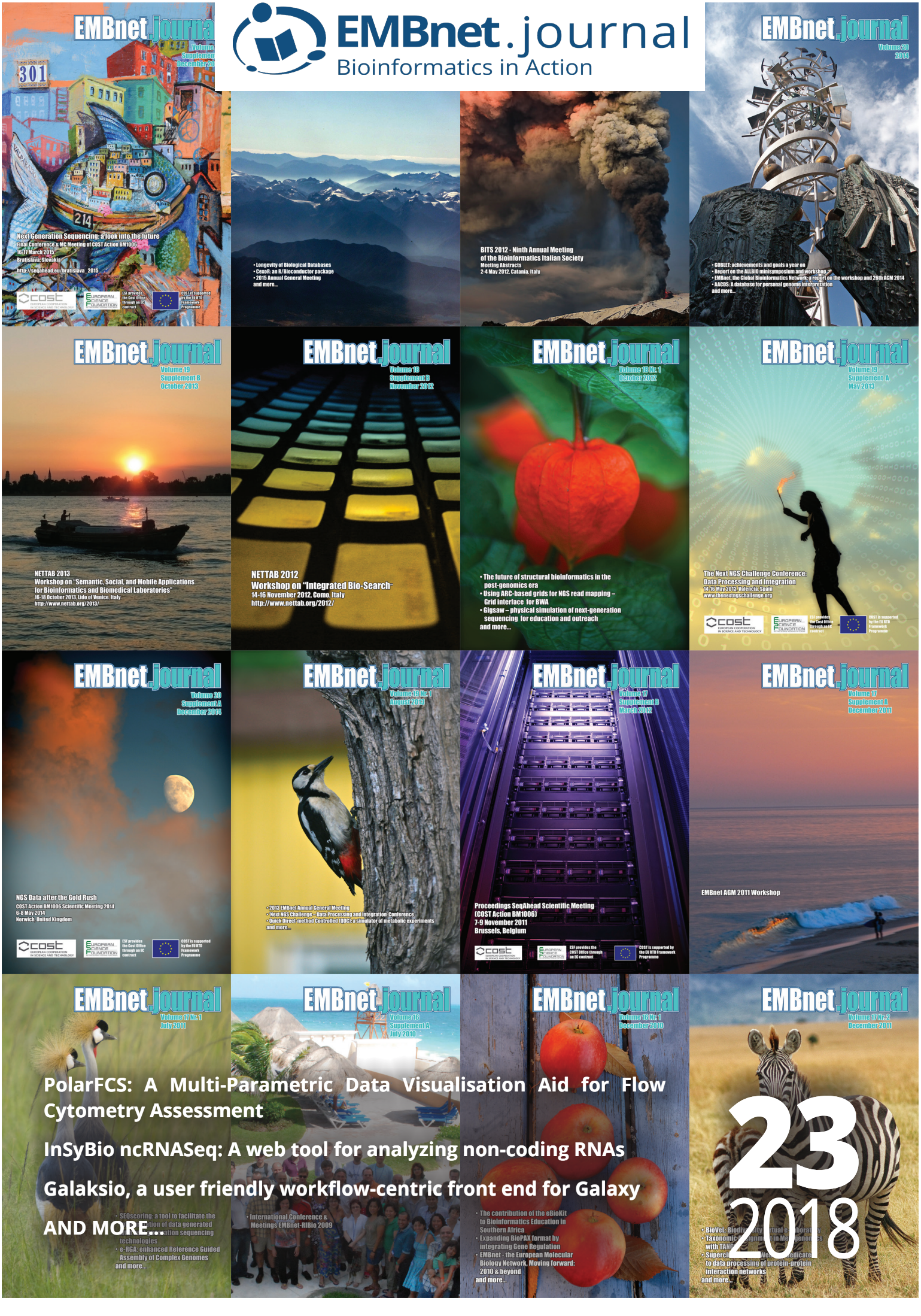

EMBnet.journal
Bioinformatics in Action
EMBnet.journal
Volume 23
Supplement
December 2018
Next Generation Sequencing: a look into the future
Final Conference & MC Meeting of COST Action BM1006
16-17 March 2015
Bratislava, Slovakia
http://seqahead.eu/bratislava_2015
• Longevity of Biological Databases
• CexoR: an R/Bioconductor package
• 2015 Annual General Meeting
and more...
BITS 2012 - Ninth Annual Meeting
of the Bioinformatics Italian Society
Meeting Abstracts
2-4 May 2012, Catania, Italy
Volume 20
2014
• GOBLET: achievements and goals a year on
• Report on the ALLBIO minisymposium and workshop
• EMBnet, the Global Bioinformatics Network: a report on the workshop and 26th AGM 2014
• AACDS: A database for personal genome interpretation
and more...
Volume 19
Supplement B
October 2013
NETTAB 2013
Workshop on "Semantic, Social, and Mobile Applications
for Bioinformatics and Biomedical Laboratories"
16-18 October 2013, Lido of Venice, Italy
http://www.nettab.org/2013/
Volume 18
Supplement B
November 2012
NETTAB 2012
Workshop on "Integrated Bio-Search"
14-16 November 2012, Como, Italy
http://www.nettab.org/2012/
Volume 18 Nr. 1
October 2012
• The future of structural bioinformatics in the
post-genomics era
• Using ARC-based grids for NGS read mapping –
Grid interface for BWA
• Gigsaw – physical simulation of next-generation
sequencing for education and outreach
and more...
Volume 19
Supplement A
May 2013
The Next NGS Challenge Conference:
Data Processing and Integration
14-16 May 2013, Valencia, Spain
www.thenextngschallenge.org
Volume 20
Supplement A
December 2014
NGS Data after the Gold Rush
COST Action BM1006 Scientific Meeting 2014
6-8 May 2014
Norwich, United Kingdom
Volume 19 Nr. 1
August 2013
• 2013 EMBnet Annual General Meeting
• Next NGS Challenge – Data Processing and Integration Conference
• Quick Direct-method Controlled (QDC): a simulator of metabolic experiments
and more...
Volume 17
Supplement B
March 2012
Proceedings SeqAhead Scientific Meeting
(COST Action BM1006)
7-9 November 2011
Brussels, Belgium
Volume 17
Supplement A
December 2011
EMBnet AGM 2011 Workshop
Volume 17 Nr. 1
July 2011
Volume 16
Supplement A
July 2010
• International Conference &
Meetings EMBnet-RIBio 2009
Volume 16 Nr. 1
December 2010
• The contribution of the eBioKit
