTector online (http://www.neuro.uu.se/fysiologi/jbgs) was implemented under the Batchelor web interface (A Lövgren et al, unpublished). It allows both file and FASTA cut-and-paste admission of sequences (5 to 1000 kilobases). Jobs are shown in an IP-number specific list. Results are downloadable as text files, and can be viewed with a stand-alone program, RetroTectorViewer.jar (downloadable from the same site), which has the full graphical capabilities of the basic RetroTector© program. Thus, a detailed analysis of any retroviral sequences found in the submitted sequence is graphically presented, and can be exported in standard formats. With the current server, a complete analysis of a 1 Megabase sequence is complete in under 10 minutes. It is possible to mask nonretroviral repetitive sequences in the submitted sequence before analysis, using host genome specific “brooms”. This increases the specificity of the analysis.

Conclusion: RetroTector online is a rational tool for retrovirological and genomic work.

Abstract 27

- A SCIENTIFIC WORKFLOW APPROACH FOR THE INTEGRATION OF BIOINFORMATICS TOOLS -

Han Youngmahn*[1], Cho Yongseong[1], Lee Sang-Joo[1]

- [1]*Korea Institute of Science and Technology Information ~ Taejon ~ Korea South*

1l) Grid technologies and Web Services

Motivation: Thanks to the rapid development of computer science and information technologies, biology work is no longer restricted to test tubes, petri dishes and pipettes. Many questions in biological research may best be answered by using extensive computational tools and resources. In the past decade "Big Science" such as the Human Genome Project has generated a vast knowledge explosion in biological field. The study of bioinformatics which emerged only a century ago has attracted vast attention in biological research by developing and utilizing a huge abundance of computer applications and statistical techniques to acquire, store, organize, analyze and visualize biological data and to facilitate biological research. However, the abundancy of bioinformatics resources brings in common problems, such as heterogeneity and incompatibility. Scientists find it slow, cumbersome and labor-intensive to establish the connections across different resources. The integration of heterogeneous bioinformatics services becomes emergent and of immense importance in this area.

Methods: Many bioinformatics studies usually require sequential analysis. For example, creating a phylogenetic tree using base or amino acid sequences consists of step by step processes, including sequence homology searching using BLAST program, multiple sequence alignment using ClustalW, editing aligned results with biological editing programs such as BioEdit or GeneDoc, and finally, creating phylogenetic trees using tree-building programs such as PHYLIP or PAUP. Thus, bioinformatics workflow system can best be an approach for effective integration of bioinformatics resources and providing seamless interfaces to facilitate bioinformatics analysis works.

Results: We have developed Bioworks system as an automated framework which enables to easily construct and execute the workflow model of complex bioinformatics analysis processes. Bioworks is based on client-server architecture. The client application provides a graphical user interface for constructing a workflow model of complex biological analysis processes and reporting intermediate results of each analysis process. The server engine not just automates the execution of workflow models, but also mitigates any interoperability issues among the bioinformatics services by the predefined data converting rules.

Abstract 28

- GIBA: A CLUSTERING TOOL FOR DETECTING PROTEIN COMPLEXES -

Moschopoulos Charalampos*[1], Pavlopoulos Giorgos[2], Likothanassis Spiridon[3], Kossida Sofia[1]

- [1]*Bioinformatics & Medical Informatics Team, Biomedical Research Foundation of the Academy of Athens ~ Athens ~ Greece* - [2]*European Molecular Biology Laboratory ~ Heidelberg ~ Germany* - [3]*Department of Computer Engineering & Informatics, University of Patras ~ Patra ~ Greece*

1C) Proteomics

Motivation: The study of protein interactions has been vital to the understanding of how proteins function within the cell. In addition, small group of proteins that interact with each other and are stable over time, called protein complexes, are extremely significant units for the harmonic function of the cells and can also provide information about the prediction of unknown proteins that participate in a protein complex.

Recently, new high - throughput methods such as microarrays, yeast two hybrid system, phage display and mass spectrometry generate enormous datasets of protein - protein interactions. Nevertheless, these methods suffer from a large error rate, where many protein interactions that exist in an organism are not recorded and yield many false positives. Moreover, only a small fraction of protein complexes has been experimentally determined due to the disability of these methods to detect all the proteins composing the under question complexes.

For these reasons, the use of computational methods in order to increase the quality of information of the biological methods and to detect protein complexes is essential. Due to the large number of protein interactions, the computational methods use the model of a graph called protein interaction graph. In such a graph, the vertices represent the proteins of an organism and the edges, the interactions between the proteins. Usually, these graphs are undirected and unweighted.

Methods: In this report, we present a new two step methodology for dealing with the protein complex detection problem. Initially, a clustering algorithm is used such as MCL, RNSC or affinity propagation. In the second step, the results are filtered based either on individual or on combination of 4 different methods (density, haircut operation, best neighbour and cutting edge). Our methodology is implemented in a user friendly tool, where the user can choose the algorithm of his preference.

Results: Extensive experiments were performed in 7 different datasets which were either derived from individual experiments or from online databases. Furthermore, we used 5 different methods in order to evaluate, as objectively as possible, the results of our experiments. We compared our method with 4 other algorithms (Mcode, HCS, SideS and RNSC with the filtering proposed from its creators) and we conclude which algorithmic combination produces the best results.

Abstract 29

- TPARVADB: A DATABASE TO SUPPORT THEILERIA PARVA VACCINE DEVELOPMENT -

VIsendi Paul[1], Bulimo Wallace[2], Ng'ang'a Wanjiku[3], Bishop Richard[4], de Villiers Etienne P.*[4]

- [1]1Center for Biotechnology and Bioinformatics, University of Nairobi ~ Nairobi ~ Kenya - [2]US Army Medical Research Unit - Kenya ~ Nairobi ~ Kenya - [3]School of Computing and Informatics, University of Nairobi ~ Nairobi ~ Kenya -[4]International Livestock Research Institute ~ Nairobi ~ Kenya

1J) Biobanks (databases and knowledgebases)

Motivation: development for East Coast Fever have been hindered due to lack of a user-friendly and specific T. parva database. We sought to develop TparvaDB, to provide a comprehensive resource to facilitate research in the development of an ECF vaccine by providing a single user-friendly database of all genome and related data for Theileria parva.
Methods: TparvaDB is based on the Generic Model Organism Database (GMOD) platform. Data was migrated from the original Manatee annotation database, reformatted, and reconfigured to populate TparvaDB. The Apollo annotation workbench and a comparative genomics pipeline were included to add functionality to TparvaDB.
Results: We have developed TparvaDB, an integrated database for T. parva based on GMOD. TparvaDB houses full genome sequences, Expressed Sequence Tags (ESTs), Massivelly Parallel Signature Sequencing (MPSS) data, vaccine candidate gene and other related data. TparvaDB consists of a web page generated using the GMOD web tool, a database implemented in MySQL using the Chado schema. Genomic EST and MPSS data were downloaded from the Manatee annotation database as MySQL dump files and converted into Gene Feature Format (GFF), for loading into Chado. TparvaDB was extended to incorporate the Apollo annotation workbench to facilitate subsequent online annotation. The database was designed to integrate data from other apicomplexan species such as T. annulata and P. falciparum to facilitate for comparative analysis. TparvaDB will greatly enhance the ongoing efforts in ECF vaccine and diagnostic.

Abstract 31

- PHD-SNP1.0: A WEB SERVER FOR THE PREDICTION OF HUMAN GENETIC DISEASES ASSOCIATED TO MISSENSE SINGLE NUCLEOTIDE POLYMORPHISMS -

Calabrese Remo*[1], Capriotti Emidio[2], Casadio Rita[1]

- [1]*Biocomputing Group, University of Bologna ~ Bologna ~ Italy* - [2]*Structural Genomics Unit, Department of Bioinformatics, Centro de Investigacion Principe Felipe (CIPF) ~ Valencia ~ Spain*

1C) Proteomics

Motivation: TSingle Nucleotide Polymorphisms (SNPs) are the most frequent type of genetic variation in humans (Collins et al., 1998). Great interest is focused on non-synonymous coding SNPs (nscSNPs) that are responsible of protein single point mutation, since mutations occurring in coding regions may have a larger effect on gene functionality. The possibility of retrieving a large dataset of annotated SNPs from the Swiss-Prot Database (Boeckmann et al., 2003) prompted the application of machine learning techniques to predict the insurgence of human diseases due to single point protein mutation starting from the protein sequence (Capriotti et al 2006).

Methods: We developed a method based on support vector machines (SVMs) that starting from the protein sequence information and evolutionary information, when available, can predict whether a new phenotype derived from a nscSNP can be related to a genetic disease in humans. The system is based on two different SVMs, one is a SVM-sequence that performs predictions relying on sequence information alone, the other is a SVM-profile performing predictions on profile features when evolutionary information is available. Merging in a unique framework the two SVMs we get a hybrid predictive method.

Results: On a recent dataset (April 2008) of 34314 mutations, 48% of which are disease related, out of 7351 proteins, we show that our method can reach more than 72% accuracy (with a correlation coefficient of 45%) in the specific task of predicting whether a single point mutation can be disease related or not. Although based on few informations, our system reaches the same accuracy, with a higher correlation, of the other web-available predictors implementing different approaches (Ramensky et al., 2002 ; Ng and Henikoff, 2003). We design a web server integrating our SVM models, called Predictor of human Deleterious Single Nucleotide Polymorphisms (PhD-SNP). The server is a user friendly resource that gives the possibility of retrieving predictions via e-mail. The submission form is very simple and the user has to paste the query sequence, to select the mutation position and the mutated residue in relative input boxes; furthermore he can choose the predictive method. Best results are obtained when evolutionary information is available and when it is possible to perform predictions using the hybrid predictive method.

Abstract 34

- COMPARISON OF ABC TRANSPORTER GENES OF PLASMODIUM SPECIES: A SEARCH IN ELUCIDATING NEW DISCOVERIES TOWARD MALARIA PARASITE ERADICATION -

OLUWAGBEMI OLUGBENGA*[1], Yah Clarence[2], Adebiyi Ezekiel[1]

- [1]*Department of Computer and Information Sciences, Covenant University ~ Ota ~ Nigeria* - [2]*Department of Biological Sciences, Covenant University ~ Ota ~ Nigeria*

1A) Genomics

Motivation: Malaria is a major public health problem associated with high mortality and morbidity rates in Sub- Saharan countries, with a spectrum of systemic complications ranging from mild and self-limiting to life-threatening. Drug resistance has posed a major problem in malaria control and occurs in areas endemic of malaria parasite.
Methods: The current research engaged the use of bioinformatics approach to seek new chemotherapeutic strategies in analyzing and proffering solutions to malaria control and eradication. Three Plasmodium species: P. berghei, P. chabaudi, P. falciparum resistance genes ((ABC transporter) putative genes) were compared using the Atermis comparative tool (ACT).
Results: There was slight variation in the up/down stream alignment within the genes likewise their phylogenetic relationships. This therefore showed that same resistance genes within a population of the same site may vary within the same drug.

