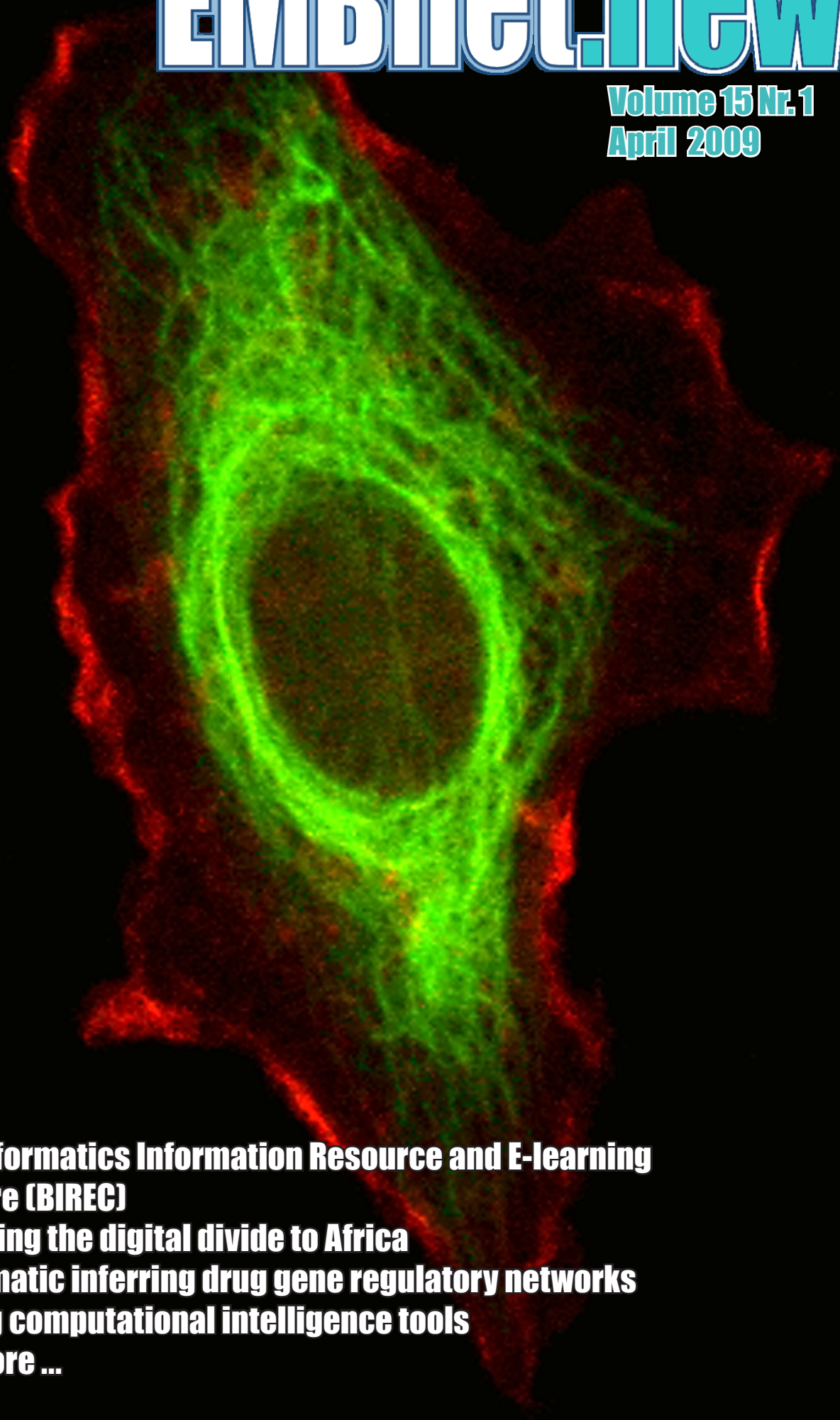
EMBnet.news

Volume 15 Nr. 1
April 2009



- Bioinformatics Information Resource and E-learning Centre (BIREC)
- Bridging the digital divide to Africa
- Automatic inferring drug gene regulatory networks using computational intelligence tools

and more ...

Editorial

In 2009 our readers will see major changes in this publication. The fist step is to structure the issues in several independent sections. The reason for that is connected with the need to separate news and announcements from the several types of contributions that we accept for publication, such as reports, technical notes, etc. The second step is to open away for publishing peer reviewed papers. As one can imagine this is far more difficult to set-up but is already under way, so that we can start this section in Volume 15, later in the year.

The present number is already structured. We hope that our readers like it better this way. We are also preparing special thematic issues with contributions on specific topics, on our request.

On its 21st year of existence, EMBnet is again adapting to change, but not loosing its focus on Bioinformatics users and their needs. EMBnet. news is on track to depict it for its readers. The number of downloads of this online publications is always increasing, a true measure that our effort is worthwhile.

We would like to remind you to visit the EMBnet website (http://www.embnet.org) where you can find instructions for authors. We welcome you to submit contributions to EMBnet.news and display relevant announcements for our community.

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

proteinspotlight

> ONE MONTH, ONE PROTEIN <

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

We provide the EMBnet community with a printed version of issue 103. Please let us know if you like this inclusion.

Contents

Editorial 2

Letters to the Editor

Win the Nobel, lose your lifetime work..... 3

EMBnet comes to Major Bioinformatics African Conference 4

Bridging the digital divide to Africa 6

Bioinformatics Information Resource and E-learning Centre (BIREC) 8

News and Announcements 9

Reports

AFBIX09: bioinformatics virtual conference..... 16

Technical Notes

Automatic inferring drug gene regulatory networks using computational intelligence tools .. 17

Protein spotlight 103 28

Node information 30

Editorial Board:

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

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

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

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

Andreas Gisel, Institute for Biomedical Technologies, CNR, Bari, IT
Email: andreas.gisel@ba.itb.cnr.it
Tel: +39-80-5929662
Fax: +39-80-5929690

Cover picture: *Confocal microscopy picture of a glioma cell stained for Vimentin (green) and Actin (red),* 1997 [© Erik Bongcam-Rudloff]

Win the Nobel, lose your lifetime work

Once upon a time, in a land far, far away, there was a scientist that worked on an edge-cutting field that no one else could understand. He was so eager to make it accessible to others that he spent a large amount of his work building tools that would make it so easy and simple to work on the bleeding edge that everybody, from the humble most to major Nobel prize winners, would be able to use them blind eyed.

Our hero wanted to share his work with everybody and made his tools available to any interested party, but as his works gained in popularity he discovered that maintaining and distributing his tools imposed a heavy tax reducing his ability to do further research. At this point he decided that the finest thing to do would be to transfer all these boring tasks to a private company that would care for distribution and maintenance, and even profit from giving support to users.

Time passed. As our hero had originally shared his works for free, everybody doing any further development had grown used to build on top of his work and then submit their improvements to the private company for distribution as well. This resulted in a nice benefit for all: scientists would concern themselves with development of new ideas, and the company would get this work for free and distribute it to its benefit. Much -you would think- as scientific journals get knowledge and content and spread it around.

In due time, our hero would receive a Nobel prize for his works in the advancement of Science.

But time passes for all. Society changed and now investors demanded even greater returns from their shares. The private company, strangled for giving more benefits resorted to more aggressive marketing: they decided that in order to increase margins they needed to become a monopoly getting rid of any competence and started to incorporate any published works and reduce access to their tools by any -even remotely possible- competitor.

Can you picture yourself as a Nobel laureate being banned from publication in a major journal if you ever publish anything on a competing publication? This is what happened to our hero: as he had continued developing his work, he was banned by the distributing company from using his own works less he be able to produce something better than what they were selling.

Our hero had won the Nobel prize. But he was unable to use the product of his own works to further develop the art.

The end?

It is a sad story that has repeated once and again. But although we often talk of scientists in a crystal tower, they are not complete fools. And as any tale, ours deserves to have a happy end too. So, let us give it a try: how would you finish this tale? Here's our take:

Many scientists, on seeing this behaviour decided to join forces and even sent papers denouncing these acts to major journals, and promoted discussion worldwide. They soon banded together and built new tools to substitute the now highly expensive behemoth that once had been free for all to share, and new scientific breakthroughs flourished everywhere. Of course, the company soon issued press releases to state it had all been a libelous misunderstanding.

Time passed by. It is now. Many of these alternatives have become new standards in scientific work, there is choice and a new golden age of invention... it has become the time for our new heroes to confront the same decision our humble Nobel had to face: how are they going to ensure continuity, maintenance and support of their works and keep doing new developments as well?

The end?

As we see it we have various solutions: you can take maintenance in your own hands -with a major tax in your time and research-, or you can defer it to someone else -with a tax on costs-. In the current context of instability and strict publication requirements, if you want to stay competitive you have no choice: you better rely on the help of others. Formerly there was only one option, a private company, but now you have more options: you may transfer your works to academic service providers, be they global, like EBI or local, like EMBnet, or you can give it away to the community in the hope someone else takes over. This would give you the best of both worlds: you will be able to meet strict publication requirements (like continued maintenance of your works and

URLs) at a minimal cost as you know that publicly funded academics do not work for a profit.

The major problem with any outsourcing solution is that you will depend on the continuity of your provider and will need to support them. As times get hard, this is getting more difficult to achieve each day, mostly so for single entities -like a private company- that may face the need to cut on lower interest services.

It is also difficult for distributed organizations, like EMBnet, and we are terribly sad to notify everybody of the final closing of the Austrian EMBnet node, that was definitively taken down last March. But, as as long as there is one node remaining, we will strive to keep continued maintenance of the services you entrusted us. We have already moved the unique services provided by EMBnet/AT to another node -and ensured the move was transparent by redirecting the URLs- so nobody should notice. Other, common services -like EMBOSS access- are available from other nodes, and Austrian scientists can rest assured that they will be able to continue using their favourite tool on other EMBnet nodes at their choice.

We think the time is coming for us all to acknowledge the value that the commitment of scientific services (centralized like EBI or distributed as EMBnet) provide to the community by avoiding aggressive commercial takeovers of tax-payer funded public developments and ensuring that the product of our shared work remains for ever free.

JR Valverde
on behalf of the EMBnet Executive Board

EMBnet comes to Major Bioinformatics African Conference

Arun Gupta[1], Nicola Mulder[2], Manuel Corpas[3]

[1] Abhyudaya BioSoft and DAVV, Indore

[2] National Bioinformatics Network Node, University of Cape Town (UCT) and University of the Western Cape, South Africa.

[3] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

Bioinformatics is a relatively affordable scientific discipline to establish as it requires intellectual capacity but not expensive laboratory facilities or equipment. This makes it a very accessible discipline to scientists in poorly resourced countries in Africa. The International Society for Computational Biology (ISCB) and the African Society for Bioinformatics and Computational Biology (ASBCB) have teamed up to organize a major meeting in Africa in 2009 focused on the theme "Bioinformatics of Infectious Diseases: Pathogens, Hosts and Vectors". This meeting, a new venture between ISCB and ASBCB and a follow on from a previous successful meeting held in Nairobi by the ASBCB, will be held this year in Bamako (Mali) from November 30 to December 3, hosted by the prestigious Malaria Research and Training Center, an important facility for malaria research in Africa. Although it will have a particular African focus, the meeting is intended to be a fully-fledged international event, encompassing scientists and students from leading institutions in the US, Latin America, Europe and Africa. By holding this event in Africa, we intend to stimulate local efforts for cooperation and dissemination

of leading research techniques to combat major African diseases.