to Bioinformatics Education in
Southern Africa
• Expanding BioPAX format by
integrating Gene Regulation
• EMBnet - the European Molecular
Biology Network, Moving forward:
2010 & beyond
and more...
Volume 17 Nr. 2
December 2011
PolarFCS: A Multi-Parametric Data Visualisation Aid for Flow Cytometry Assessment
InSyBio ncRNASeq: A web tool for analyzing non-coding RNAs
Galaksio, a user friendly workflow-centric front end for Galaxy
AND MORE...
23
2018

EMBnet.journal
Bioinformatics in Action

Contents

Editorial

The year 2017 was full of new revolutionary biotechnology advances, not only a massification of new High Throughput Sequencing technologies as the ones created by Oxford Nanopore but also an increase in the usage of these technologies by researchers working on all areas related to Life Sciences, from biologists to veterinaries, farmers and medical doctors.

The low price of sequencing has also brought bioinformatics (or at least the need of) to households. Today any person interested in obtaining an accurate family genealogy tree or having some basic knowledge about genetic predisposition to develop a specific disease can do that for around 100€. There are many companies today offering single nucleotide polymorphism (SNP) chip screening, exome sequencing and even whole genome sequencing (WGS) to the public. As an example, a company offers WGS for less than the magic 1000U$ level, a science fiction price a few years ago.

Today companies are not only providing genome sequencing services for human patients but also for pets like dogs, cats and horses. In the coming years, we are going to see business ideas using bioinformatics in many new fields; with only imagination as a limit.

This explosion in widespread usage of biotechnologies is creating a much more prominent need in education, in new easy of use analysis tools and also of an increasing number of people with expertise in the field of bioinformatics.

This issue of EMBnet.journal has articles about examples of new initiatives, analysis tools and network activities that reflect the scenario described above. The EMBnet network will during 2018 celebrate 30 years since its foundation. Therefore, the EMBnet.journal is working hard to create a more prosperous 2018 issue and full of interesting articles.