Keywords: ABC transporter genes, Plasmodium, eradication

Abstract 35

- THE EFFECT OF SINGLE MUTATIONS ON THE CARRIER ACTIVITY OF THE DICARBOXYLATE CARRIER (DIC) OF S. CEREVISIAE: IN VITRO VALIDATION OF PREDICTIONS OF PROTEIN STABILITY CHANGES -

Ferramosca Alessandra *[1], Mirto Luisa[2], Tasco Gianluca[3], Tartarini Daniele[2], Zara Vincenzo[1], Aloisio Giovanni[2], Casadio Rita[3]

- [1]Di.S.Te.B.A ~ University of Salento, Lecce ~ Italy - [2]SPACI Consortium, University of Salento, Lecce ~ & NNL/CNR-INFM, Lecce ~ Italy - [3]Biocomputing Group ~ University of Bologna ~ Italy

1D) Molecular structure prediction, modelling and dynamic

Motivation: A basic problem of structural biochemistry studies is to which extent a mutation will affect the stability, and then the function of the protein. From this point of view, an important aspect regards the function of metabolite mitochondrial carriers, in relation to the role that such proteins cover in some mitochondrial pathologies. Sequence studies have shown that the PX(D/E)XX(K/R) signature is characteristic of all mitochondrial carriers, and possibly involved in the transition from the open to closed states, corresponding to the active/inactive state of the carrier.

In this study our approach is to combine predictions and experimental validation adopting as a test case the dicarboxylate carrier (DIC) of S. cerevisiae. Because DIC structure is unknown, we integrated a routinely expert dependent strategy in a automatic tool based on a Grid infrastructure to facilitate the generation of carrier models, including site directed mutagenesis.

Methods: Transport activity of mutagenized proteins was measured in vitro in reconstituted systems. In parallel in silico experiments protein stability was predicted by using I-Mutant3, available at http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi. DIC structure was computed by homology modelling using a service that integrates several software and data involved in a Grid infrastructure, available at https://sara.unile.it/cgi-bin/bioinfo/enter.

Results: Protein stability changes upon mutations can be both predicted and/or tested in vitro, by monitoring the function of mutagenized proteins in reconstituted systems. We found that mutations in the DIC carrier signature reduced or blocked nearly completely the protein transport activity, in agreement with the functional role of the carrier motif. Then the question is whether the observed effect on the activity can be also related to a change on protein stability upon mutation. This was tackled by computational methods. According to our predictions protein destabilization would correlate with the observed loss of protein activity. Our data therefore corroborate the finding that single point mutations may hamper protein stability when placed in functionally relevant structural position, and further add to the possibility of predicting a priori whether the mutation destabilize the protein.

Abstract 38

- ANALYSING GENE EXPRESSION PATTERNS IN THE METABOLIC NETWORK OF NEUROBLASTOMA TUMOURS WITH WAVELET TRANSFORMS -

Schramm Gunnar[1], Gaarz Andrea[1], Seitz Hanna[2], Oswald Marcus[2], Eils Roland[2], Konig Rainer*[2]

- [1]IPMB, Bioquant, University of Heidelberg ~ Heidelberg ~ Germany - [2]Institute of Computer Science, University of Heidelberg ~ Heidelberg ~ Germany

1B) Transcriptomics

Motivation: Neuroblastoma tumours show a very heterogeneous clinical picture ranging from rapid growth with fatal outcome to spontaneous regression or differentiation into benign ganglioneuroma. Specific treatment is crucial and can be supported by understanding the molecular functionality of the tumour.
Methods: We have performed gene expression profiling with microarrays for this tumour and mapped it onto the metabolic network to define biochemical pathways that show a discriminative regulation behaviour between the different tumour types. We used an established method (König et. al., BMC Bioinformatics, 2006) that calculates Haar wavelet transforms on adjacency matrices of the network. These wavelet transforms are normally applied to pattern recognition on images. The method was further developed applying heuristic solutions for the grid arrangement problem.
Results: With this we were able to evaluate all KEGG maps in respect to their ability to discriminate neuroblastoma tumours of patients with favourable and unfavourable outcome. The most significant patterns were found for e.g. purine, pyrimidine metabolism, and one-carbon-pool-by-folate indicating increased nucleotide production for proliferation. Furthermore, we found an interesting significant pattern in the glutamate metabolism indicating a potential switch like behavior of the aggressive tumour. Especially the glutamate and one carbon pool metabolism suit well for further analysis by drug treatment and knock down experiments in the laboratory to define drug targets for the aggressive tumours.

Abstract 39

- INFERRING THE ASSOCIATION OF GENOMIC EXPRESSION AND COPY NUMBER VARIATION -

Orsini Massimiliano[1], Capobianco Enrico*[1]

-[1]*CRS4 Bioinformatics Laboratory ~ Pula (CA) ~ Italy*

1K) Biological data integration

Motivation: MicroRNAs are small non-coding RNAs (~22 nucleotides) regulating target gene expression via cleavage or translational inhibition. Lu et al (Nature, 2005) showed that most of microRNAs have differential expression (gain or loss) values in tumour samples, and other studies have mapped tissue-specific cancer signatures (Volinia et al, PNAS, 2006). The location of microRNAs in relation to copy number variation (CNV) has also been recently addressed (Lamy e al, Brit J Cancer, 2006) to reveal possible possible correlation patterns in three types of cancers, but unexpectedly clear signatures could not be established. We hypothesize the same kind of possible associations in brain cancers, and suggest a model representation to specify and test relationships among variables. The rationale is that cancers are also characterized by chromosomal aberration that may be predictive of disease outcome, and many by somatically acquired copy number changes, including loss of heterozygosity (LOH) at multiple loci. These aberration are strongly associated with clinical phenotype including patient outcome.

Methods: The model approach starts from a general genetic signature G which may depend on two general factors, expression X and variation Y, such that G = A(X,Y) + E. We thus propose a flexible stochastic model where observable and latent variables can be combined, specified and tested. Nowadays, high throughput array-based methods deliver huge amounts of data for expression, genotyping and CNV leading to a parallel assessment of multiple genomic alterations. We have developed a micro-target warehousing system by tissue, miRWare, aimed to allow coordinated inference, and a tool for automatic annotation of regions highlighted in CNV experiments, Magellano which returns gene structure, SNPs, disease association and expression profiles of each gene in a selected genomic region.

Results: As an example from Magellano, the region 3p14.4-p25.3 (often identified in neuroblastoma) contains 599 genes with 22 microRNAs in part differentially expressed in neural tissue cancer. Brain cancers offer a wealth of data due to the richness of microRNA expression and the tissue-specific stem cell differentiation. We have thus reduced the scale of the analysis compared to Lamy et al who looked at colon, prostate and bladder cancers, and have instead emphasized brain-specific characterization from cell line evidence.

Abstract 40

- EMBOSS ON THE GRID WITH EMBOSS-GUI -

Valverde Jose R*[1]

- [1]*Scientific Computing Service ~ Madrid ~ Spain*

1l) Grid technologies and Web Services

Motivation: The availability of a Grid port of EMBOSS provides a ready solution for some common Bioinformatics tasks requiring large computing power as that provided by the Grid (specially analysis at the genomic level), but at a significant cost in learning the command line. A web interface that hides Grid and EMBOSS complexity can greatly empower users needing to perform these tasks.
Methods: Due to the nature of EMBOSS implementation, the easiest way to make these complex tasks easily usable is to exploit EMBOSS generic command interface,resulting in all EMBOSS applications being adapted for Grid use by acting only at one single level. We have analyzed various solutions to implement a web based GUI for EMBOSS, and finally settled for EMBOSS-Explorer (aka EMBOSS-GUI) as our initial target.
We have analyzed and tested various different implementation approaches, which finally led us to generate a fork of Luke McCarthy original project in order to satisfy Grid policy requirements.
Results: Deciding on an initial target for a web user interface to EMBOSS was a difficult decision were technical and subjective factors. Once we had settled on EMBOSS-GUI as our initial target, the initial adaptation to make it run jobs on the Grid was relatively simple, but in order to accommodate it to Grid policies we had to fork a new project to include user authentication.
Our experience using EMBOSS-GUI for Grid work shows that this solution is wanting in some features required for modern, advanced users and sheds light for future developments using other user interfaces. It also evidences the impact of current Grid policies on the way bioinformatics analysis are normally carried out.
In this work we discuss not only the technical problems faced and the solutions developed but also reflect on how other off line factors such as social, personal and vital issues, usability and policies affect development of modern Bioinformatics solutions.

Abstract 41

- TRANSCRIPTION FACTORS REGULATION OF HUMAN GENE NETWORK -

Krivosheev Ivan*[1], Du Lei[2]

- [1]Department of Bioinformatics, Harbin Institute of Technology ~ Harbin ~ China - [2]Department of Bioinformatics, Harbin Medical University ~ Harbin ~ China

1E) System Biology

Motivation: With increasing variety of molecular networks, such as regulatory, interaction etc, finding relations between structural features of these networks and their biological significance attracts attention of many researchers. While abundant quantity of studies represented connections between protein networks and topological parameters, few approaches have been applied to transcriptional regulations of coexpressed gene networks. Here, we focus on how the number of transcription factors (TFs) correlate with gene network properties.
Methods: We applied three graph-theoretical characteristics - node degree, betweenness centrality and pairwise disconnectivity - to the analysis of gene coexpression networks from HapMap human gene expression data. The networks were constructed by using ARACNE algorithm. We revealed hundreds of genes, each is subject to massive regulation by several TF. We examine the relationship between topological features of each gene and the number of TFs regulating it with the Spearman coorelation and statistical evaluation.
Results: We demonstrate that for human coexpressed gene networks, betweenness centrality and node degree are negatively correlated with according TF number. Our study provides global insights into the effects of TFs regulation in human gene interactions.