Program

The meeting will consist of a 4-day conference followed by 2 days of practical workshops. The first 3 days of the meeting will include keynote presentations by 6 invited speakers from around the world. The last day of the conference will be a dedicated KAUST (King Abdullah University of Science and Technology) day focused on the topic "Systems view of biological organisms". KAUST has secured 20 full fellowships (valued at up to $1,700 each) to cover travel expenses, registration and accommodation for Africans attending the ISCB Africa ASBCB conference. Highly accomplished researchers will present the 2 days of post-conference tutorial workshops. Erik Bongcam-Rudloff, chairman of EMBnet, will give a keynote presentation and a tutorial during a workshop day.

Participation

We expect that most participants will come from Africa, and that the majority will be on the level of junior faculty, PhD students and post-doctoral researchers. Travel fellowships will be awarded to African researchers and students to cover travel and local expenses, with priority given to those selected for oral presentations through peer review of submitted research paper submissions. Several other travel fellowships will be secured for non-African participants. Fellowships will be entirely dependent on the funds available. Hence, prospective participants are encouraged to seek their own sources of funding.

Submissions

The conference will consist of a single track with oral and poster presentations. There will be a call for submission of abstracts (papers of up to 1000 words will be required for oral presentations); authors will indicate whether they aspire to give an oral or poster presentation. Graduate students, young investigators and all African researchers involved in the field will be strongly encouraged to submit an abstract describing their work. Submissions from all other parts of the globe are invited as well. The Scientific Committee will evaluate submissions, and the (few) selected abstracts from the oral presentation track will be invited for oral presentation, while others will be invited for poster presentation. Using this model we have previously established a powerful cadre of networking scientists – who have demonstrated that by meeting, they can establish networks of collaboration between Africa and the US and EU, as well as between African countries.

Proceedings

The first ASCBCB conference was published in a special issue[1,2] of Infection, Genetics and Evolution and). The organizers are in the process of negotiating and securing a publication for the proceedings of this 2009 conference.

Summary

The "ISCB-Africa ASBCB Joint Conference on Bioinformatics of Infectious Diseases: Pathogens, Hosts and Vectors" will provide a forum for discussion, fostering of new collaborations and the development of this nascent field in Africa, promoting more effective networking and novel research initiatives on the continent. It is expected that there will be an important presence of EMBnet key members and people from African EMBnet nodes. A key result will be to improve regional training as a direct consequence of sharing training techniques and material relevant to hosts, pathogens and diseases. The meeting will also facilitate links between young and emerging scientists from Africa, established leading African scientists and distinguished colleagues from the international community. The conference will be followed by two days of workshops for students and trainers, for which we have received commitment from several leading African, European and American researchers to serve as organizers and presenters.

Website for further details:

https://www.iscb.org/iscb-africa

1 http://dx.doi.org/10.1016/j.meegid.2008.09.002
2 http://dx.doi.org/10.1016/j.meegid.2008.09.003

Bridging the digital divide to Africa

Ian Moore
ICT manager for the International Livestock Research Institute (ILRI) and the World Agroforestry Centre (ICRAF). Nairobi, Kenya

Africa has always felt disconnected, or at most connected by a thin thread, to the digital world. In the past, many projects attempting to connect African countries by fibre optic cable have floundered at an early stage.

The IDRC map "The Internet: Out of Africa" on Fig. 1 shows the status of Internet connectivity per capita in 2002.

The larger the circle over a country the more bandwidth per person was available from within the country, mostly from satellite connections.

Only four fibre optic submarine cables landed on African soil and SAT3, the main West African cable, was not used to full capacity for many years due to poor infrastructure within the countries and poor management and marketing by incumbent telecommunication monopolies.

Since then, the availability of satellite connectivity has grown enormously but little has changed in terms of the fibre optic cables that connect Africa to the rest of the world.

The good news!

But all that is about to change! The second map "Sub-Saharan Africa undersea cables (2011)" from our friend Steve Song's blog site shows the eight undersea cable projects that are already underway and will be commissioned before the end of 2011.

The thickness of the line indicates the comparative bandwidth that will be made available. The West coast, in particular South Africa, Nigeria

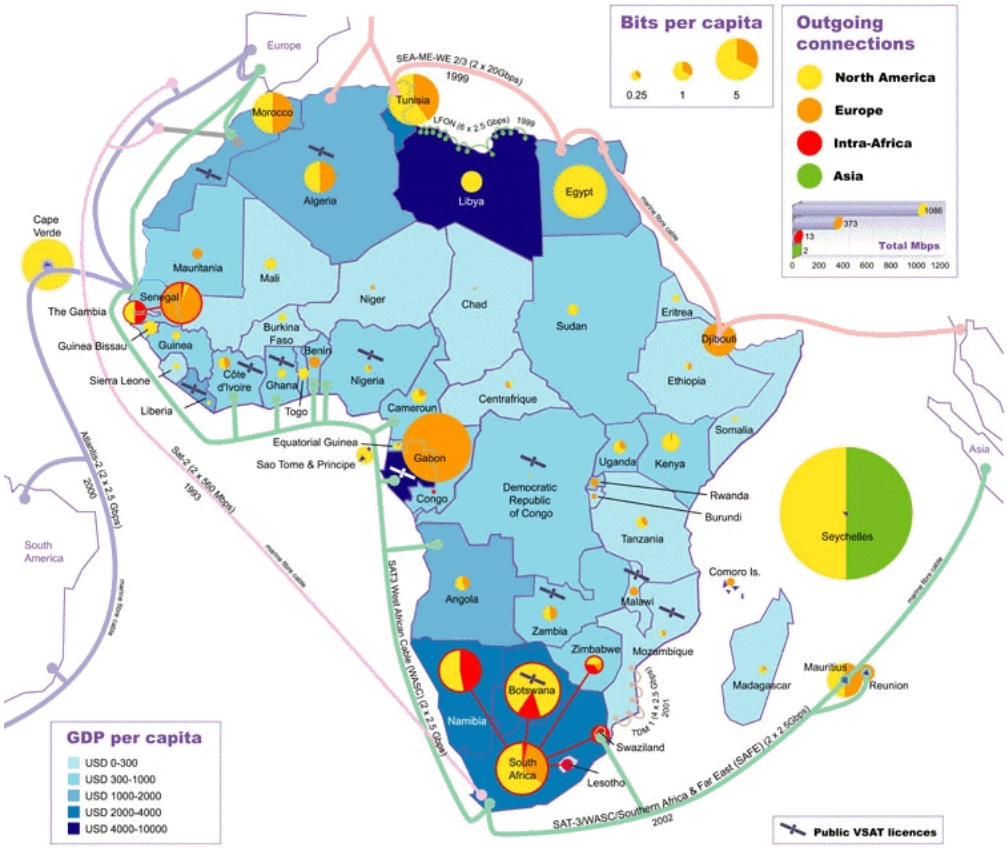


Figure 1. The IDRC map "The Internet: Out of Africa". in 2002.

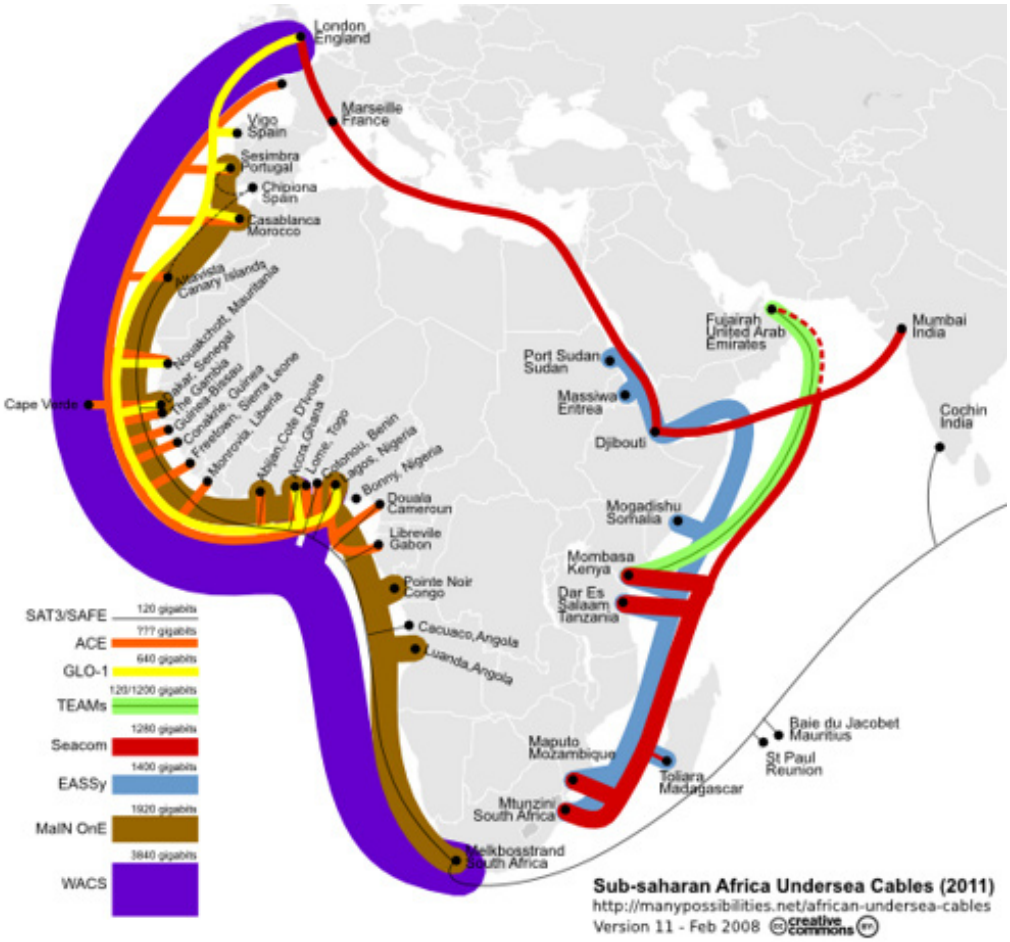


Figure 2. Sub-Saharan Africa undersea cables (2011).

and Ghana, are set to benefit most from this revolution, but the East coast will also be connected for the first time!

When compared to the thin black line of the original SAT3 cable you'll see that the planned explosion in available bandwidth driven by the telecommunication companies is huge.

Those who read the Kenyan newspapers will know that the red SEACOM cable is due to be commissioned in Kenya at the end of June 2009 and that the green TEAMS cable is not far behind. See: "Seacom steps up cable marketing" Daily Nation (Kenya) 23 February 2009[1].

And at the end of February 2009, the government of Ethiopia finally commissioned the cross border connection to Djibouti. This provides a much needed alternative to the unreliable fibre route through Sudan. It also means that Ethiopia can benefit from the SEACOM cable and eventually the blue EASSY cable that has been plagued and delayed by political infighting among the consortium members.

Sources:

1. Steve Song "Sub-Saharan Africa undersea cables (2011)"[2]
2. Canadian International Development Research Centre (IDRC) "The Internet: Out of Africa"[3]

1 http://www.nation.co.ke/business/news/-/1006/534028/-/j1s664z/-/index.html

2 http://manypossibilities.net/2009/02/undersea-cables-update-2/

3 http://www.idrc.ca/en/ev-6568-201-1-DO _ TOPIC.html

Bioinformatics Information Resource and E-learning Centre (BIREC)

An Initiative of the EMBnet Pakistan National Node

Nazim Rahman[1,2], Shahid Nadeem Chohan[1,3], Raheel Qamar[1,4]

[1] Department of Biosciences, COMSATS Institute of Information Technology, Islamabad, Pakistan
[2] CEGG, Faculty of Genetic Medicine and Development, University of Geneva, Switzerland
[3] School of Natural Sciences, University of Western Sydney, Australia
[4] Shifa College of Medicine, Islamabad, Pakistan

Introduction

Only two decades ago the bottleneck in life sciences research was scarcity of data. At that time the challenge was to generate new data. Nowadays, thanks to the high-throughput technologies, enormous quantity of data is available to our life scientists which are continuously being added on. The challenge is to extract useful knowledge from this vast collection of data made available internationally through the Internet.

This enormity of data poses another consequent challenge. It makes it difficult for the bioinformatics user to comprehend ever more complex tools databases, new functionalities, new algorithms, updated tools, new database fields, new layouts, and other complexities. Increasingly life scientists are asking questions such as:

- Which tool should I use? Why? And How?
- What are the limitations of this tool?
- Can I couple this tool with another? If so, how and under what conditions?
- How the data is generated and how reliable is it for my research?
-

Answers to such simple questions are difficult to find. Consequently, precious time and resources are wasted.

Moreover, there is an inherent problem with life sciences data. The researchers specializing in different domains are often interested in looking at the same data from different perspectives. For example, a researcher might only be interested in the nitrogen binding capabilities of a molecule while another would be interested in the molecule's role in a pathway. Thus presenting the same data to very different audiences from different perspectives is a daunting task. The Internet is full of thousands of resources and tutorials but they often target only a specific domain of life science researchers and rarely provide a global view. The desired information is available but it is not connected.

In an attempt to start addressing these questions, we have created BIREC.

Organization of BIREC

BIREC (Bioinformatics Information Resource and E-learning Center) was created to provide free, dynamic, collaborative, and solution-oriented bioinformatics information resources and an online self-learning platform. Our goal was to provide maximum amount of information in one place, update it regularly and present the information in an audience-centric format. This would allow students and researchers alike to master essential bioinformatics skills.

The simplest solution is to reclassify the data available on the Internet to facilitate bioinformatics users. Hierarchical classification methods are incapable of classifying highly interconnected data especially when classification changes with audience. Tag-based classification is ideal given these circumstances. In BIREC, all the content are classified using keyword tags, the contents are then organized dynamically based upon these tags. This method permits linking of relevant video tutorials (screen casts), FAQ, newsfeeds, eBooks, case studies, cheat sheets or any new type of data we would add in the future to be aggregated dynamically on one page accessible upon a click by the user.

At BIREC, we provide the following:

Resources: introduces bioinformatics databases and software.

Case Study: step-by-step instructions to solve a discrete problem.

Strategy: a collection of case-studies or methodology to solve complex problems.

eBooks: online books on specific topics.

Tutorials: tutorials on databases, tools, algorithms or methodology.

Cheat sheets: a small document to be used as a reference.

Newsfeeds: collection of news items relevant to specific topics or domains.

Relevant linked keywords are present on every page. They serve to dynamically reorganize content for the user.

Conclusion

BIREC provides an integrated access to Bioinformatics resources as well as systematic links to other available resources, elsewhere on the Internet. The focus of BIREC is on providing quality content in a user-friendly manner. New content is continuously being added and the older content is revised based on user feedback.

BIREC can be accessed at the following URL: http://www.birec.org/

For content submission requests, error corrections, and user feedback please contact: editor@birec.org.

BIREC is a project of the EMBnet Pakistan National Node at COMSATS Institute of Information Technology, Islamabad, Pakistan.

Announcement

GTPB

The Gulbenkian Training Programme in Bioinformatics

10th Edition

The GTPB runs continuously since 1999, more than 1200 international attendees have benefited from hands-on intensive training courses with optimized conditions for fast acquisition of skills. The Instituto Gulbenkian de Ciência in Oeiras, Portugal proudly hosts the Portuguese node of the EMBnet since 1991 and the GTPB since 1999.

As a yearly programme, GTPB tries to improve on course quality using the accumulated expe-

Training room

MEPA09 Group photo

rience and feedback. Themes for which there is continued demand for training are revisited each year, GTPB constantly introduces new ones, following recent developments and applications. The current edition (10th) offers a total of 19 courses.

Portugal is a nice and sunny destination and local subsistence costs are very moderate. This combination has attracted many students in Europe Africa and Asia.

The courses, held in a very efficient training room, also cover cross disciplinary themes such as Biostatistics, Computer Science, Modeling, etc., making them also attractive to established Bioinformaticians that never had the chance of exploring some of these "foundation-type" subjects formally or in sufficient depth.

The course website (http://gtpb.igc.gulbenkian.pt), has up-to-date information on the programmed activity throughout the year.

Course fees depend on the duration but our standard rate is Euro 80.00 per day including lunch.

Every attendee receives a CDROM that fully documents each course.

Course by the Finnish EMBnet Node

DNA expression microarray data analysis using R and Bioconductor

This course will cover the basics of R and Bioconductor. Participants are expected to have some background knowledge on DNA microarrays, although the basic datatypes and concepts are introduced on the course. No previous knowledge of R or Bioconductor is expected.

The course consists of lectures and hands-on exercises, the emphasis being on the exercises. The last day of the course is reserved for work with your own data. This is a good opportunity to practise R and Bioconductor even if you haven't conducted any microarray experiments, since we can provide you with some demo data.

The course will run daily from 9.00 to 16.00, except that the course will start on the first day at 10.00.

If you're not at all familiar with R, please read this 8-paged introduction to R[1] before attending the course.

Download lectures[2], exercises[3], data[4] and R-2.6.2[5].

Program

Day 1 (with approximate times)
10.00-12.00 Basics of R and Bioconductor
13.00 -14.30 Affymetrix preprocessing and quality control
14.30-16.00 Illumina preprocessing and quality control

Day 2
9.00-11.00 cDNA preprocessing and quality control
11.00-12.00 Basic statistics
13.30-16.00 Finding differentially expressed genes

Day 3
9.00-11.30 Annotation
12.30-14.30 Clustering and visualization
14.30-15.00 Experimental design
15.00-16.00 Wrap-up

Day 4
9.00-16.00 Work with your data

Date: 02.06.2009 10:00 - 05.06.2009 16:00

Location: Premises of CSC, Keilaranta 14, Keilaniemi, Espoo.

Language: English

Lecturers: Dario Greco (HY), Panu Somervuo (HY), Jarno Tuimala (CSC)

Price: for Finnish academics 240 EUR / for governmental research institute staff 380 EUR / for others 520 EUR, prices excl. VAT 22%

The fee includes course materials and morning and afternoon coffee.

Registration

Registration form:
http://www.csc.fi/csc/kurssit/arkisto/bioc2009/event_registration_form
Registration by 01.05.2009 23:55.

Additional information

There are 24 seats on the course, and seats are allocated on a "first come, first served" basis. Course will be arranged if there are at least 12 registrations by 1st of May, 2009. Participants will get a course certificate after attending the course.

A confirmation email will be sent to the participants about one week before the course. Participants can cancel their registration at latest three busines days before the course without extra costs. Cancellation after that is possible, but the whole course fee will be charged.

Bills will be sent to the participants after the course by mail (not by email) or as an electronic bill.

For information, please contact Jarno Tuimala (09-457 2226 at Jarno.Tuimala [at] csc.fi).

1 http://extras.csc.fi/biosciences/courses/bioconductor/quick_R.pdf
2 http://extras.csc.fi/biosciences/courses/bioconductor/all-lectures.pdf
3 http://extras.csc.fi/biosciences/courses/bioconductor/all-exercises.pdf
4 http://extras.csc.fi/biosciences/courses/bioconductor/data.zip
5 http://extras.csc.fi/biosciences/courses/bioconductor/R-2.6.2h.zip

Computer superpower strengthens attempts to combat common diseases

Ingela Nystrom
Uppsala University, Uppsala, Sweden

New large-scale sequencing technology will revolutionize biomedical research in the coming decade. Uppsala University's entity UPPMAX is now expanding its operations and providing researchers with a powerful system for large-scale compute and storage of data, which can lead to new breakthroughs in research on our public-health disorders.

Among other things, the new large-scale sequencing technology offers researchers the opportunity to understand the impact of the genome on the genesis of common diseases. Questions can be posed in a different way with more large-scale methods. For instance, the technology renders it possible to map all the bacteria in a person's mouth, to see why one individual develops malaria while another does not, and how the malaria parasite adapts in order to elude people's immune defences. It can also involve cataloguing all DNA modifications in a cancer cell. Furthermore, scientists have found regions in our genes that increase the risk of various common diseases such as cancer, diabetes, obesity, and autoimmune diseases. These regions were identified through the use of so-called SNP chips, but since then it has proven difficult to find the actual mutations that cause disease.

"The new sequencing methods supported by this funding offer tremendous potential for finding many of these mutations. Knowledge about the mutations and disease mechanisms will enable development of better, more targeted drugs," says Kerstin Lindblad-Toh, professor of comparative genomics.

Since extremely large quantities of data are produced in such studies, many terabytes of both data storage and primary memory in order to be able to deal with and analyze these data. Kerstin Lindblad-Toh, in collaboration with

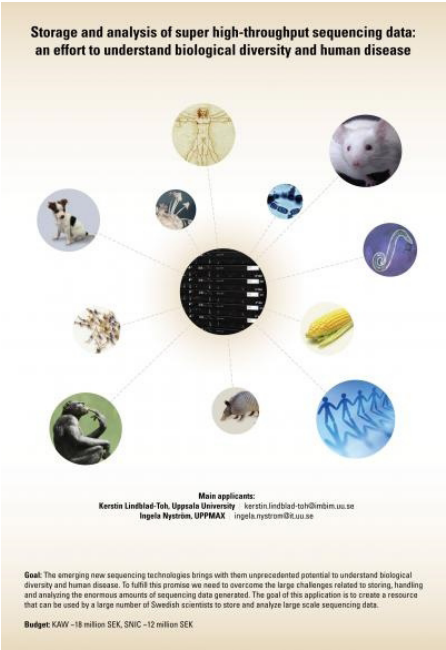


Figure 1. Uppsala University is aiming to use computer power for life science.