Abstract 44

- MOLMETH: THE MOLECULAR METHODS DATABASE -

Lagercrantz Erik[1], Oelrich Johan[2], Martinez Barrio Alvaro[3], Bongcam-Rudloff Erik*[1], Landegren Ulf[2]

- [1]Department of Animal Breeding and Genetics, SLU ~ Uppsala ~ Sweden - [2]Department of Genetics and Pathology, Uppsala University ~ Uppsala ~ Sweden - [3]Linnaeus Centre for Bioinformatics, Uppsala University ~ Uppsala ~ Sweden

1J) Biobanks (databases and knowledgebases)

Motivation: MolMeth, short for molecular methods, is a database system that catalogs laboratory protocols and methods for the life sciences. It is of particular value for large-scale applications in biobanks and systems biology, but also provides value in scientific communication about molecular procedure in general. It is designed to meet a growing need for structure in protocol specifications while offering convenience for contributors and easy access for end users. Structured protocols offer several advantages over current "flat file" protocol databases, such as allowing protocol presentation be adapted for different purposes. It also provides a foundation for automated reasoning regarding protocols.

Methods: The system presents itself as a web site for searching, retrieving and viewing protocols. Registered users can submit and modify their protocols, and modifications result in new versions with distinct, permanent URL:s. There is also a web service, which allows third party applications to retrieve structured versions of protocols.

The database stores various properties related to each protocol, including a unique accession number, information about materials and available suppliers, versioning information, user comments, references and related entries. The submitter of a protocol is allowed to specify a publication date and rudimentary access rights.

The basic structure of a protocol is modular, meaning that it can be built as a hierarchy of (sub-) protocols, combining steps into different protocols without duplicating common parts. Each protocol is also viewed as a function, which transforms an input to some output, specified using well defined ontologies.

Results: The modularity saves effort for authors when protocols have steps in common, or when a protocol is part of another, more extensive protocol. Protocols that are split into modules are still presented with contiguous instructions in a hierarchical list of steps, adapted for a specific setting if desired.

The MolMeth team hopes that that the computational abilities arising from structured protocols will allow the system to automatically suggest steps for protocol authors or, given a start condition and a goal, even suggest entire protocols by combining smaller protocols from the database. It is already clear that structured protocols will play a role in the development of harmonised standards in several pan-European research infrastructures.

Read more: www.molmeth.org

Abstract 45

- A GREEDY ALGORITHM FOR HAPLOTYPE INFERENCE BY PURE PARSIMONY -

Poormohammadi Hadi*[1], Eslahchi Changiz[1], Kargar Mehdi[2], Pirhaji Leila[3], Pezeshk Hamid[4], Sadeghi Mehdi[5]

- [1]Faculty of Mathematics, Shahid Beheshti University, Tehran, Iran ~ Tehran ~ Iran - [2]Department of Computer Engineering, Sharif University of Technology, Tehran, Iran ~ Tehran ~ Iran - [3]Department of Biotechnology, College of Science, University of Tehran, Tehran, Iran ~ Tehran ~ Iran - [4]Center of Excellence in Biomathematics, School of Mathematics, Statistics and Computer Sciences ~ Tehran ~ Iran -[5]National Institute for Genetic Engineering and Biotechnology, Tehran, Iran ~ Tehran ~ Iran

1A) Genomics

Motivation: Haplotype are important information in the study of complex diseases and drug design. However, due to technological limitations, genotype data rather than haplotype are usually obtained. Thus, haplotype inference from genotype data using computational methods is of interest for many researchers.
Methods: There are several models for inferring haplotypes. One of the most important models is haplotype inference by pure parsimony (HIPP), consisting of finding the minimum number of haplotypes that can resolve all given genotypes. HIPP is an NP-hard problem. In this paper we propose a new greedy algorithm for this problem. The greedy algorithm accurately predicts an efficient Haplotype for inferring the remaining genotypes in each step.
Results: Results of applying our algorithm on a variety of biological and simulated data show that it is very effective with a high accuracy compared to other algorithms.
Also a new measure for evaluating the effectiveness of the algorithms is introduced. This measure is based on the pure parsimony approach which seeks to find the minimum number of haplotypes for resolving the input genotypes.

Abstract 47

- GENE REGULATORY NETWORKS IN BACTERIOPHAGES -

Klucar Lubos*[1], Stano Matej[1], Hajduk Matus[1]

- [1]*Institute of Molecular Biology SAS ~ Bratislava ~ Slovakia*

1E) System Biology

Motivation: Complex approach to the study of biological systems is of increasing importance. In order to achieve this task immense amounts of data is needed, which requires computer preprocessing of these data. An important role in this process play biological databases which store records in an easily accessible form for whole scientific community. A system approach to biological data integration and consequent construction of predictive models is the most valuable outcome of systems biology.
Methods: phiSITE database is built upon the MySQL database (version 4.0) and PHP (version 4.3). For visualization of Gene Regulatory Networks BioTapestry program (version 2.1.0, www.biotapestry.org) was used. The simulations of Gene Regulatory networks (GRNs) were run on the Dizzy simulation engine (version 1.11.4, http://magnet.systemsbiology.net/software/Dizzy/).
Results: We have developed phiSITE - database of gene regulation in bacteriophages (www.phisite.org). To date it contains detailed information about almost 500 cis-regulatory elements from 42 bacteriophages. Based on the phiSITE data we defined GRNs for four phages: Enterobacteria phage lambda, Mycoplasma virus P1, Enterobacteria phage Mu and Bacillus phage GA-1 used for visualization in BioTapestry viewer. Next, we created a scaffold of gene regulatory network model of Enterobacteria phage lambda. The model is written in SBML and it is simplified to the level of transcriptional control. We omitted Paq and Pi promoters since they do not influence the simulation significantly. Because phiSITE database does not contain the exact kinetic data of transcriptional processes, these were not specified in the model (experimentally obtained values can be added to the model when desired). We launched the model under two different conditions. In the first test, there were no phage proteins present - values for initial amounts of all protein species were set to 0. In the second test we simulated lysogeny conditions by high initial concentrations of CII (100) and CIII (40) proteins. Both, deterministic and stochastic simulations in Dizzy simulator, produced similar results that correspond with progress of lambda infection in a living bacterial cells even with the lack of exact kinetic data. This fact refers to the robust nature of lambda gene regulation. This work was funded by APVT-51-025044 grant from Slovak Research and Development Agency.

Abstract 48

- THE DIVERGENCE OF EXPRESSION PATTERNS OF DUPLICATED GENES IN ORYZA SATIVA -

Li Zhe[1], Zhang He[1], Gao Ge[1], Luo Jingchu*[1]

- [1]*College of Life Sciences, Peking University ~ Beijing ~ China*

1A) Genomics

Motivation: Genome-wide duplication is ubiquitous during the diversification of the angiosperms, and gene duplication is one of the most important mechanisms for evolutionary novelties. As an indicator of functional evolution, the divergence of expression patterns following duplication events have drawn great attention in recent studies.
Methods: Here, using large-scale whole-genome microarray data, we systematically analyzed expression divergence of genes arising through whole-genome and small-scale duplication events, in the rice (Oryza sativa ssp. japonica) genome.
Results: Our results shown that duplicates created by whole-genome duplication that retained in colinear segments shown more similar expression patterns than those created by small-scale duplication. We propose that such difference could largely be explained by sequence divergence. Further analysis suggested sequence divergence plays important roles in modeling the divergence of expression patterns, and the mode of duplication had less effect on the divergence of expression patterns.

Abstract 50

- PHYLOGENETIC DATABASE QUERY SYSTEM FOR SPECIES DETERMINATION -

Vicario Saverio*[1]

-[1]CNR - ITB ~ Bari ~ Italy

1F) Molecular biodiversity, DNA Barcode and metagenomics

Motivation: Species diagnose is still a complex and knowledge intensive activity. This does not allow society at large to benefit of all advantage of correct systematic information. This limits the recognition of the benefit and interest of systematics. The barcode initiative try to set up a standardized and automatic protocol of species diagnose based on standardized molecular sequences and a database of sequences belonging to known species. Here we implement a protocol that uses explicit phylogenetic inference to treat barcode data for identification. The protocol try to balance the need of fast answer typical of database query tools with the need to have a robust phylogenetic inference that would give answer in term of probability. We explored the efficacy of this protocol under various conditions of speciation and sampling

Methods: The data that we use to test our protocol are the set of the barcode quality sequences available on GenBank/EMBL/DDBJ for lepidopterans. This 3523 sequences were organized in a database grouped few groups based on a priori phylogenetic knowledge. For each group a Bayesian inference based on realistic evolutionary model was performed. A hierarchical query system was build to place an unknown sequence first in the correct group and then in the phylogenetic tree of the chosen group, taking account the uncertainty of the inference. The placement in the pre-computed phylogenetic trees was based on three different methodologies. The system was tested in a cross validation framework and the different topological placement methods compared.

Results: The protocol performed quite well with overall high accuracy (>.95) although error concentrate in species with problematic phylogenetic pattern (polyphyly or paraphyly of the species sequences) or species with very few representatives and distant sister taxa.

in conclusion the methods although rather efficient is very dependent from the phylogenetic pattern for the marker under examination.

Abstract 51

- COMPARING THE BIOCHEMICAL NETWORKS OF HUMAN AND RODENT CELLS INFECTED WITH DIFFERENT PLASMODIUM SPECIES -

Fatumo Segun[1], Schramm Gunnar[2], Adabiyi Ezekiel[1], Eils Roland[2], Konig Rainer*[2]

- *[1]Department of Computer and Information Sciences, College of Science and Technology, Covenant University ~ Ota ~ Nigeria - [2]IPMB, Bioquant, University of Heidelberg ~ Heidelberg ~ Germany*

1E) System Biology

Motivation: There are about 156 species of Plasmodium which infect vertebrates. Only four of these species infect human: Plasmodium falciparum, Plasmodium vivax, Plasmodium ovale and Plasmodium malariae. Other species infect vertebrates including birds, reptiles and rodents. The four rodent malaria parasites are Plasmodium berghei, Plasmodium yoelii, Plasmodium chabaudi and Plasmodium vinckei. Since there is a high sequence similarity between human and rodents, we have studied the similarities and differences between the parasites that infect these two organisms, in respect to the differences of the hosts.

Methods: In this paper, a computational biochemical approach was employed to identify chokepoints in the four selected species of Plasmodium. A well established method that detects such enzymes in the metabolic networks which uniquely produce or consume a metabolic compound (Yeh. et al., Genome Res., 2004, 14, 917-924) was applied to select these bottlenecks of the networks. These chokepoints were used for discriminating and grouping Plasmodium species.

Results: There existed several common chokepoints enzymes to all the species. We identified an average of 178 chokepoints enzymes in each of these Plasmodium species which are common to all of them. Interestingly, we detected chokepoints which are only common to particular species. These chokepoints helped to partition the parasites into two groups reflecting their dependencies to the hosts. This analysis shows that the differences between the discovered biochemical networks of the Plasmodium species are not only due to lack of knowledge but mainly because of the parasite-host dependencies. Finally, we propose host specific drug targets which have some evidence when compared to the literature.