UPPMAX director Ingela Nyström, has led the research team at Uppsala University that was recently granted SEK 13 million from the Knut and Alice Wallenberg Foundation (KAW) to construct a computing system that meets the needs of the sequencing platform that already exists at Uppsala University. At present there are three new technology sequencing machines, which make Uppsala University the largest player in the Nordic countries in this field. This position will now be strengthened even further.

Large-scale computations and large-scale storage, processing, and analysis of data play an ever greater role in many scientific fields. The Swedish National Infrastructure for Computing (SNIC) announced, together with KAW, that funding was available for resources for framework of SNIC's general resources. KAW contributed investment capital and SNIC operational and user support. Two of eight applications submitted were granted.

"Uppsala University hit the jackpot, since both projects have Uppsala researchers involved. One is for research on new energy materials that Professor Olle Eriksson is conducting, with colleagues, and the other is this visionary genome-sequencing research," says Ingela Nyström.

The new computer system for DNA sequencing will be located at Uppsala Multidisciplinary

Center for Advanced Computational Science (UPPMAX) and will be run by the Center's systems experts. Since its establishment in 2003, UPPMAX has provided researchers, both locally and nationally, with computational power from a number of computer clusters. A previous allocation from SNIC in 2008 is earmarked for a new cluster of some 2000 computing cores and is under procurement. This new grant will add yet another cluster.

"What's more, our activities are expanding to include extensive data storage, some 500 terabytes. Some data will have to be stored up to ten years, which places special demands on the technology," says Ingela Nyström.

SNIC has six member centers in Sweden (from north to south: HPC2N in Umeå, UPPMAX in Uppsala, PDC in Stockholm, NSC in Linköping, C3SE in Göteborg, and Lunarc in Lund). With the two latest grants to UPPMAX, Uppsala University will play a significant role in providing Sweden's researchers with adequate infrastructure.

A dozen leading researchers collaborated on the application in order to attain the common goal of satisfying the need to be able to deal with the enormous quantities of data from modern sequencing technology. The researchers, Kerstin Lindblad-Toh, Ulf Gyllensten, Ann-Christine Syvänen, Leif Andersson, Siv Andersson, Rolf Ohlsson, Claes Wadelius, Erik Bongcam-Rudloff, Helgi B. Schiöth, Hans Ronne, and Joakim Lundeberg, all work with different biological problem complexes and submitted a joint application in order to have the new technologies function as efficiently as possible.

Read more about the project[1] oriented toward new energy materials.

For more information, please contact

Professor Kerstin Lindblad-Toh
kersli@broad.mit.edu
phone: +1 617 223 7476
(currently in the US)

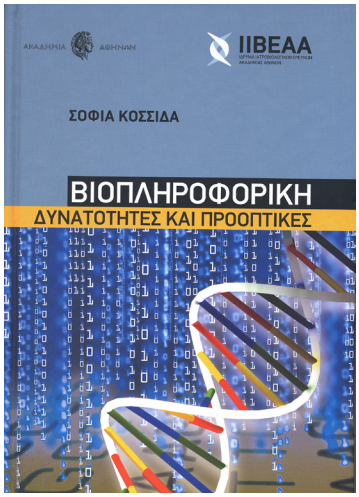
UPPMAX director Ingela Nyström
ingela.nystrom@it.uu.se
cell phone: +46 (0)70-167 9045

UPPMAX chief systems expert Jukka Komminaho
Jukka.Komminaho@it.uu.se
phone: +46 (0)18-471 1024.

1 http://www.uu.se/press/pm.php?typ=pm&id=498

Bioinformatics: potentials and perspectives

The EMBnet is proud to announce the publication, by Dr. Sophia Kossida, node manager of the EMBnet National Node in Greece, of the book entitled "Bioinformatics: potentials and perspective".



The book[1] has been written by Sophia with the contribution of the scientific team at Biomedical Research Foundation of the Academy of Athens[2] and represents the first Greek book written by Greek scientists concerning bioinformatics.

The 14 chapters of this book are assembled harmonically in order to offer the reader the basic knowledge to understand the specialized issues of bioinformatics such as proteomics, system biology, biological databases etc. This book could be useful for students that want to be familiarized with the basic concepts of bioinformatics, educators who want to have their knowledge updated and researchers who work on the same or similar scientific issues.

1 http://www.bioacademy.gr/bioinformatics/bioinfo _ book/bioinformatics _ home.htm
2 http://www.bioacademy.gr/bioinformatics/

Max Planck Institute for Informatics

max planck institut informatik

Max Planck Scholarship for Women in ComputerScience

Description

Max Planck Scholarship for Women in Computer Science

Requirements

PhD in computer science.

Application deadline

October 31st 2009

Contact

csscholar@mpiinf.mpg.de

The Max Planck Institute for Informatics (MPI-INF) is inviting applications from female postgraduate computer scientists working in computational biology for a two-year research scholarship of up to €72,000. The MPI-INF encourages women to excel in computer science and become active role models and leaders in the field. Scholars will be chosen based on the strength of their academic background and research credentials.

The Max Planck Society (MPG) is the premier basic research organization in Germany with more than 30 Nobel Prize winners, including Prof. Christiane Nüsslein-Volhard - one of the few female Nobel Laureates. The research institutes of the Max Planck Society have an international reputation as "Centers of Excellence" for foundational research. Research of the computational biology division at the Max Planck Institute for Informatics ranges from medical bioinformatics, with a focus on HIV, HCV, Influenza and polygenic diseases, to the prediction and analysis of protein structures, molecular docking and drug screening as well as analysis of genomics, transcriptomics and metagenomics data. Additional research fields in the division include functional prediction of proteins based on genomics and proteomics data as well as analysis of protein-protein interactions, genetic variation and epigenomics. Research will continue to span new computational methods as well as innovation in biology and medicine.

Applications should include a resume, cover letter, publication list and two references. Please refer to "MPI-INF Scholarship" in your application. The MPI for Informatics strives to provide a family-friendly work environment with options for day care of smaller children and other means of support for working mothers.

Send applications via Email to:

csscholar@mpi-inf.mpg.de

Max Planck Institute for Informatics
Campus E1 4
66123 Saarbrücken
Germany
WWW: http://www.mpi-inf.mpg.de/

Workshop by the Italian EMBnet node

Next Generation Sequencing

September 16-18, 2009 - Bari (Italy)

Next Generation Sequencing (NGS) methods are revolutionizing genomic and functional-genomics research. However, the manipulation and interpretation of the huge amounts of data produced present significant computational challenges.

The workshop will provide an overview of current ultra-high throughput sequencing platforms and introduce NGS data analysis strategies for genomic sequencing, transcriptome analysis ChIP-Seq and other applications.

Topics will include:

- Next Generation Sequencing platforms (Roche 454, Illumina Solexa, ABI SOLiD)
- Strategies for mapping NGS reads to reference sequences
- Genome re-sequencing and assembly, SNP discovery
- Clustering of reads from expressed sequences (peak finding)
- Integration of gene expression data and genome annotation
- Estimation of gene expression levels and profiles
- Analysis of splicing and alternative splicing
- Small RNA discovery and characterization
- Analysis of genome methylation
- Chip-seq data analysis

The workshop will include seminars to illustrate the theoretical basis of the computational approaches and analysis workflows based on real examples.
Practical sessions will follow, where the participants will use state-of-the art tools for real data analysis.

Workshop Scientific and Organizing Committee

Graziano Pesole, University of Bari and ITB-CNR, Bari (Workshop Chair)
Ernesto Picardi, University of Bari
Elisabetta Sbisà, ITB-CNR, Bari
Sabino Liuni, ITB-CNR, Bari
Apollonia Tullo, ITB-CNR, Bari
Domenica D'Elia, ITB-CNR, Bari
Tiziana Castrignanò, CASPUR, Rome
David Horner, University of Milan
Giulio Pavesi, University of Milan
Flavio Mignone, University of Milan

Workshop Secretary and Organization

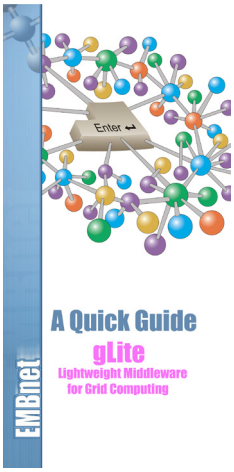
Angela Evangelista, University of Bari

The number of participants is limited to 80 for the lectures and to 30 for the practical sessions. The registration form is available at the Workshop web site:
http://mi.caspur.it/workshop_NGS09/index.php

For information please contact
a.evangelista@biologia.uniba.it

A Quick Guide to gLite 3 commands

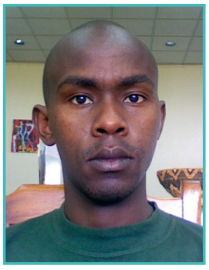
The EMBnet is proud to announce the publication, by José R. Valverde, node manager of the EMBnet National Node in Spain, of the Quick guide to gLite 3 commands[1].



gLite (http://www.glite.org/) is the middleware used in the EGEE Grid. It allows you to manage jobs and data on the Grid. The goal of this guide is to get you quickly started and is not an exhaustive list of available commands.

1 http://www.embnet.org/QuickGuides

AFBIX09: bioinformatics virtual conference



Nelson Ndegwa
ILRI, Nairobi, Kenya

The Afbix'09 virtual conference on Bioinformatics, which is the first of its kind in Africa, was held on the 19th-20th February 2009, and was organized by the Regional Student Group Africa (RSG Africa), Regional Student Group Morocco (RSG Morocco) – affiliates of the International Society for Computational Biology Student Council[1] (ISCBSC), the Bioinformatics Organization[2] and the African Society of Bioinformatics and Computational Biology[3] (ASBCB).

The conference had five virtual hubs drawn from SANBI-South Africa, ILRI-Kenya, SMBI-Morocco, Covenant University-Nigeria and University of Notre Dame-USA.

The ILRI-Kenya hub had a total of 40 participants drawn from various institutions and organizations in Kenya, and was sponsored by the ILRI/BecA Bioinformatics platform headed by Dr. Etienne deVilliers. The two day conference had keynote addresses and poster sessions from scientists and students drawn from Africa and abroad and followed four themes;

Figure 1. Prof. Erik Bongcam-Rudloff addressing participants during the conference break. He visited ILRI as part of his capacity building activities in Bioinformatics.

1 http://www.iscbsc.org/
2 http://wiki.bioinformatics.org/Bioinformatics _ Organization
3 http://www.asbcb.org/

Figure 2. Dr. Etienne de Villiers, the Bioinformatics Group leader at the ILRI/BecA bioinformatics platform[1], informing participants on an upcoming two weeks Introduction to bioinformatics course , which they conduct annually with Prof. Erik Bongcam-Rudloff at ILRI.

1 http://hpc.ilri.cgiar.org/training.html

1. Structural Biology Applied to Infectious Diseases.
2. Applied Genomics to Infectious Diseases.
3. Career development in Bioinformatics & opportunities for researchers in Africa.
4. Proteomic applications to Tropical diseases.
5.

Among the objectives of the conference was to provide scientists from developing countries with an opportunity to publish in leading journals, thus high quality articles presented during the conference will be submitted to the Journal of Bioinformatics and Computation Biology (JBCB) and Bioinformation for publication.

During the conference Dr. Erik Bongcam-Rudloff, the chair of EMBnet, addressed the participants at the ILRI-Kenya hub about EMBnet and its contribution to Bioinformatics capacity building in Africa.

More detailed information on the conference: http://wiki.bioinformatics.org/Afbix09.

Automatic inferring drug gene regulatory networks using computational intelligence tools

Alexandru G Floares[1,2]

[1] Solutions of Artificial Intelligence Applications, Cluj-Napoca, Romania
[2] Department of Artificial Intelligence, Oncological Institute Cluj-Napoca, Cluj-Napoca, Romania

Email: alexandru.floares@iocn.ro

Abstract

Background: Mathematical modelling gene regulating networks is important for understanding and controlling them, with various drugs and their dosage regimens. The ordinary differential equations approach is probably the most sensible. Unfortunately, this is also the most diffcult, 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.

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 (GP) and neural networks feedback linearization (NN FBL). 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 di®erential 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 di®erential 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.

Background

The deluge of complex, high-dimensional data in bioinformatics is continuously increasing; however, our modelling capacity is much smaller and increasing only slowly|particularly in fields using high-throughput techniques such as genomics, transcriptomics, and proteomics. Knowing which genes are expressed, when, where, and to what extent is important for understanding organisms, as well as for controlling genes through adequate drug dosage regimen development. The regulation of gene expression is achieved through complex regulatory systems|gene regulatory networks (GRNs)|which are networks of interactions among DNA, RNA, proteins, and small molecules.

Usually the concentration profiles are not available for all these molecular species because they are not measured simultaneously in the same experiment. Inferring GRN models only from microarray time-series data is one of the most challenging tasks of bioinformatics. Various formalisms, such as Bayesian networks, Boolean networks, differential equation models, qualitative differential equations, stochastic equations, and rule-based systems, have been used, each approach having its merits and drawbacks (see [1], [2], [3] for reviews).

The ordinary differential equations (ODE) approach tries to elucidate a deeper understanding of the exact nature of the regulatory circuits and their regulation mechanisms. In a pharmacogenomic context, it allows the design of controls that are optimal, individualized drug dosage

regimens ([4, 5]). Unfortunately, this is also the most difficult, tedious, expensive, and time-consuming approach. The models are high-dimensional systems of nonlinear coupled stiff ODEs.

The number of parameters is extremely large, and many of them have unknown values. Although in principle one can find the best set of parameter values by sampling the whole parameter space, many degenerate solutions may be expected. The correlations between parameters and that biological systems have built-in regulation mechanisms that make them robust to changes in many of their parameter values. These facts suggest that it is the network structure rather than the precise value of the parameters that confers stability to the system.

We introduced the more general concept of drug-gene regulating networks (DGRNs), where the regulation is exerted not only by transcriptions factors (TFs), like in GRN, but also by drugs ([6]). We proposed a reverse-engineering (reversing) algorithm, based on linear genetic programming ([7]), producing ODE systems from data, and we applied it to DGRN and GRN ([6]), as well as to the subthalamopallidal neural networks of basal ganglia ([8]).

This algorithm automatically discovers the structure, estimates the parameters, and allows the identification of the molecular mechanisms involved. It starts from either experimental or simulated time-series data and produces systems of ODEs. We tested the algorithm on data from model simulation because this allows evaluation of the resultant model against the true known structure. The accuracy of the results and the CPU time were very good, most of the experiments reaching the best performances. This is mainly because the problem of reversing systems of ODEs is reduced to the problem of reversing individual algebraic equations. Therefore, evaluating the fitness function is not necessary to integrate the ODE system. This approach drastically reduces the execution time. Also, the possibility of incorporating common domain knowledge reduces the structure search space and further speeds up the algorithm. Other studies, proposing GRN reverse-engineering algorithms based on evolutionary computation, require integration of the ODE systems hundreds or thousands of times for each generation ([9], [10], [11], [12]). Similar methods have been proposed in the past, but most of them require a predefined model structure, however, and are limited to parameter estimation. For example, the S-system model refers to a particular type of ODE system in which the component processes are power-law functions ([13], [14]). Despite the elegance and computational simplicity of the S-system model, this formalism has its limitations for biochemical networks (e.g., [15]).

The reversing ODE systems (RODES)algorithm can reveal if some information from the input set is either missing or not related to the output. This is possible because RODES requires the temporal series of all variables of the system to infer an accurate mechanistic model. It also means that it does not discover false input-output relations.

For this report we investigated GRNs and DGRNs with regulated mRNA transcription and unregulated mRNA degradation, when either the temporal series of the regulatory transcription factor (in GRN) or those of the drug-receptor complex (in DGRN) are missing. Doing so requires extending RODES such that it can cope with missing information. The tricky solution consists of transforming the modelling problem in a tracking control problem. The measured mRNA temporal series become the desired or reference trajectories. The problem is to find the control(s) such that the plant output|the solution of the mRNA ODE|tracks the desired trajectory with an acceptable accuracy level. For this goal we used the neural network version of feedback linearization (see [16] and the literature cited there). These controls are the missing variables of the (D)GRNs that are identified in this way. To the best of our knowledge, this is the first realistic reverse-engineering algorithm, based on linear GP and NN FBL, for large DGRNs and GRNs.

Methods

DGRN and GRN fundamental ODE patterns and building blocks

The rate at which the concentration of a protein changes inside a cell depends mainly on the following:

1. the rate at which its mRNA is produced and degraded,
2. the rates at which the mRNA molecules are translated, and
3. the rate at which the protein itself degrades.

Usually, these rates have the same mathematical form as the pharmacokinetic (PK) blocks de-

scribing the drug movement into, within, and out of the body:

1. Zero order: $dX/dt = k$, where k is a zero-order rate constant and X is the concentration of the drug;
2. First order: $dX/dt = k \cdot X$, where k is a first-order rate constant and X is as above; and
3. Michaelis-Menten: $dX/dt = V_m \cdot X/(K_m + X)$, where V_m is a maximum rate, K_m is the Michaelis constant, and X is as above.

Some of the rates of mRNA production are regulated by TFs in GRN. In a pharmacogenomic context, new regulatory interactions, or exogenous control factors represented by drugs, are added. In ([6]) we introduced the more general concept of DGRNs. If the regulation is restricted to transcription factors, the network is a GRN. If the regulation can be exerted by transcription factors and by drug-related compounds, e.g., drug-receptor complexes, the network is a DGRN. Thus, GRN could be considered a subset of DGRN. The mathematical descriptions of the mechanisms of regulation, by TFs and drug-related compounds, are the same. They have the form of the most common pharmacodynamic (PD) blocks, describing the relationship between drug doses or concentration and effects:

1. Linear (stimulation [+] or inhibition [-]) model: $E = E_0 \pm S \cdot C_e$
2. Log-linear (stimulation [+] or inhibition [-]) model: $E = E_0 \pm S \cdot \log (C_e)$
3. Ordinary ($\gamma = 1$) or sigmoid ($\gamma > 1$) Emax (stimulation [+] or inhibition [-]) model:

$$E = E_0 \pm E_{max} \cdot C_e^{\gamma} /(C_e^{\gamma} + EC_{50}^{\gamma})$$

where E is the effect variable; E_0 is the baseline effect; E_{max} is the maximum drug-induced effect, also called capacity; EC_{50} (sometimes IC_{50} [50% inhibitory concentration] is used instead of EC_{50} for inhibitory effect) are the plasma concentration at 50% of maximal effect, also called sensitivity; S is the slope of the line relating the effect to the concentration; C_e is the concentration to which the effect is related, and γ is the sigmoidicity factor (Hill exponent). GRN and DGRN ODE systems models have one ODE for each mRNA and protein, corresponding to transcription and translation. Their structure results from the following:

1. summing up the pharmacokinetic blocks and
2. multiplying the rate constants of the regulated processes by pharmacodynamic blocks.

Usually, it is assumed that other processes, such as diffusion and transport, are fast with respect to transcription and translation and may thus be ignored. Thus, the rate of change in a specific mRNA concentration ($mRNA$), and in the translated product concentration (e.g., a transcription factor, TF, in our case) are

$$\frac{dmRNA}{dt} = k_{sm} \cdot R_s - k_{dm} \cdot R_d \cdot mRNA \quad (1)$$

$$\frac{dTF}{dt} = k_{sTF} \cdot mRNA - k_{dTF} \cdot TF \quad (2)$$

where k_{sm} is the rate at which mRNA is produced and k_{dm} is the mRNA degradation rate constant; k_{dTF} is the TF degradation rate constant, and k_{sTF} is the average TF translation rate constant. R_s and R_d are generic notation for different regulatory factors of mRNA synthesis and degradation, respectively. Usually, R_s represents TFs regulating mRNA synthesis and R_d represents drug-related compounds, e.g., a drug-receptor complex. A regulatory factor $R_{s,d} = 1$ indicates no regulation, and an $R_{s,d}$ having the form of one of the pharmacodynamic blocks indicates the action and the mechanism of action of a regulatory factor.

Equations 1 and 2 together with the above PK and PD blocks form fundamental ODE patterns or building blocks of the (D)GRN models. This common domain knowledge can be simply used to reduce the structure search space of the algorithm, and to identify the biochemical mechanisms involved, in the resultant model. Three cases, of increasing complexity, are possible:

1. unregulated mRNA transcription and unregulated mRNA degradation,
2. regulated mRNA transcription and unregulated mRNA degradation, and
3. regulated mRNA transcription and regulated mRNA degradation.

For unregulated transcription and degradation ($R_s = 1$, $Rd = 1$) all variables (mRNAs) are available and one can use RODES to automatically infer the corresponding ODE. For regulated transcription and missing information about the TFs, RODES must be extended and this is the main goal of this report. Usually, while the equations' structure is known, and the parameters' values can be found in the literature or in public databases, only the temporal series of the mRNAs are

available from microarray experiments, but not those of the TFs.

RODES algorithm: no missing information

The goal of the proposed algorithm for reverse engineering is threefold:

1. to automatically identify the structure of accurate ODE systems models of GRN and DGRN,
2. to automatically estimate their parameters, and
3. to identify the biochemical and pharmacological mechanisms involved.

The RODES algorithm starts from experimental or simulated time-series data. The name of the algorithm is related to its results, not to the biological systems investigated. This is because we successfully applied it to various biological networks: the subthalamopallidal neural network of the basal ganglia ([8]) and the vascular networks of tumors (work in progress).