Abstract 52

- STATISTICAL ASSESSMENT OF DISCRIMINATIVE FEATURES FOR PROTEIN-CODING AND NON CODING CROSS-SPECIES CONSERVED SEQUENCE ELEMENTS -

Creanza Teresa Maria*[1], Horner David S.[2], D'Addabbo Annarita[1], Maglietta Rosalia[1], Mignone Flavio[3], Ancona Nicola[1], Pesole Graziano[4]

- [1]ISSIA-CNR ~ Bari ~ Italy - [2]Dipartimento di Scienze Biomolecolari e Biotecnologie, Universita' di Milano ~ Milano ~ Italy - [3]Dipartimento di Chimica Strutturale e Stereochimica Inorganica, Università di Milano ~ Milano ~ Italy - [4]Dipartimento di Biochimica e Biologia Molecolare, Universita' di Bari ~ Bari ~ Italy

1A) Genomics

Motivation: The annotation of whole genomes through the identification of coding and regulatory regions is one of the major challenges in the current research in molecular biology. One important topic is identifying the protein coding elements in the set of the mammalian conserved elements. Many features have been proposed for automatically distinguishing coding and non-coding conserved sequence elements (CSEs) making so necessary a systematic statistical assessment of the relevance of single and groups of features in addressing this issue, conditionally to the compared species and to the sequence lengths.

Methods: In our study, we evaluated the relevance of various comparative (based on pairwise cross-genomic comparisons) and intrinsic (based on single-species sequences) features in distinguishing coding from non coding CSEs among human, rat and mouse species by using associative and predictive methods. In order to study the influence of the sequence lengths on the feature performances, the predictive study was performed on different accurately rearranged data sets with coding and non coding alignments in equal number and equally long with an ascending average length. We used Fisher's linear classifiers trained on single as well as groups of features and estimated their prediction accuracies by using multiple cross validation strategy. The statistical significance and power of the estimated prediction accuracy were evaluated by using non parametric permutation tests. Moreover, by using Kolmogorov-Smirnov non parametric tests we investigated if adding intrinsic features to the comparative ones could improve in a statistically significant way the performances of classifiers.

Results: We found that the most discriminant feature was a comparative measure indicating the proportion of synonymous nucleotide substitutions per synonymous sites. Moreover, linear discriminant classifiers trained by using comparative features in general outperformed classifiers based on intrinsic ones. It results that the combination of comparative features is more powerful in the classification of protein coding sequences while the inverse is true for the intrinsic features independently on sequence length. Finally, the prediction accuracy of classifiers trained by using comparative features increased significantly by adding intrinsic features to the set of input variables.

Abstract 55

- IMGT/V-QUEST: AN ALGORITHM FOR IMMUNOGLOBULIN AND T CELL RECEPTOR SEQUENCE ANALYSIS -

Brochet Xavier*[1], Lefranc Marie-Paule[2], Giudicelli Véronique[1]

- [1]IMGT, LIGM, IGH, UPR1142 ~ Montpellier ~ France - [2]IMGT, LIGM, IGH, UPR1142 ~ Montpellier ~ France

Motivation: The molecular synthesis of the immunoglobulin (IG) and T cell receptor (TR) is particularly complex and unique since it generates an extraordinary diversity of the IG and TR repertoires (1012 antibodies and 1012 TR per individual) which results from several mechanisms at the DNA level: the combinatorial diversity of the variable (V), diversity (D) and joining (J) genes, the N-diversity and, for IG, the somatique hypermutations. IMGT/V-QUEST has been developed for the standardized analysis of IG and TR nucleotide sequences.

Methods: IMGT/V-QUEST identifies the closest V, D, J genes and alleles using pairwise alignment and comparison to expertly annotated and standardized data from the IMGT reference directory which is based on IMGT-ONTOLOGY. The algorithm proceeds in 3 steps for the V genes and alleles identification. (1) it identifies a model sequence by aligning the user sequence to a set of the IMGT reference directory comprising ungapped germline V gene sequences (gaps according to the IMGT numbering are stored for the next step), without allowed insertions or deletions. (2) it gaps the user sequence with the positions of the stored IMGT gaps of the model sequence. (3) it identifies the closest germline genes and alleles by the highest similarity score between the gapped user sequence and the complete IMGT reference directory. An optional step detects potential insertions and deletions in the user sequence by Smith and Waterman alignment with the closest germline genes and alleles. If insertions/deletions are detected, the steps for V gene identification are performed again. The J genes and alleles identification proceeds in 2 steps: the beginning of the J is determined by alignment with the IMGT reference directory. Then the closest germline J genes and alleles are identified by similarity evaluation. At last, the algorithm integrates IMGT/JunctionAnalysis for a detailed analysis of the V-J and V-D-J junctions and an accurate D genes and alleles identification.

Results: IMGT/V-QUEST provides a standardized, complete and accurate characterization of the rearranged IG and TR nucleotide sequences. IMGT/V-QUEST is widely used for the study of the IG and TR repertoires and for antibody engineering. It has been recommended by the European Research Initiative on chronic lymphocytic leukemia (ERIC) for the evaluation of the V genes mutational status.

Abstract 56

- HTC FOR ASPIC: A DISTRIBUTED WEB RESOURCE FOR ALTERNATIVE SPLICING PREDICTION AND TRANSCRIPT ISOFORM CHARACTERIZATION -

D'Antonio Mattia[1], Paoletti Daniele[1], Carrabino Danilo[1], D'Onorio De Meo Paolo[1], Sanna Nico[1], Castrignano' Tiziana*[1], Anselmo Anna[2], D'Erchia Anna[3], Licciulli Flavio[4], Mangiulli Marina[3], Mignone Flavio[2], Pavesi Giulio[2], Picardi Ernesto[3], Riva Alberto[5], Rizzi Raffaella[6], Bonizzoni Paola[6], Pesole Graziano[3]

- [1]CASPUR ~ Rome ~ Italy - [2]University of Milan, Dipartimento di Scienze Biomolecolari e Biotecnologie ~ Milan ~ Italy - [3]University of Bari, Dipartimento di Biochimica e Biologia Molecolare ~ Bari ~ Italy - [4]Istituto Tecnologie Biomediche del Consiglio Nazionale delle Ricerche ~ Bari ~ Italy - [5]Department of Molecular Genetics and Microbiology, University of Florida ~ Gainesville ~ United States - [6]DISCo, University of Milan Bicocca ~ Milan ~ Italy

1B) Transcriptomics

Motivation: : Alternative splicing (AS) affects the great majority of intron-containing genes and thus is a major mechanism in the expansion of transcript and protein complexity in eukaryotes. Recent descriptions of the functional implications of AS in tissue-specificity, different biological processes and tumor development has generated an explosion of interest and activity in this field.
Methods: In order to analyse the transcriptome and proteome complexity of multicellular organisms and detect the genes specifically involved in human health and disease, we developed a software platform for high-throughput large-scale alternative splicing analysis and transcript isoform characterization.
This platform, HTC for ASPic, provides independent, flexible and scalable high-throughput large-scale alternative splicing analysis and transcript isoform characterization. It integrates computational intensive algorithms we developed previously [Castrignanò et al. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W440-3.][Castrignanò et al. Bioinformatics. 2008 Apr 3.] with suitable web services and databases.
Results: The software system has been optimized programming multi-threaded powerful Java client for data preprocessing and several distributed application servers for intensive computation. HTC for ASPic divides the input into parallel tasks without dependency and therefore it scales linearly with the number of processors. The system is also fault-tolerant.
The web resource is available free of charge for academic and non-profit institutions.

Abstract 57

- GIBBS FREE ENERGY CHANGES OF BIOCHEMICAL REACTIONS INFERRED FROM REACTION SIMILARITIES -

Rother Kristian*[1], Hofmann Sabrina[2], Bulik Sascha[2], Hoppe Andreas[2], Holzhuetter Herrmann-Georg[2]

- [1]Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology ~ Warsaw ~ Poland - [2]Computational Biophysics Group, Institute of Biochemistry, Charite Universitätsmedizin ~ Berlin ~ Germany

1E) System Biology

Motivation: An indispensable prerequisite for the thermodynamic and kinetic modeling of biochemical reaction networks is to assign a reliable value for the standard Gibbs free energy change (DeltaG0) to each reaction and transporter. However, for genome-wide metabolic networks experimental DeltaG0 values are scarce. Here we propose a novel computational method to infer the unknown DeltaG0 value of a reaction from known DeltaG0 values of chemically similar reactions.

Methods: To quantify the chemical similarity of biochemical reactions we have established a detailed classification procedure that assigns 3304 different chemical attributes to atomic groups occurring in presently characterized biochemical metabolites. Changes in these attributes between the substrate and product molecules are tracked on a per-atom basis and similarities between these reaction-specific attribute changes are assessed by the Tanimoto coefficient (T) assuming values between 0 (complete dissimilarity of reactions compared) and 1 (identity of reactions compared).

Results: Testing our method across a set of 1546 biochemical reactions 216 of which being covered by experimentally determined DeltaG0 values - the root-mean-square distance (RMSD) between predicted and measured DeltaG0 values amounted to 8.0 kJ/mol, if a minimum similarity of T>0.6 to reactions with known DeltaG0 values is assumed. This value is significantly smaller than the RMSD of 10.5 kJ/mol achieved with the commonly used group contribution method. However, for less similar reactions, the group contribution method produces a more accurate predictions and a combination of both approaches is proposed. Clustering all reactions of a given metabolic network according to chemical similarity allows to identify minimal sets of reactions for which DeltaG0 values yet have to be experimentally determined in order to make reliable predictions of DeltaG0 values for the remaining reactions.

Abstract 58

- IS IT POSSIBLE TO PREDICT PROTEIN BEHAVIOR IN HYDROPHOBIC INTERACTION CHROMATOGRAPHY AND AQUEOUS TWO-PHASE SYSTEMS USING ONLY THEIR AMINO ACID COMPOSITION? -

Salgado J. Cristian*[1], Asenjo Juan A.[1], Andrews Barbara A.[1]

-[1]*Centre for Biochemical Engineering and Biotechnology, Department of Chemical Engineering and Biotechnology, University of Chile ~ Santiago ~ Chile*

1D) Molecular structure prediction, modelling and dynamic

Motivation: The prediction of the partition behavior of proteins in hydrophobic interaction chromatography (HIC) and aqueous two-phase systems (ATPS) using mathematical models based on amino acid composition was investigated. Predictive models were based on the average surface hydrophobicity (ASH), which is estimated by means of models that require the 3D structure of proteins and by models that use only the amino acid composition. These models were evaluated in a set of 12 proteins with known experimental retention time in HIC and 11 with known partition coefficient in ATPS. Our results indicate that the prediction based on the amino acid composition is feasible for both separation systems, even though the quality of the prediction depends strongly on the operational conditions. In the case of ATPS the best results were obtained by the model which assumes that all of the amino acids are completely exposed. An increase in the predictive capacity of at least 54% with respect to the models which use the 3D structure of the protein was obtained in this case. However, best prediction in HIC was obtained by the model based on a linear estimation of the amino acidic surface composition. This model required additional tuning, but its performance was 5% better than that obtained by the 3D structure model.