The result, but also the simulated data generator, is an ODE system, $dX/dt = f(X)$. In the time-series data, at any given discrete time point, t, where $t = 1, 2, \ldots, T$, dX/dt is equal to $f(x)$ at the same time point t. Equivalently, for any individual ODE of the system, dXi/dt (at t) $= fi(X)$ (at t), where $i = 1, 2, \ldots, n$ is the number of variables. For simulated data, the true structure of the (D)GRN models is known, and this allows a faithful evaluation of the predicted models. We therefore used simulated time-series data to test our algorithm. RODES is based on genetic programming, as a machine learning method, and consists of the following steps:

1. Compute the time derivative of each variable, $dXi=dt$, at all discrete time points t:
   (a) differentiate each variable with respect to time for simulated data;
   (b) first a function to smooth the data, and then differentiate it, for noisy experimental data.
2. Build input-output pairs, (Xi; $dXi=dt$), at the corresponding discrete time points t:
   (a) use all variables supposed to belong to the right-hand side of the reconstructed ODE as inputs
   (b) use the time derivative of one of the variables as output, if the GP implementation accepts many inputs but only one output, or
   (c) use the time derivatives of all the variables as output, if the GP implementation accepts many inputs and many outputs.
3. Build training, validation (optional, to avoid overfitting), and testing sets from the input-output pairs.
4. Initialize a population of randomly generated programs, coding mathematical models relating the inputs Xi to the output(s) dXi/dt.
5. Perform a tournament contest:
   (a) Randomly select four programs and evaluate their fitness (mean squared error)|how well they map the input data Xi to the output data dXi/dt.
   (b) Select two programs as winners and the other two as losers.
   (c) Copy the two winner programs and transform them probabilistically by:
      i. exchanging parts of the winner programs with each other to create two new programs (crossover) and/or
      ii. randomly changing each tournament winner to create two new programs (mutation).
   (d) Replace the loser programs with the transformed winner programs. The winners of the tournament remain in the population unchanged.
6. Repeat steps 5(a)-5(d) until a program is developed that predicts the behaviour sufficiently.
7. Extract the ODE model from the resultant program or directly use it.

Steps 1-3 reduce the problem of reversing a system of coupled ODEs, $dX/dt = f(X)$, in that of reversing individual, decoupled, algebraic equations, $dX_i/dt = f_i(X)$. Even though the output is in reality a time derivative, dX_i/dt, the algorithm is simply searching for an algebraic equation relating the inputs to the output, at each discrete time point t. The corresponding relation is the predicted function, $\hat{f}_i(X)$, for the right-hand side of each differential equation of the system.

This approach drastically reduces the CPU time of the algorithm, by orders of magnitude, because in step 5(a) the fitness evaluation does not require the integration of the ODE system. More precisely, one can use a fitness function based on (e.g., [16]):

$$E_j = \sum_{i=1}^{n}\sum_{t=1}^{T}(\hat{X}_i(t) - X_i(t))^2 \quad (3)$$

where j is the number of programs, n is the number of variables, T is the number of sampling points, $\hat{X}_i(t)$ is the numerically calculated time course of the variable X_i at time t from the ODE system predicted by the program j, and $X_i(t)$ represents the experimentally or simulated time course of X_i at time t. Therefore, for every program's fitness calculation, at each generation, the ODE system must be numerically integrated. We used a fitness function of the form

$$E_j = 1/T \sum_{i=1}^{T} (d\hat{X}_i(t)/dt - dX_i(t)/dt)^2 \quad (4)$$

where j and T are as above, $d\hat{X}_i(t)/dt$ is the time derivative at time point t of the variable X_i predicted by the program j, and $dX_i(t)=dt$ represents the time derivative at time t of the experimental or simulated variable X_i calculated in step 1 of the algorithm.

While the time needed to integrate a system of ODE seems negligible, during ¯tness evaluation the integration has to be executed hundreds or thousands of times per generation. These, and the results of our present and previous studies ([4], [5], [8]), suggest that RODES will scale up well, as required by modern high-throughput biomedical techniques. We used a linear version of a steady-state genetic programming proposed by Banzhaf (see [7] for a detailed introduction and the literature cited there). In linear genetic programming the individuals are computer programs represented as a sequence of instructions from an imperative programming language or machine language. Nordin introduced the use of machine code in this context (cited in ([7])). The major preparatory steps for GP consist of determining

1. the set of terminals (see Table 1),
2. the set of functions (see Table 1),
3. the fitness measure (see equation 4),
4. the parameters for the run (see Table 1),
5. the method for designating a result, and
6. the criterion for terminating a run.

The function set, also called instruction set in linear GP, can be composed of standard arithmetic or programming operations, standard mathematical functions, logical functions, or domain-specific functions. Using simple and common domain knowledge, such as the set of mathematical functions that appear in the models, e.g., arithmetic functions but not trigonometric function, is enough for RODES to find the proper structure of the reconstructed equations, also greatly increasing execution speed. The terminals are variables and parameters. In microarray experiments, the number of mRNAs is usually of the order of 102 or 103 after filtering, but the number of the clusters of genes with similar temporal signatures is small One needs only to discover this small number of prototype ODE structures. All the equations have one of these prototype structures, and the equations in the same cluster have the same structure but different parameter values. We still do not know which are the input variables for each mRNA ODE equation. From the fundamental ODE patterns of DGRN, we know that the equations for each mRNA (see equation (1)) contains a synthetic and a degradation term. The inputs variables for these mRNA equations are:

1. the mRNA concentration|in a degradation term proportional with mRNA concentration|for unregulated transcription and degradation,

Table 1. Genetic programming parameters.

Parameter	Setting
Population size	500
Mutation frequency	95%
Block mutation rate	30%
Instruction mutation rate	30%
Instruction data mutation rate	40%
Crossover frequency	50%
Homologous crossover	95%
Program Size	80-128
Demes	
Crossover between dem	0%
Number of demes	10
Migration rate	1%
Dynamic Subset Selection	
Target subset size	50
Selection by age	50%
Selection by di±culty	50%
Stochastic selection	0%
Frequency (in generation equivalents)	1
Function set	f+, -, *, /g
Terminal set	64 = j + k
Constants	j
Input	k

2. the concentration of a transcription factor (for GRN) and/or of a drug related compound (for DGRN)|in a PD block (R_s in eqn (1) multiplying a PK block (the constant mRNA synthesis)|for regulated transcription and unregulated degradation, and
3. as above but also the concentration of a drug-related compound (for DGRN), contained in a PD block (R_d in eqn (1) multiplying a PK block (the linear mRNA degradation)|for regulated transcription and regulated degradation.

The RODES version described in this section requires all inputs to be available. This condition is certainly true for the ¯rst situation but is usually false for the second and the third. The next section will extend RODES to cope with the second situation.

Because we know the structure of this ODE (see eqn (1)), this is also the route to automate the discovery of the biochemical and pharmacological molecular mechanisms involved. Analysing the resultant equations, one can easily identify

1. cellular processes such as syntheses and degradations and their mechanisms as PK blocks,
2. the presence of regulation and
   (a) which are the regulated processes|their rate constants are multiplied by PD blocks,
   (b) which are the regulatory factors|transcription factors for GRN, drugs, or both for DGRN|the corresponding PD blocks can be functions of the TF concentrations or drug-related compound concentrations, respectively (see eqn (2)),
3. the regulation mechanisms|by looking at the corresponding PD blocks and at the rate constants they are multiplying.

There are situations in which the PK/PD mechanisms in the resultant mathematical model need to be clarified. When we have the product of two or more constants, in the symbolic form of the model, the algorithm will find only one numerical value. Using elementary domain knowledge, one can easily and clearly identify the PK/PD mechanisms (see ([5] for details)).

RODES - neural network feedback linearization for missing variable identification

RODES can reveal if some information from the input set is either missing or not related to the output. This is possible because RODES requires the temporal series of all variables of the system, to infer an accurate mechanistic model. It also means that it does not discover false input-output relations. We used these capabilities to automatically discover the subthalamopallidal network's structure of the basal ganglia and to investigate its connectivity, in ([8]). They can also be used in (D)GRN to remove variables unrelated to the output and to indicate when information is missing from the input set. We investigated the case of regulated mRNA transcription and unregulated mRNA degradation, when either the temporal series of the regulatory transcription factor (in GRN) or those of the drug-receptor complex (in DGRN) are missing. The first step consists of applying RODES as described in the preceding section. A low performance, i.e., a low R^2 and a high fitness (see eqn (4)), which is not increasing after a sufficiently long time, is an indicator of a missing variable(s). This requires extending RODES such that it can cope with missing information. The tricky solution consists in transforming the modelling problem in a tracking control problem:

1. The measured mRNA temporal series (in the experimental case) or the solution of the complete mRNA ODE (in the simulated case), becomes the desired or reference trajectory.
2. The ODE system with known structure (see equations (1) and (2)) and missing variable(s) becomes the plant to be controlled.
3. The missing variables become the control inputs and are incorporated in the known place of the mRNA ODE|the position of the regulatory factor R of the transcription in equation (1).

The problem is to find the control(s) such that the plant output|the solution of the mRNA ODE|tracks the desired trajectory with an acceptable accuracy, while all the states and the control remain bounded in a physiological range.

It is tempting to speculate that this might be similar to the problem faced by the real living systems during evolution. This idea is corroborated by the fact that regulation appears to evolve on a faster time scale than the coding regions of

the genes. For example, related animals, such as mice and humans, have similar genes, but the transcription regulation of these genes is quite different.

Feedback linearization can be considered one of the most important nonlinear control design strategies developed in the last few decades ([16]). This approach algebraically transforms a nonlinear dynamic system into a linear dynamic system, by using a static-state feedback and a nonlinear coordinate transformation, based on differential geometric analysis of the system. Because our goal is to automate the modelling process, we intended to use a computational intelligence technique. The massive parallelism, natural fault tolerance, and implicit programming of neural networks suggest that they may be good candidates. We successfully applied neural network feedback linearization, based on multilayer perceptrons (MLPs), to complex pharmacogenomic systems to find adequate drug dosage regimens ([4], [5]).

Owing to the reformulation of the modelling problem as a control problem, similar to the aforementioned investigations, an NN FBL approach seems adequate and feasible. We use the NARMA-L2 version of input-output feedback linearization (Narendra cited in ([16]); see also the literature cited in ([16]), in which the output becomes a linear function of a new control input. Fortunately, the prerequisites of the approach, represented by the equations' structure and parameters, are usually known. The control input is the unknown regulator: a pharmacodynamic block containing either the TF concentration in GRN, a drug-related compound, or both in DGRN. A random input, between zero and the estimated maximal value, is injected into the model at random intervals. In this approach a neural network model of the \plant" is first identified. The NN model structure is the standard nonlinear autoregressive moving average (NARMA) model, adapted to the feedback linearization of affine systems|the controller input is not contained in the nonlinearity. We want the system output represented by the mRNA to track a reference trajectory. To determine the model order, we use the lag-space method [17]. The time-series data were obtained by simulation. The number of hidden layers is one for all neural networks. The number of neurons in the hidden layer, of the two MLPs, is 9. The activation functions are tangent hyperbolic for the hidden layer and linear in the output layer for all NNs. The parameters and their values are presented in detail in Table .

Table 2. Neural network feedback linearization parameters.

Parameter	Setting
Network Architecture	
Number of hidden layers	1
Size of hidden layer	9
Sampling tim	0.01
Number of delayed plant inputs	3
Number of delayed plant outputs	2
Training Data	
Training samples	10,000
Minimum plant input	0
Maximum plant output	60
Minimum time interval value	0.1
Maximum time interval value	1
Minimum plant output	0
Maximum plant output	1
Training epochs	100

We investigate the prediction errors by cross-validation on a test set. We used Bayesian regularization [18], a training function that updates the weight and bias values according to Levenberg-Marquardt optimization. It minimizes a combination of squared errors and weights and then determines the correct combination so as to produce a network that generalizes well. We start with different random initial conditions to avoid ending in \bad" local minima. In NN FBL the controller is simply a rearrangement of the neural network plant model. The time-series of the missing variables are identified as the control inputs, and the complete equation is thus reconstructed.

Three important problems can be approached by the proposed methods:

1. finding the unknown transcription factor pro¯le in a drug{gene regulatory network, using the measured mRNA pro¯le as a reference trajectory;
2. finding the unknown drug{receptor complex pro¯le in a drug{gene regulatory network, using again the measured mRNA pro¯le as a reference trajectory; and
3. finding the optimal/individualized drug-receptor complex profile, corresponding to the optimal drug dosage regimen, capable of

constraining the mRNA profile to track a desired therapeutic objective, in a pharmacogenomic context.

To illustrate the algorithm, we selected realistic models of corticoid pharmacogenomics, based on experimental high-dimensional microarray time-series data (see [17] and the literature cited there). Because of the mechanistic and mathematical similarities between transcription regulation by TFs and by drug-receptor complexes, the example is illustrative for both situations. The rationales for this choice are the following:

1. The models are mechanism based, or knowledge driven; i.e., the authors hypothesized the biochemical and pharmacological mechanisms involved and these decided the structure of the ODE models.
2. The models are also data driven; i.e., the authors fitted the parameters of the ODE models to the data.
3. Therefore, even though we use simulated data, these data are quite realistic because we simulated mechanistic models fitting the data very well.

These rationales allow us to test RODES capabilities to automatically identify the structure, estimate the parameters, and determine the corresponding mechanisms, even when information is missing from the input dataset. The authors drastically reduced the dimensionality of the data, by first applying filtering techniques eliminating 7282 probe sets of the 8799 total (82%), and then clustering techniques, resulting in six clusters with similar temporal signatures. After this, they hypothesized the PK/PD mechanisms involved, built the mathematical structure of the six models, and fitted the models' parameters to the data. We investigated and reversed the ODE models of all pharmacological processes, linear and nonlinear, from drug administration to genes stimulation, inhibition, or both. For brevity, we will show only the results for some simple but representative models.

RODES: no missing information

The simple RODES version can be applied only when the temporal series of all variables are available. In practice, this is true for unregulated genes, or for regulated genes when one knows the temporal series of the TF in GRN, or of the drug-receptor complex in DGRN. For the translation equation (see the TF ODE (2)), this implies that the mRNA profile of the TF is known. Although the last two situations are not common, we used them as examples because the equations are more complex and more representative for the present investigation. The induced production of mRNA by drug was described mechanistically as follows:

$$\frac{dmRNA}{dt} = k_{sm} \,.\, (1 + S \,.\, DR_N) - k_{dm} \,.\, mRNA \qquad (5)$$

where the enhancement of transcription rate k_{sm} is dependent on the the drug-receptor complex in the nucleus D_{RN} with a linear efficiency constant of S; k_{dm} is the mRNA degradation rate constant. Equation (5) is a particular case of the more general fundamental ODE pattern of DGRN (see eqn (1)). As we previously stated the equation structure results from the following:

1. summing up pharmacokinetic blocks|a zero-order mRNA production (k_{sm}) and a first-order mRNA degradation ($k_{dm} \cdot mRNA$), and
2. multiplying the rate constants of the regulated processes|the zero-order mRNA production (k_{sm})|by a pharmacodynamic block, a linear stimulation ($1+S \cdot DRN$), with the basal level $E_0 = 1$.

For an example, the authors proposed this mechanism for alpha-2-Macroglobulin (Accession No: X13983mRNAat) [17], and the result of fitting the parameters of the corresponding structure to the data is as follows:

$$\frac{dmRNA}{dt} = 0.038 \,.\, (1 + 0.27 \,.\, DR_N) - 0.038 \,.\, mRNA \qquad (6)$$

A typical equation for a transcription factor TF, resulting from fitting the parameters of an ODE, having the structure of the equation (2), is:

$$\frac{dTF}{dt} = 0.09 \,.\, mRNA_p - 0.09 \,.\, TF \qquad (7)$$

As we stated previously, the preceding equation results by summing up PK blocks, describing here the kinetics of transcription and degradation. The main parameter settings of the GP component of RODES, implemented in Discipulus software by RML Technologies, Inc., are presented in Table 1. The function set, which were selected based on domain knowledge, drastically reduced the structural search space of the algorithm. The reversing result given by RODES for equation (6) was

$$\frac{dmRNA}{dt} = 0.038 + 0.0102 \,.\, DR_N - 0.038 \,.\, mRNA \qquad (8)$$

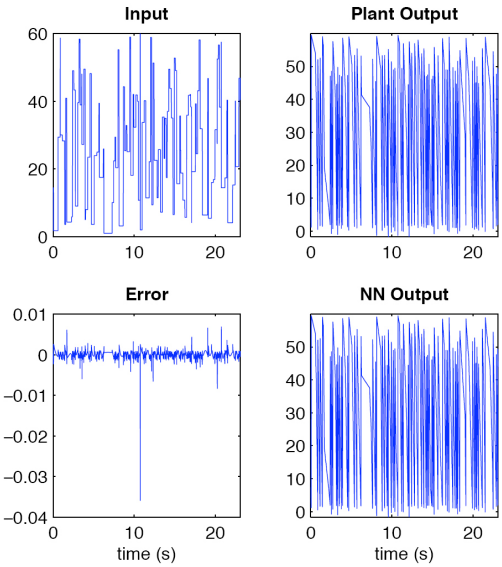


Figure 1. The NN FBL performance (explanations in text).

By factoring the ¯rst term of this equation, we clearly identify the equation (6). It is also easy to identify the mechanism involved: the ¯rst term is a linear stimulation of the zero-order mRNA transcription and the second is a linear degradation of mRNA. The results of applying RODES to equation (7) were

$$\frac{dTF}{dt} = 0.09 \,.\, mRNA - 0.09 \,.\, TF \qquad (9)$$

The equation (7) and the mechanisms involved are easily identified. The average performance measure was $R^2 = 0.99$ (maximum R^2 is 1) for test data, after about 15 minutes, and about 20,000,000 programs evaluation on a PC Athlon 64 £2, 4800+, and 4 GB of RAM.

RODES: Missing Information

Most often the temporal series of the TF or of the drug-receptor complex is not known for regulated genes. This is also true for the temporal series of the mRNA of the TF in proteomics experiments, when one wants to reconstruct the TF ODE (see eqn (2)). In this situations, RODES clearly indicates that information is missing from the input set, and one has to use the extension based on neural network feedback linearization. We used the following equation for illustration:

$$\frac{dmRNA_R}{dt} = k_{sm} \,.\left(1 - \frac{DR_N}{IC_{50} + DR_N}\right) - k_{dm} \,.\, mRNA_R \qquad (10)$$

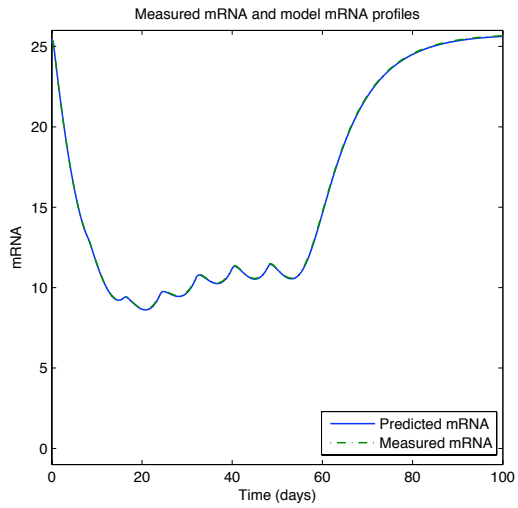


Figure 2. Neural network model identification (explanations in text).

where $mRNA_R$ is the receptor mRNA, $DR(N)$ is the drug-receptor complex in the nucleus, k_{sm} is the rate constant of the receptor mRNA synthesis, k_{dm} is the rate constant of receptor mRNA degradation, and IC_{50} is the concentration of $DR(N)$ at which the synthesis rate of receptor mRNA drops to 50% of its baseline value. A typical equation for a transcription factor TF, resulting from fitting the parameters of an ODE, having the structure of the equation (1), is

$$\frac{dmRNA_R}{dt} = 2.90 \,.\left(1 - \frac{DR_N}{26.2 + DR_N}\right) - 0.1124 \,.\, mRNA_R \qquad (11)$$

To simulate the missing-information case, we disabled the DRN variables from the input set to the algorithm. RODES indicated that information is missing by a law of performance: $R2 = 0.02248$, and the fitness function $= 0.523767$, which is not increasing after a long time, 24 hours and 967,832,980 programs evaluated. Then we applied the neural network feedback linearization presented. The performance of the neural network plant model identification are presented in Fig 1.

The controller is a simple rearrangement of the neural network plant model, and its performance is represented in Fig 2. Because of the very low tracking error, the reference trajectory, which is the \measured" mRNA profile or the mRNA resulting from complete mRNA ODE simulation, and the mRNA output of the controlled system cannot be distinguished. The corresponding controller is related to the missing variable. More

precisely, the regulator R to be identified appeared in equation (1) as $R = DRN=(IC_{50} + DR_N)$. This is because it was clear from the inspection of the `measured" mRNA pro¯le that the TF or drug regulation is nonlinear and inhibitory. So, by applying this method one can identify the PD block containing the variable. Identifying the variable requires further investigations, but is possible (results not shown).