ACKNOWLEDGEMENTS: FONDECYT PostDoctoral Research Project 3070031.

KEYWORDS: Amino acid composition; Mathematical model, hydrophobic interaction chromatography and aqueous two-phase systems.

Abstract 60

- MODELING HETEROCYST PATTERN FORMATION IN CYANOBACTERIA -

Gerdtzen Ziomara P.[1], Salgado J. Cristian*[1], Osses Axel[2], Asenjo Juan A.[1], Rapaport Ivan[2], Andrews Barbara A.[1]

- [1]Millennium Institute for Cell Dynamics and Biotechnology, Department of Chemical Engineering and Biotechnology, University of Chile ~ Santiago ~ Chile - [2]Millennium Institute for Cell Dynamics and Biotechnology, Department of Mathematical Engineering, Center for Mathematical Modeling (UMI 2807-CNRS), University of Chile ~ Santiago ~ Chile

1E) System Biology

Motivation: In this paper we study the process by which vegetative cells of cyanobacteria differentiate into heterocysts in the absence of nitrogen. We propose a simple network which captures the complexity of the differentiation process and the role of all variables involved in this cellular process. Specific characteristics and details of the system's behavior such as transcript profiles for ntcA, hetR and patS between consecutive heterocysts are studied. The proposed model is able to capture one of the most distinctive features of this system: a characteristic distance between two heterocysts, with a small standard deviation according to experimental variability. The system's response to knock out and over expression of patS and hetR was simulated in order to validate the proposed model against experimental observations. In all cases, simulations show good agreement with reported experimental results. The model also shows that refractability of heterocysts to the action of PatS is not required in order to achieve the characteristic differentiation pattern observed in cyanobacteria.

ACKNOWLEDGEMENTS: Millennium Scientific Initiative ICM P05-001F.

KEYWORDS: cyanobacteria, heterocyst, mathematical modeling, cell differentiation and gene network.

Abstract 61

- IDER ASSOCIATED PHYSIOLOGICAL NETWORK IN MYCOBACTERIA -

Ranjan Sarita[1], Ranjan Akash*[2]

- [1]1LEPRA-Blue Peter Research Centre, ~ Hyderabad ~ India - [2]Sun Centre of Excellence in Medical Bioinformatics, EMBnet India Node, CDFD, ~ Hyderabad ~ India

1E) System Biology

Motivation: Transcription regulators play an important role in coordinating gene expression of physiologically related genes in an organism. Each transcription factor recognizes specific DNA sequence close to promoter regions and modulated gene expression by binding to these sequences. Using bioinformatics, we can learn the DNA sequence pattern associated with these transcription factors and predict genome wide targets of these regulators. Taking operonic context of the predicted targets it is possible construct a model of physiological interaction network that reveals the genome encoded biology associated with the transcriptional regulator. We have taken this approach to model IdeR (iron dependent regulator) associated physiological network in mycobacteria.

Methods: Using a profile based approach the DNA sequence pattern associated with IdeR regulator was recognized. The IdeR associated DNA binding pattern was subsequently used to predict genome wide targets of IdeR in sequenced genomes of mycobacteria. Taking operonic context of the predicted targets we construct a model of physiological interaction network of the IdeR. These interactions were modeled using Cytoscape.

Results: A model IdeR (iron dependent regulator) associated physiological network in mycobacteria constructed. Using comparative genomics approach we have also explored the relative conservation IdeR associated physiological network in different sequenced species of mycobacteria.

Abstract 62

- CORYNEDB: A DATABASE OF IN SILICO IDENTIFIED OPERONS AND TRANSCRIPTIONAL UNITS OF CORYNEBACTERIA -

Ranjan Sarita*[1], Savala Narendra Kumar[2], Ranjan Akash[2]

- [1]LEPRA-Blue Peter Research Centre, ~ Hyderabad ~ India - [2]Sun Centre of Excellence in Medical Bioinformatics, EMBnet India Node, CDFD ~ Hyderabad ~ India

1B) Transcriptomics

Motivation: Corynebacteria are a diverse group of microorganism belonging to the phylum Actinobacteria. Species belonging to Corynebacterium genus are gram positive, non motile acid fast staining bacilli. Corynebacteria are found in a wide range of different ecological niches such as vegetables, soil, cheese smear, skin, and sewage. Some of the species, such as Corynebacterium diphtheriae and Corynebacterium jeikeium, are important pathogens while others, such as Corynebacterium glutamicum, are of immense industrial importance as they are extensively used to produce amino acid using fermentation process. Corynebacteria are also closely related to other important group of pathogens called mycobacteria.

Methods: We have applied our in house developed operon prediction approach (Ranjan et al BMC Bioinformatics 2006; 7(Suppl 5): S9) to identify operons and transcriptional unit in sequenced genomes of corynebacteria. This approach involved orientation analyses, Intergenic distance analyses, transcriptional terminators analyses and conserved gene cluster analyses.

Results: We have predicted transcriptional units and operons in six sequenced genomes of corynebacteria. The predicted operons and transcriptional units are organized as relational database called CoryneDB. CoryneDB has information about corynebacterial genes and in silico predicted transcriptional units and operons. This database would assist the scientific community, to hypothesize functional linkages between operonic genes of corynebacteria, their experimental characterization and validation.

Abstract 63

- DOOPSEARCH: A WEB-BASED TOOL FOR FINDING AND ANALYZING COMMON CONSERVED MOTIFS IN THE PROMOTER REGIONS OF DIFFERENT CHORDATE AND PLANT GENES -

Sebestyén Endre[1], Nagy Tibor[2], Barta Endre*[2]

-[1]*Agricultural Research Institute of the Hungarian Academy of Sciences ~ Martonvásár ~ Hungary* -
[2]*Bioinformatics Group, Agricultural Biotechnology Center ~ Gödöllo ~ Hungary*

1A) Genomics

Motivation: The comparative genomic analysis of a large number of orthologous promoter regions from chordate and plant genes revealed thousands of conserved motifs. Most of these motifs are different from any known transcription factor binding sites (TFBS). To identify new potential TFBSs with in silico analysis methods, we need a tool to be able to search among the conserved motifs. The result of a given search is expected to provide a list of genes, which are associated with a certain conserved motif and might be regulated by a transcription factor recognising the motif and binding to it. To test and confirm the association of a conserved motif with certain types of genes, we can perform a Gene Ontology (GO) analysis on the gene list.
Methods: Based on our DoOP database, we used different taxonomic groups to extract conserved motifs either from the human genome annotation based chordate or the Arabidopsis thaliana based plant database. We have developed a C program called MOFEXT, for performing gapless alignments and fast searches in the different motif collections. The FUZZNUC program from the EMBOSS package has also been implemented to search in the promoter sequences of the DoOP database. We slightly modified the GeneMerge program to use it for the GO analysis of the results. To handle the web-based queries efficiently, all data are stored in a MySQL database. We have developed several PERL modules (available from CPAN) to carry out the querying of the MySQL database, the MOFEXT searches, the GO analysis and the graphical presentation of the results.
Results: We have developed a new web page called DoOPSearch (http://doopsearch.abc.hu) for the analysis of the conserved motifs in the promoter regions of chordate or plant genes. We used the orthologous promoters of the DoOP database to extract conserved motifs from different taxonomic groups. The advantage of this approach is that different sets of conserved motifs may be found depending on how broad the taxonomic coverage of the underlying orthologous promoter sequences is (consider e.g. primates vs. mammals). The DoOPSearch web page allows the user to search these motif collections or the promoter regions of DoOP with user supplied query sequences or any of the conserved motifs from the DoOP database. The gene lists obtained can be further analyzed using the modified GeneMerge program.

Abstract 64

- BRAGOMAP - A NEW PERL SCRIPT FOR HIGH THROUGHOUTPUT BLAST RESULTS ANALYSIS INCLUDING GO AND MAPMAN AUTOMATIC ANNOTATIONS -

Woycicki Rafal*[1], Gutman Wojciech[2], Przybecki Zbigniew[1]

- [1]Department of Plant Genetics, Breeding and Biotechnology, Faculty of Horticulture and Landscape Architecture, Warsaw University of Life Sciences ~ Warsaw ~ Poland - [2]Department of Biochemistry, Faculty of Agronomy and Biology, Warsaw University of Life Sciences ~ Warsaw ~ Poland

1A) Genomics

Motivation: Analyzing of sequences similarities is the first and most important method used to find out the function of unknown nucleotides.
Searching of homologs should be done carefully not to loose any important ones. Having thousands of results from various long-read sequencing projects (genomic polymorphons or BAC ends), the by-hand ability to retrieve interesting (to our goal) similarities in hundreds of Blast results decreases rapidly.
Decreasing the number of retrieved sequences by giving more stringency in e-value threshold or displaying less results could lead to false deductions.
Functional genomics, proteomics and metabolomics could give us answers to the role of nucleotide sequences. It makes the need to annotate as much of the homologies as we can, to proper molecular function, biological process and cellular component (as its proposed by widely accepted Gene Ontology Consortium annotations or MapMan mappings by Max-Planc-Institute).
Methods: To facilitate fast retrieval of interesting Blast homologies and making right deductions about the biological role of sequences, in big sequencing projects, the new Perl script BRAGOMAP was written. The program make use of some of BioPerl modules as well as the power of regex text-mining in the Perl itself.
Results: The script gives us the possibility to find interesting sequence similarities by using keywords and giving points for each one found. It collects all important information from the GenBank data and puts it in different columns of tab-delimited file for further use.
If we were interested (for example) in flower differentiation genes we could use the keywords (flower, ovule, anther, etc.) and/or filter all the homologies isolated from flower tissues in a special development stage. We can also filter results by choosing similarities to interesting genes or protein products. This script retrieve also all standard information from the Blast and GenBank files as Description, ACC no., E-value, Similarity positions, Query Length, Percent of Similarity etc.
Automatic GO and MapMan annotations are done by looking for genes, protein products and /or DB references in the proper mappings files.
Here we present the usefulness of the script in analyzing sequence similarities and annotations mapping of 3855 BAC ends obtained from the HindIII BAC genomic library of cucumber (Cucumis sativus L., line B10).