Conclusions

ODE systems are one of the most sophisticated approaches to modelling gene regulatory networks and drug-gene regulatory networks, the superset of GRN we have recently proposed; it is also one of the most difficult. This report extended RODES, our algorithm for reverse-engineering gene networks, on the basis of linear genetic programming, by adding a neural network feedback linearization component. This component enables RODES to deal with missing information. The RODES algorithm automatically discovers the structure of drug-gene regulatory networks and simultaneously estimates its parameters. It takes experimental time-series data or simulated data as input and produces a computer program that models the underlying ODE system. The resulting program can even be deconstructed to identify the biochemical and pharmacological mechanisms involved. The algorithm is applied to simulated data based on a realistic corticoid pharmacogenomic model, and it faithfully reproduces the source ODE system. The execution time of RODES is on the order of a few minutes. This high speed of execution can be attributed mainly to the fact that its fitness evaluation does not require integration. Common domain-specific knowledge can be easily incorporated in the algorithm. At the beginning of the run this is used to select only the mathematical functions expected to be found in the final models. This reduces the structural search space and further speeds up the algorithm. At the end of the run, domain knowledge can be used to identify the biochemical and pharmacological mechanisms from the resultant model. In its basic form, RODES can reveal if some information from the input set is either missing or not related to the output. This is because the basic version requires the temporal series of all variables of the system for inferring an accurate mechanistic model. It also means that it does not discover false input output relations. Here we focused on the case of regulated mRNA transcription and unregulated mRNA degradation, when either the temporal series of the regulatory transcription factor (in GRN) or those of the drug-receptor complex (in DGRN) are missing. Doing so requires extending RODES such that it can cope with missing information. The tricky solution consists of transforming the modeling problem in a tracking control problem. The measured mRNA temporal series becomes the desired, or reference, trajectory. The problem is to find the control(s) such that the plant output|the solution of the mRNA ODE|tracks the desired trajectory with an acceptable level of accuracy. These control inputs are the missing variables that can be identi¯ed in this way, thus completing the automatic reconstruction of the ODE system model. Previous work on RODES has reported similar results in the related field of neural networks. To the best of our knowledge, this is the first time that a reverse-engineering algorithm based on linear genetic programming and neural network feedback linearization has been applied to gene networks. We suggest the algorithm can reverse-engineer ODE systems in any scientific field with a proper use of domain knowledge.

List of abbreviations

DGRN Drug{Gene Regulating Networks
GP Genetic Programming
GRN Gene Regulatory Networks
MLP Multi-Layer Perceptron
NARMA Nonlinear Autoregressive Moving Average
NN FBL Neural Networks Feedback Linearization
ODE Ordinary Di®erential Equations
PD pharmacodynamic
PK pharmacokinetic
RODES Reversing Ordinary Di®erential Equations Systems
TF Transcriptions Factor

References

1. Jong HD: Modeling and Simulation of Genetic Regulatory Systems: A Literature Review. Journal of Computational Biology 2002, 9:67-103.
2. Gardner TS, Faith JJ: Reverse-engineering transcription control networks. Physics of Life Reviews 2005, (2):65-68.

3. Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D: How to infer gene networks from expression profiles. Molecular Systems Biology 2007, 3(78):1-10.
4. Floares AG: Genetic Programming and Neural Networks Feedback Linearization for Modeling and Controlling Complex Pharmacogenomic Systems. In Fuzzy Logic and Applications, 6th International Workshop, WILF 2005, Revised Selected Papers, Volume 3849 of Lecture Notes in Computer Science. Edited by Bloch I, Petrosino A, Tettamanzi A, Crema, Italy: Springer 2005:178-187.
5. Floares AG: Computational Intelligence Tools for Modeling and Controlling Pharmacogenomic Systems: Genetic Programming and Neural Networks. In Proceedings of the 2006 IEEE World Cogress on Computational Intelligence. Edited by Yen GC, Wang L, Bonissone P, Lucas SM, Vancouver, CA: IEEE Press 2006:7510-7517.
6. Floares AG: Automatic Reverse Engineering Algorithm for Drug Gene Regulating Networks. In Proceedings of The 11th IASTED International Conference on Artificial Intelligence and Soft Computing, Palma de Mallorca, Spain 2007.
7. Brameier M, Banzhaf W: Linear Genetic Programming. Genetic and Evolutionary Series, Springer 2007.
8. Floares AG: Reverse Engineering Algorithm for Neural Networks Applied to the Subthalamopallidal Network of the Basal Ganglia. In Proceedings of the International Joint Conference on Neural Networks, Orlando, Florida, USA 2007.
9. Sakamoto E, Iba H: Inferring a System of Differential Equations for a Gene Regulatory Network by using Genetic Programming. In Proceedings of the 2001 Congress on Evolutionary Computation CEC2001, COEX, World Trade Center, 159 Samseong-dong, Gangnam-gu, Seoul, Korea: IEEE Press 2001:720-726, [http://citeseer.ist.psu.edu/sakamoto01inferring.html].
10. Kikuchi S, Tominaga D, Arita M, KTakahashi, Tomita M: Dynamic modeling of genetic networks using genetic algorithm and s-system. Bioinformatics 2003, 19(5):643-650.
11. Noman N, Iba H: Reverse engineering genetic networks using evolutionary computation. Genome Informatics 2005, 16(2):205-214.
12. Cho DY, Cho KH, Zhang BT: Identification of biochemical networks by S-tree based genetic programming. Bioinformatics 2006, 22(13):1631-1640.
13. Savageau MA: Biochemical System Analysis: a Study of Function and Design in Molecular Biology. Reading, MA: Addison-Wesley 1976.
14. Voit EO: Computational Analysis of Biochemical Systems. Cambridge University Press 2000.
15. Beard DA, Qian H, Bassingthwaighte JB: Stoichiometric Foundation of Large-Scale Biochemical System Analysis. In Modelling in Molecular Biology, Springer Natural Computing Series. Edited by Ciobanu G, Rozenberg G, Springer 2004:1-19.
16. Spieth C, Worzischek R, Streichert F: Comparing evolutionary algorithms on the problem of network inference. In GECCO '06: Proceedings of the 8th annual conference on Genetic and evolutionary computation, New York, NY, USA: ACM Press 2006:305-306.
17. Almon RR, Dubois DC, Jin JY, Jusko WJ: Pharmacogenomic Responses of Rat Liver to Methylprednisolone: An Approach to Mining a Rich Microarray Time Series. The AAPS Journal 2005, 7:156-194.
18. MacKay DJC: Bayesian Interpolation. Neural Computation 1992, 4(3):415-447, [http://citeseer.ist.psu.edu/article/mackay91bayesian.html].

about the blues

Vivienne Baillie Gerritsen

Every living being has devised a way to protect its embryos. Humans lodge them in wombs. Fungi protect them in spores. Butterflies keep them in cocoons. Nature's imagination has no limits. In order to keep life going, she has thought up hundreds – if not thousands – of ways of protecting her little ones. Some of her inventions are colourful indeed. Certain species of frog are capable of whipping up bright pink or orange foams in which are embedded their eggs, thereby hidden from predators or sheltered from challenging weather. A certain type of Malaysian tree frog, known as *Polypedates leucomystax* or the Java whipping frog, whisks up foam while it is mating, which gradually turns into a greenish blue on its surface. To what end? No one really knows. But we do know what it is that makes the foam blue: ranasmurfin.

Red Eyed Tree Frog

by Alison Zapata

Courtesy of the artist

Ranasmurfin is a protein. It was discovered in the biofoam that the Java whipping frog creates with a whizz of its hind legs as it is in the process of mating. Biofoams are singular entities. They look like foam that you get on the top of a beer. Or the froth you created in your mouth as a child and let dribble to impress your friends. It's a very comfortable habitat to be living in. Light, airy and soft, it is still strong enough to protect you from challenging conditions, such as hostile weather, microbes or hungry predators. Biofoams are full of different kinds of molecules – amongst which many proteins – which have diverse roles: nutrition, adhesion, strength, protection, hydration etc.

These particular foams are whipped up very close to the water's edge – and are left to hang off a branch above the surface or are stuck to reeds in a pool for example. And when the tadpoles are ready, all they have to do is let themselves drop into the water. Biofoams come in many colours: pink, orange, cream-coloured or colourless. *Polypedates leucomystax* biofoam comes in either one of these colours but has the singularity of gradually turning into a greenish-blue. Such a colour is not commonly found in nature, so it hardly comes as a surprise that the protein which makes the blue colour is not a common protein either. It seems that it is the sunlight or perhaps the atmosphere – or indeed both – which gives this particular tinge to the Java whipping frog's foam. And the protein was named after the Smurfs – the little blue gnome-like people created by the Belgian cartoonist Peyo – who first appeared in comic books in the late 1950s.

How can ranasmurfin become blue? Ranasmurfin is a dimer of two medium-sized monomers which have an arrangement of alpha-helical motifs that has never been described before. What is more, the monomers are linked by way of a chromophore, whose centre is most probably a zinc ion. It is this chromophore which confers on ranasmurfin its blue colour when in contact with the atmosphere. In fact,

ranasmurfin's blue chromophore echos the structural makeup of known blue dyes such as indophenol which is used to colour denim jeans for example. Surprisingly, researchers found that the blue colour in ranasmurfin persists even when the protein itself has been completely denatured!

Why is ranasmurfin blue? Ranasmurfin is found in substantial levels in the Java whipping frog's biofoam, so it must be there for a good reason. It could well be involved in mechanical properties such as conferring stability to the foam or making it more adhesive. But this could hardly account for its blue colour. Perhaps this is a question to which there is no answer. Take a rainbow for instance. Physicists can readily explain why a rainbow shines red, orange, yellow, green, blue, indigo and violet. Yet no one could seriously claim that they know to what end a rainbow shows off its multi-coloured arch. Besides being beautiful, perhaps there is no other purpose. As for this particular biofoam, perhaps blue is a colour which is disagreeable to predators. On the other hand, it may help the biofoam to blend into the environment better thus making it discrete. Whatever the reason may be, it has not been found yet.

The crystallographic study of a protein such as ranasmurfin turned out to be very precious since DNA sequencing on its own could not have predicted the chromophore which forms in the dimer's middle. Indeed, the chromophore results from a modification which occurs once the protein has been synthesized. This simply highlights the necessity to study a protein from all angles possible. What is more, scientists are only beginning to become acquainted with the ins and outs of biofoams. These are turning out to be intricate and specialized worlds of their own where embryos can develop harmoniously within a space designed for light, comfort, air, proper hydration as well as protection against predators and harmful sunrays. Perhaps biofoams will inspire a scientist or two in the creation of environments which could sustain the development of embryos other than tadpoles. Science fiction? Perhaps. In the meantime, let's just admire the palette of colours Nature offers us every day.

Cross-references to Swiss-Prot

Ranasmurfin, *Polypedates leucomystax* (Common tree frog) : P85511

References

1. Oke M., Ching R.T., Carter L.G., Johnson K.A., Liu H., McMahon S.A., White M.F., Bloch C. Jr, Botting C.H., Walsh M.A., Latiff A.A., Kennedy M.W., Cooper A., Naismith J.H.
   Unusual chromophore and cross-links in ranasmurfin: a blue protein from the foam nests of a tropical frog
   Angew. Chem. Int. Ed. Engl.:47:7853-7856(2008)
   PMID: 18781570

2. McMahon S.A., Walsh M.A., Ching R.T., Carter L.G., Dorward M., Johnson K.A., Liu H., Oke M., Bloch C. Jr, Kennedy M.W., Latiff A.A., Cooper A., Taylor G.L., White M.F., Naismith J.H.
   Crystallization of Ranasmurfin, a blue-coloured protein from Polypedates leucomystax
   Acta Crystallogr. Struct. Biol. Cryst. Commun. 62:1124-1126(2006)
   PMID: 17077494

National Nodes

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

Australia
RMC Gunn Building B19, University of Sydney, Sydney

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

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

CASPUR
Rome, Italy

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

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:
May 20, 2009