Abstract 65

- IMPROVING GENE RANKING METHODS USING FUZZY CLUSTERING AND GENE ONTOLOGY -

Mohammadi Azadeh*[1], Saraee Mohammad Hossein[1]

- [1]*Advanced Database Syetems, Data Mining and Bioinformatics Research Laboratory, Department of Electrical and Computer Engineering,Isfahan University of Technology ~ Isfahan ~ Iran*

1A) Genomics

Motivation: Microarray technology allows simultaneously monitoring the expression levels of thousands of genes. An important analysis task for this data is identification of the genes which are significant or mostly associates with a disease. Identification and selection of such genes from thousands of genes in microarray experiments, is called gene selection. A variety of approaches have been proposed for gene selection. A group of methods are ranking methods which measure the discriminatory power of each gene according to a test-statistic and select the genes with highest score as discriminatory genes. Although these kinds of methods have low computational complexity they have two drawbacks. First, they don't consider the correlation between genes and consequently the selected genes have redundancy. Second these methods don't utilize the biological knowledge.
Methods: In this paper we have hybridized fuzzy clustering and gene ontology with ranking methods such as t-test fisher, information gain and TNOM, to solve the mentioned problems. In the proposed method genes are clustered based on their gene expression profiles and their biological knowledge obtained from gene ontology, after that a test statistic is applied on genes to rank them. Since the genes in a cluster represent similar genes, we allow only a limited number of genes from a cluster to be selected, in this way the correlation amongst selected genes decrease considerably.
Results: We have applied the proposed method on colon dataset and compared the result of our method to some ranking methods such as t-test, fisher, information gain and TNOM. The results show that Coupling clustering and gene ranking methods can identify gene subset that has lower redundancy and improves classification accuracy, compared with simple gene ranking methods.

Abstract 66

- GENEFINDER: "IN SILICO" POSITIONAL CLONING OF TRAIT GENES -

Martínez Barrio Álvaro*[1], Lagercrantz Erik[2], Burvall Sofia[3], Bongcam-Rudloff Erik[2]

- [1]The Linnaeus Centre for Bioinformatics, Uppsala University, Biomedical centre, P.O. Box 598, SE-75124 ~ Uppsala ~ Sweden - [2]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Biomedical centre, P.O. Box 597, SE-751 24 ~ Uppsala ~ Sweden - [3]Uppsala University ~ Uppsala ~ Sweden

1K) Biological data integration

Motivation: : Positional cloning of trait genes is very laborious and the amount of information on gene function in different organisms is increasing so rapidly that it is hard for a research group to collect all relevant information from a number of data sources without performing a big number of tedious, non-automatized and time consuming searches.
Methods: A specific application known as GeneFinder has been designed and implemented in order to collect information of a trait locus controlling a specific phenotype mapped to a certain chromosomal region. The information retrieved consists of information on gene function, disease conditions, tissue expression and even predicted gene homologies on several other species. All this amount of information is then further combined with a specially devised ranking algorithm. Finally, the results of the search are presented to the user in a friendly ranked list doing easy its interpretation. To achieve this, a web interface to the GeneFinder webservice was also developed. Our distributed application is publicly available and free to use.
Results: We show how our implementation works, both with its web interface and programmatically with the webservice API, in a very short number of steps when searching an example candidate region. An online server for the web interface is available at http://genefinder.ebioinformatics.org/.

Abstract 70

- STORING, CLASSIFICATION AND BETA-LACTAMASE SEQUENCE MOLECULAR PATRONS DETECTION -

Rodríguez Luis[1], Emiliano Barreto*[1], Ramón Mantilla[1], Reguero Maria Teresa[1]

- [1]*Bioinformatics Center, Biotechnology Institute, National University of Colombia ~ Bogotá ~ Colombia*

1J) Biobanks (databases and knowledgebases)

Motivation: Recently, beta-lactamases have gained attention due to their role in bacteria resistance to beta-lactamic antibiotics. Beta-lactamase identification and classification are big challenges for the molecular epidemiology research, due to the few differences in their DNA sequences. Consequently, is difficult to classify new sequences in the beta-lactamases classes and families already described. The currently used classification is based in the Lahey Clinic and the Pasteur Institute bioinformatics network. Despite this, a bioinformatics system focused on storing outbreaks, sequences, phylogenetic and bibliography data isn't available. The current work was intended to develop it.
Methods: The bioinformatic system BLA.id was implemented in Linux Suse 10 box, using MySQL and including Perl and PHP written scripts and some C ++ programs. The user web interface was built using PHP, MySQL and R. The system automatically retrieved information from international databases such as UniProt, EMBL and PubMed. BLA.id data was manually annotated and reviewed. The classification used was based on lists from Lahey Clinic (http://lahey.org/studies/), Pasteur Institute (http://www.pasteur.fr/recherche/genopole/PF8/betalact_en.html) and single international journal reports.
Results: We were able to develop a system that has beta-lactamases sequences actually reported properly classified. It allows the direct report of new sequences into the data base and their accurate classification. Our bioinformatics system has leaded us to the detection of some classification inconsistencies in beta-lactameses' sequences already reported. It also has permitted the phylogenetic classification and the identification of conserved patterns that are useful for molecular epidemiologic approaches. Moreover, some problems such as the existence of sequence synonyms can be easily solved on this structured solution. BLA.id is accessible via web on http://co.embnet.org/BLA.id/.

Abstract 71

- DETECTING HYDROPHOBIC CLUSTERS IN 3D STRUCTURES AT ATOMIC LEVEL -

Alexeevski Andrei[1], Sergei Spirin*[1], Karyagina Anna[2]

- [1]Belozersky Institute, Moscow State University ~ Moscow ~ Russia - [2]Gamaleya Institute of Epidemiology and Microbiology ~ Moscow ~ Russia

1D) Molecular structure prediction, modelling and dynamic

Motivation: The hydrophobic effect is one of the main natural forces stabilizing structures of macromolecules and their complexes. The basis of the hydrophobic effect is that nonpolar atoms tend to dispose together in a water environment.
Methods: Our approach is based on the supposition that every set of closely located nonpolar atoms leads to a force trying to keep their close location. This suggestion dictates an "atomic" level of investigating hydrophobic effect, because, for example, such polar aminoacid residues as lysine, arginine, or glutamate contain nonpolar atoms, which can participate in hydrophobic interactions with other residues or ligands.
Results: We introduce an algorithm for detecting clusters of nonpolar atoms in structures of macromolecules, and a web service implementing this algorithm. The web service is available at http://monkey.belozersky.msu.ru/npidb/cgi-bin/hftri.pl. The program can be used in two modes. First, it can detect hydrophobic clusters in an entire structure, for example, the main hydrophobic core of a protein globule and minor hydrophobic areas at the surface of a molecule. Second, as its input it can take two parts of a macromolecular complex (e.g., two interacting protein molecules) and detect hydrophobic clusters at the interface of those parts. The latter can help to investigate hydrophobic interaction between two molecules.
Also we suggest a procedure to detect conserved hydrophobic cores in a family of related proteins and conserved hydrophobic interactions in a family of related complexes. Programmatic implementation of the procedure is a subject of our current efforts.
We used our program to describe conserved hydrophobic clusters in a number of structural families (e.g., homeodomains and V-set domains), for analysis of inter-subunit interaction in capsids of icosahedral viruses, for analysis of DNA-protein interaction in a number of complexes of DNA-binding proteins with DNA (a part of results was published in: Karyagina et al. J. Bioinf. Comp. Biol. 2006; 4(2); 357-372).

Abstract 73

- REGEXPBLASTING, A REGULAR EXPRESSION RULES BASED ALGORITHM TO CLASSIFY A NEW SEQUENCE -

Rubino Francesco[1], Attimonelli Marcella*[1]

- [1]*Department of Biochemistry and Molecular Biology, Bari, Italy*

1A) Genomics

Motivation: One of the most frequent usages of bioinformatics tools concerns the functional characterization of a newly produced nucleotide sequence (a query sequence) by applying Blast or FASTA against a set of sequences (the subject sequences). However, in some specific context it could be useful to compare the query sequence against a cluster such as a multialignment. The purpose of the RegExpBlast tool is i) to associate to each multialignment a pattern, defined through the application of regular expression rules; ii) to automatically characterize the submitted nucleotide sequence on the basis of the function of the sequences described by the pattern better matching the query sequence.

Methods: Regular expressions are tools used to represent every possible character variation in a sequence alignment, much more powerful than the classic consensus sequences that leaves out a great quantity of information about that site, like possible SNPs, insertions and deletions.

The RegExpblasting tool is organized in two sections: A and B. RegExpBlasting section A produces the RegExp pattern describing each of the considered multialignments by extracting the regular expression which represents all the variations of the group of sequences available through their multialignment. RegExpblasting section B allows the classification of a new sequence by comparing it to each of the patterns defined in section A and reports as output the matching multialignments where the new sequence can be easily added.

The algorithm section A scans all the multialignment sites and produces a regular expression according to the below described criteria:

1) if a site contains only one type of nucleotide, only this will be included in the regular expression;
2) if the types of nucleotides in a site are two or more, all this nucleotides will be enclosed in a character class to represent every nucleotide variation;
3) if a site contains one or more gaps the meta character “?” will be added to represent this case.

The resulting pattern represents every sequence composing the multialignment.

Section B compares the query sequence with the patterns defined in section A and produces as output the multialignments associated to the best matching patterns.

Results: An application of these algorithms is used in the "characterize your sequence" tool available in the PPNEMA resource [1] http://www.ppnema.uniba.it. PPNEMA is a resource of Ribosomal Cistron sequences from various species grouped according to nematode genera. PPNEMA allows the retrieval of plant nematode multialigned sequences or the classification of a new nematode rDNA sequence by applying RegExpblasting. The same algorithm supports automatic updating of the PPNEMA database also.

References: F.Rubino, A.Voukelatou, F.De Luca, C.De Giorgi and M.Attimonelli “PPNEMA: a database of the RNA cystron from Plant Parasitic nematodes.” Int.J.Plant Genomics, Sp.Issue on Bioinforamtics, 2008, in press.

Abstract 74

- POSSIBLE ROLE FOR PROXIMITY OF GENES IN THEIR EXPRESSION IN RICE AND ARABIDOPSIS -

Shahmuradov Ilham A.*[1], Akbarova Yagut Yu.[3], Aliyev Jalal A.[2], Qamar Raheel[1], Chohan Shahid Nadeem[1], Solovyev Victor V.[4]

- [1]Dept. of Biosciences, COMSATS Institute of Information Technology, Islamabad, Pakistan - [2]Bioinformatics Laboratory, Institute of Botany, Baku, Azerbaijan - [3]College of Medicine and Health Sciences, Sultan Qaboos, University, Muscat, Sultanate of Oman - [4]Dept. of Computer Science, Royal Holloway, University of London, Egham, UK

1A) Genomics

Motivation: In contrast to prokaryotes, the proximity of genes in eukaryotic genomes has not previously been known to play any significant role in their expression profiles. However, two recently reported phenomena in eukaryotes including human, mouse, yeast and a few plant species indicate such a role. These phenomena are: (1) transcription of the Head-to-Head (H2H) adjacent genes from the shared promoter; and (2) chimeric mRNAs and proteins produced via alternative transcription termination, splicing and translation of the Tail-to-Head (T2H) neighboring genes. To further verify this evidence at the genomic scale, we searched through the genomes of rice and Arabidopsis for the presence of H2H and T2H gene pairs at a distance of less than 800 bp and 1000 bp, respectively.
Methods: Gene ontology data for rice and Arabidopsis were obtained from the genome annotations and TAIR WEB-site, (ftp://ftp.Arabidopsis.org/home/tair/Genes/Gene_OntologyATH_GO.20031202.txt), respectively. Analysis of the genome annotations of rice and Arabidopsis was performed using computer programs ARGAN and OSGANn specially developed by us for this task. The pair-wise comparison of amino acid sequences has been carried out by BLAST program. Search for promoters and statistically significant open reading frames were performed by TSSP and BESTORF programs, respectively (www.softberry.com).
Results: In the rice genome, 580 H2H and 1,386 T2H gene pairs were found, while 1,898 H2H and 6,618 T2H gene pairs were found in the Arabidopsis genome. The short spacers between H2H genes in both the genomes may serve as potential bidirectional promoters, though they did not reveal any significant evolutionary conservation, However we found striking conservation of intergenic region between the same gene pair in Arabidopsis and Brassica napus with the experimentally confirmed bidirectional promoter. Further studies suggest that “non-stopping” transcription and alternative splicing of some of these T2H pairs may result in chimer transcripts. We obtained cDNA support for 106 and 105 rice and Arabidopsis T2H pairs, respectively, to be transcribed into chimeric or read-through mRNA(s). Analysis of the protein coding potential of such putative transcription-induced chimer genes revealed putative chimer proteins having significant similarity with known plant proteins.

Abstract 75

- A BIOINFORMATICS PLATFORM TO STORE AND ANALYZE ALTERNATIVE SPLICING EVENTS DETECTED BY EXON ARRAYS -

Licciulli Flavio[1], Picardi Ernesto[2], Pesole Graziano*[1], Calogero Raffaele[3], Delle Foglie Gianfranco[1], Grillo Giorgio[1], Liuni Sabino[1]

- [1]Istituto Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Bari, Italy - [2]Dipartimento di Biochimica e Biologia Molecolare "E.Quagliariello", Università di Bari, Bari, Italy - [3]Bioinformatics and Genomics unit, Dipartimento di Scienze Cliniche e Biologiche, Università di Torino, Orbassano, Italy

1K) Biological data integration

Motivation: Current studies suggest that more than 90% of human multi-exon genes are subjected to alternative splicing, a key molecular mechanism in which multiple transcripts may be generated from a single gene. Such transcripts may encode proteins with different biological functions or may be subjected to different post-transcriptional regulations through variants of 5' and 3'UTRs. As a consequence, the alternative splicing greatly increases the complexity of the human transcriptome and proteome.

Last generation microarray platforms such as exon arrays now offer a more detailed view of the gene expression profile providing information on the alternative splicing pattern.

Exon arrays, in fact, have a capacity of more than six million data points and have been designed to interrogate approximately one million exons including all annotated exons (from RefSeqs) in addition to computationally and empirically identified exons (including those supported by ESTs and gene predictions as well).

Exon arrays significantly differ from traditional microarray platforms in the number and placement of the oligonucleotide probes, and allow researchers to perform two different but related analyses at both gene and exon level. In particular, the accession to genome-wide exon data can provide answers to challenging issues concerning the identification of tissue-specific exons, biomarkers and isoforms involved in different pathological mechanisms.

In this context and in order to investigate tissue-specific, developmental stage-specific as well as disease-related alternative splicing events, we posed our attention to the GeneChip Human Exon 1.0 ST Array system by Affymetrix. One significant feature of this innovative Affymetrix platform is that individual probes are designed along the full-length of human genes. Moreover, an exon-level probe set is made of 4 probes and more than one exon-level probe set may fall in the same exon. All generated probe sets related to a specific gene are then grouped in transcript clusters. Annotation details such as genome coordinates for each feature and the relationships among exons, probe sets and transcript clusters are provided and frequently updated by Affymetrix.

Current protocols to manage and analyze Affymetrix Exon Array data are not completely defined. Affymetrix proposed a workflow based on pre-filtering of the

expression data, transformation of exon-level intensity data in gene-level normalized values called splice index (SI) and statistical validation based on an ANOVA based method based on measuring differences between exon level signal and aggregate gene level signal called MiDAS (Microarray Detection of Alternative Splicing).. Unfortunately, exon array workflows do not include facilities to provide a simple and basic visualization of the putative alternative splicing events (ASEs) to facilitate the interpretation of their biological significance.

With the aim to fill this gap and, thus, provide a user-friendly representation of alternative splicing events in different experimental conditions we built a robust bioinformatics platform based on a data-warehouse approach.

Methods: According to data-warehouse standards suggested by Kimball and Inmon (Inmon 1995. Prism Solutions), our bioinformatics platform can manage a huge quantity of data and, in the meantime, facilitate their biological interpretation by using Business Intelligent (BI) techniques. In its first version, the platform integrates experimental data produced by Affymetrix Exon 1.0 ST Arrays and related annotations from several public and specialised databases such as AspicDB (Castrignanò et al. Nucleic Acids Res. 2006; 34,W440-443), Gene, GO and KEGG. At the heart of our system is the mapping of Exon Array probe sets along all transcripts predicted and annotated in ASPicDB in order to provide powerful and stimulating insights at genomic and transcriptomic level. Gene, transcript and exon annotation data are, therefore, integrated with experimental result data and other well-established databases to setup the primary database of the data-warehouse architecture (Staging Area). Specific DataMarts aggregating data stored in the main Staging Area are then built to navigate and analyze this huge quantity of data. Finally, BI tools such as multidimensional queries, graphics and reports are used to simplify the biological interpretation of Exon Array results.

Results: At this stage of development the Staging Area contains data from publicly available Exon Array experiments. In particular, it stores normalized intensities (at gene and exon level) from experiments carried out on eleven different human tissues and some normal and tumour colon samples. In addition user-provided data may be uploaded in the platform in order to analyze and visualize experimental results in an alternative splicing “perspective”.

An heat map visualization of the hybridization signal at gene, transcript or exon level is provided in all the conditions sampled by experiments stored or uploaded by the user. In this way the user can retrieve and visualize the differential expression of alternative isoforms collected in ASPicDB, in different conditions, using also GeneOntology terms, biochemical pathways (KEGG) and other functional annotation as search criteria.

Abstract 76

- EASYCLUSTER: A FAST AND EFFICIENT GENE-ORIENTED CLUSTERING TOOL FOR LARGE-SCALE TRANSCRIPTOME DATA -

Picardi Ernesto*[1], Pesole Graziano[2]

- [1]*Dipartimento di Biochimica e Biologia Molecolare "E.Quagliariello", Università di Bari, Bari, Italy* - [2]*Istituto Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Bari, Italy*

1H) Text and data mining

Motivation: Expressed sequence tags and full-length cDNAs represent an invaluable source of evidence for inferring reliable gene structures and discovering potential alternative splicing events. In newly sequenced genomes, these tasks could not be easily feasible for several limitations such as the lack of appropriate training sets. However, when expression data (mainly ESTs and FL-cDNAs) are available, they can be consistently used to build EST clusters related to specific genomic transcribed loci. Common strategies employed in the last years, including procedures implemented in UNIGENE, TIGR Gene Index and STACK, are based on sequence similarity mainly detected by means of BLAST, Blat or d2_cluster programs. Moreover, genomic sequences are not always taken into account during the cluster reconstruction. As a consequence, EST sequences can be erroneously grouped leading to inconsistent results when merged in consensus transcripts. In order to improve existing procedures for cluster building and facilitate all downstream annotation analyses, we developed a simple but efficient system to generate gene-oriented clusters (Unigene-like) of ESTs every time a genomic sequence and a pool of related expressed sequences are provided. Our procedure, named here EasyCluster, takes into account the spliced nature of ESTs after an ad hoc genomic mapping. EasyCluster has proved to be highly reliable after a benchmark on a manually curated set of human EST clusters. Therefore, it can be suitably applied for clustering EST in the case a genome sequence is available but not a reliable gene annotation.

Methods: Differently from other existing clustering methods based on similarity searches, EasyCluster takes advantage from the well-known EST-to-genome mapping program GMAP (Wu and Watanabe. Bioinformatics 2005; 21,1859-1875). The main benefit of using GMAP is that it can perform a very quick mapping of whatever expressed sequence onto a genomic sequence attended by an alignment optimization. In particular, GMAP can detect splicing sites according to a so defined "sandwich" dynamic programming that is organism independent.

Given a genomic sequence and a pool of ESTs/FL-cDNAs, EasyCluster first builds a GMAP database of the genomic sequence to speed-up the mapping and a local EST database storing all provided expressed sequences. Subsequently, it runs GMAP program and parses results in order to create an initial collection of pseudo-clusters. Each pseudo-cluster is obtained by grouping ESTs according to the overlap of their

genomic coordinates on the same strand. In the next step, EasyCluster refines the EST grouping by running again GMAP on every pseudo-cluster and by including in a cluster only expressed sequences sharing at least one splice site. Finally, for each generated cluster EasyCluster produces a graphical representation in pure HTML code for a simple inspection of results by eyes. EasyCluster is written in python programming language and works on all unix-based platforms where GMAP can be installed.

Results: The EasyCluster program provides EST/FL-cDNA clusters ready to be used in gene prediction pipelines and to detect alternative splicing events. In order to investigate the reliability of EasyCluster we tried to group 256 spliced ESTs related to eleven human homeobox genes (family HOXA) located in the chromosome 7. The same pool of ESTs was used as input in wcd, a new and computationally efficient program to build EST clusters based on sequence similarity (Hazelhurst et al. Bioinformatics. 2008; 13,1542-1546).

EasyCluster was able to reconstruct eleven clusters corresponding to each homeobox gene. In contrast, wcd predicted only nine groups where two of them were related to more than a gene. In particular, ESTs supporting HOXA3 and HOXA4 genes and HOXA9 and HOXA10 genes were clustered together. Our simple results, therefore, demonstrate the reliability of EasyCluster and, in general, of genome-based EST clustering programs over widespread systems based on sequence similarity.

Given the simplicity, flexibility and portability of our system, we are planning to introduce EasyCluster in a more complex pipeline to facilitate the genome-wide detection of the alternative splicing in newly sequenced genomes.

Abstract 77

- GNPIS, THE PLANT INFORMATION SYSTEM OF INRA URGI BIOINFORMATICS PLATFORM -

Steinbach Delphine*[1], Alaux Michael[1], Kimmel Erik[2], Durand Sophie[2], Pommier Cyril[2], Luyten Isabelle[2], Mohellibi Nacer[2], Verdelet Daphne[2], Quesneville Hadi[2]

- [1]*Bioinformatics ~ Versailles ~ France* - [2]*INRA URGI bioinformatics ~ Versailles ~ France*

1K) Biological data integration

Motivation: : URGI (Unité de Recherche Génomique-Info) is an INRA bioinformatics unit dedicated to plants and pest genomics. Created in 2002, one of its mission is to develop and host a genomic and genetic information system called GnpIS, for INRA plants of agronomical interest and their bioagressors. It hosts a bioinformatics platform which belongs to the ReNaBi network and is labelled RIO/IBISA 2007. The URGI maintains an efficient computing environment and offers services covering database conception, software engineering, and bioinformatics. Since 2007, the unit hosts a research team which work is focused on repeats detection and analysis (REPER pipeline). The work presented here will show the description of the actual information system, its development and modular evolution during time, data results since 2000 and details concerning the state of the databases interoperability project.

Methods: GnpIS: information system:

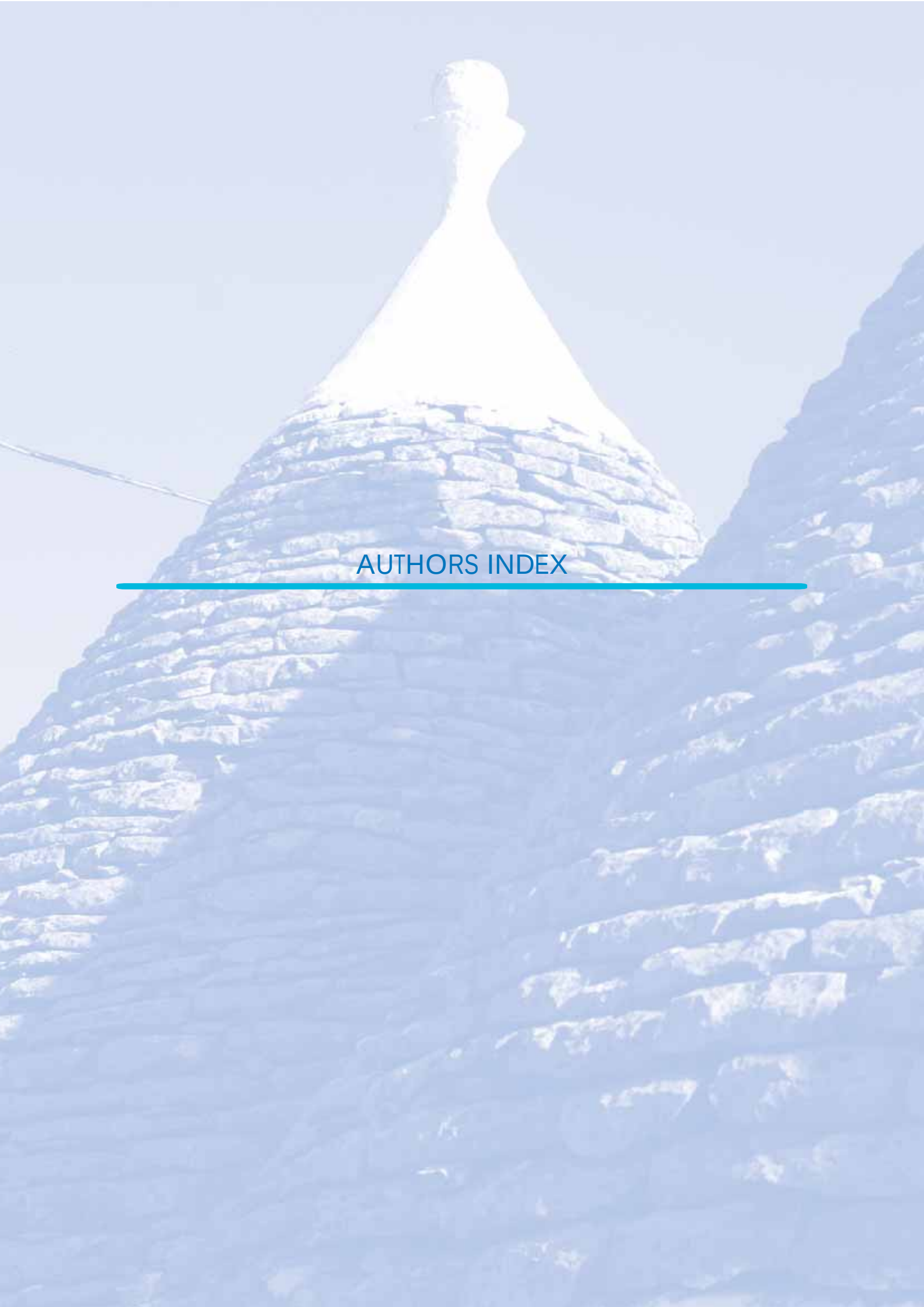
The URGI information system called GnpIS, is a web based system composed of several applications (in Java and Perl) built above a relational database that includes integrated schemas for sequence data, annotation data, mapping data, expression data, proteomic data and SNP data. Since 2005, a new module concerning genetic resource data (SiReGal) was added to the system. Data are submitted by the laboratories through an automatic Web submission tool which allows the checking and the data bulk loading. Web interfaces allow the biologists to query and visualize the data and navigate through them. The ongoing developments are the creation of an interoperability between the genomic and genetic databases modules to allow integrated queries involving all kind of data together and also to be able to skip from one thematic to the other transparently (GnpGenome, GnpSnp, GnpMap, GnpSeq). 2 technologies are in test (in 2008), Biomart (EBI) and Hibernate/Lucene technology, JAVA J2EE technology.

Aster is the key and central module of the information system. It allows the interoperability between the modules.

Results: For all the resources, the databases are available for query either on a public access (http://urgi.versailles.inra.fr), either with an account for partners before publication.

References: See: http://urgi.versailles.inra.fr/about/publications

Grants: Genoplante, ANR Genoplante and INRA



AUTHORS INDEX

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 (CIByB). University of Chile, Santiago

China
Centre of Bioinformatics, Peking University, Beijing

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

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

Cuba
Centro de Ingeniería Genética y Biotecnolgía, La Habana

Finland
CSC, Espoo

France
ReNaBi, French bioinformatics platforms network

Greece
Biomedical Research Foundation of the Academy of Athens, Athens

Hungary
Agricultural Biotechnology Center, Godollo

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

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

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

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

The Netherlands
Dept. of Genome Informatics, Wageningen UR

Norway
The Norwegian EMBnet Node, The Biotechnology Centre of Oslo

Pakistan
COMSATS Institute of Information Technology, Chak Shahzaad, Islamabad

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

Portugal
Instituto Gulbenkian de Ciencia, Centro Portugues de Bioinformatica, Oeiras

Russia
Biocomputing Group, Belozersky Institute, Moscow

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

South Africa
SANBI, University of the Western Cape, Bellville

Spain
EMBnet/CNB, Centro Nacional de Biotecnología, Madrid

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

Sweden
Uppsala Biomedical Centre, Computing Department, Uppsala

Switzerland
Swiss Institute of Bioinformatics, Lausanne

Specialist Nodes

EBI
EBI Embl Outstation, Hinxton, Cambridge, UK

ETI
Amsterdam, The Netherlands

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

IHCP
Institute of Health and Consumer Protection, Ispra. Italy

ILRI/BECA
International Livestock Research Institute, Nairobi, Kenya

LION Bioscience
LION Bioscience AG, Heidelberg, Germany

MIPS
Muenchen, Germany

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

for more information visit our Web site

www.embnet.org

EMBnet.news
ISSN 1023-4144

Dear reader,

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

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

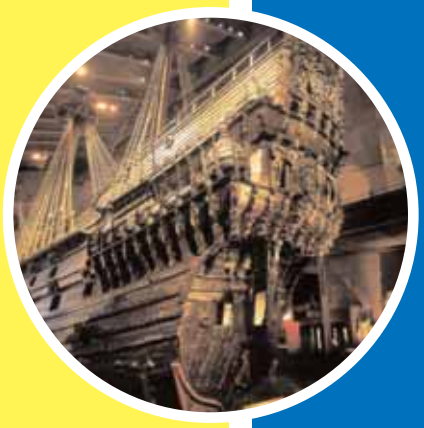
Publisher:
EMBnet Executive Board
c/o Erik Bongcam-Rudloff
Uppsala Biomedical Centre
The Linnaeus Centre for Bioinformatics, SLU/UU
Box 570 S-751 23 Uppsala, Sweden
Email: erik.bongcam@bmc.uu.se
Tel: +46-18-4716696

Submission deadline for the next issue:
November 20, 2008

17th Annual International Conference
on Intelligent Systems for
Molecular Biology (ISMB)
& 8th European Conference on
Computational Biology (ECCB)

ISMB ECCB STOCKHOLM 2009

JUNE 27 – JULY 2



Mark Your Calendar!

Gunnar von Heijne, Honorary Conference Chair
Stockholm University, Stockholm, Sweden

Eugene Myers, Conference Co-chair
HHMI Janelia Farm Research Campus, Ashburn, VA, USA

Marie-France Sagot, Conference Co-chair
INRIA Grenoble — Rhône-Alpes Research Centre
Montbonnot-Saint Martin, France

ISCB International Society for Computational Biology

An Official Conference of
the International Society
of Computational Biology

http://www.iscb.org/ismbeccb2009/

Platinum Sponsors:

Gold Sponsors:

Silicon Graphics S.p.A.

http://imgt.cines.fr

Silver Sponsors:

Consiglio Regionale della Puglia - Italy

Student travel fellowships:

Supporters Sponsors:

Italy

Southern Partnership for Advanced Computational Infrastructures - Italy

Bari-Italy

Italy

Institute for Biomedical Technologies

Consortium of Universities Italy

EMBnet