Erik Bongcam-Rudloff
Editor-in-Chief
erik.bongcam@slu.se

http://dx.doi.org/10.14806/ej.23.0.913

EMBnet.journal
Bioinformatics in Action

Palopoli et al. (2017) EMBnet.journal 23, e883
http://dx.doi.org/10.14806/ej.23.0.883

Reports

A report on the "International Society for Computational Biology - Latin America (ISCB-LA)" Bioinformatics Conference 2016

Nicolas Palopoli[1,2], Alexander Miguel Monzon[1], Gustavo Parisi[1], Ariel Chernomoretz[3], Fernán Agüero[4]

[1] Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, CONICET, Buenos Aires, Argentina; [2] Fundación Instituto Leloir-IIBBA-CONICET, Buenos Aires, Argentina; [3] Integrative Systems Biology Laboratory, Fundacion Instituto Leloir & IFIBA (FCEN-UBA / CONICET), Buenos Aires, Argentina; [4]Instituto de Investigaciones Biotecnológicas – Instituto Tecnológico de Chascomús (IIB-INTECH), Universidad Nacional de San Martín (UNSAM) – CONICET, Buenos Aires, Argentina.

Abstract

The fourth edition of the "International Society for Computational Biology - Latin America (ISCB-LA)" Bioinformatics Conference, jointly organized with the Argentine Association for Bioinformatics and Computational Biology (A^2B^2C), was held in Buenos Aires, Argentina, on November 2016. The ISCB-LA 2016 conference aimed at showcasing the latest findings in the field from researchers at all career levels in the region and beyond, and promoting collaborations to further develop Bioinformatics and Computational Biology in Latin America. Here we present a brief report on the main activities held during the Conference and their successful outcomes.

Introduction

The 4th edition of the Bioinformatics Conference "International Society for Computational Biology - Latin America (ISCB-LA)[1]" took place on 21-23 November 2016 at Universidad Católica Argentina (UCA) in Buenos Aires, Argentina. This biennial meeting gathers students and researchers from across Latin America, plus participants and renowned invited speakers from the rest of the world, to present and discuss their work on the development and application of computational methods to advance biological knowledge. More than 300 delegates had the opportunity to network with and learn from colleagues in keynote lectures, selected oral presentations, poster sessions and sponsored tech talks, plus the associated workshop and student symposium.

Figure 1. ISCB-LA 2016 conference group photography. Photo by Franco L. Simonetti.

In the last decade, the International Society for Computational Biology[2] has spread its reach to the scientific communities outside Europe and USA. Latin America has become an active asset for ISCB, as demonstrated by the ISCB-LA meetings and the number of Regional Student Groups (RSG) of the ISCB Student Council[3] that have flourished lately. Previously joining as conference sponsor, this was the first ISCB-LA edition co-organized by the Argentine Association for Bioinformatics and Computational Biology[4]. It also served as the 7th annual edition of the Argentine Conference on Bioinformatics and Computational Biology. A^2B^2C is a non-profit organization established in 2009 that is successfully promoting the growth of Bioinformatics in Argentina and guiding the career development of students and young researchers in the country. In particular, the emergence of A^2B^2C has been crucial to create and strengthen bonds between groups working far from each other in a vast country, as well as fostering connections among scientists in Argentina and neighboring countries such as Chile, Uruguay and Brazil.

Main conference

Highlights of ISCB-LA 2016 were the five keynote lectures presented by invited speakers. Dr. Ruth Nussinov opened the Conference discussing the structural properties of Ras proteins related with cancer development. The day finished with Dr. Emilio Kropff who explained how the internal GPS in a mammalian brain controls the speed

[1] http://www.iscb.org/iscb-latinamerica2016
[2]https://www.iscb.org
[3]http://rsg.iscbsc.org
[4]http://www.a2b2c.org

Article history
Received: 21 February 2017
Published: 10 April 2017

and displacement of movement, highlighting the special role of the newly discovered speed cells. The second day had Dr. Ana Amador speaking about the use of computational models to predict neural activity of birds based on their singing patterns. It was closed by the EMBO Lecture keynote presentation delivered by Dr. Søren Brunak in which he presented a 20-year-long study on 6 million of Danish patients to uncover correlations among drugs, disease and genetic information. The last keynote, by Dr. Seán I. O'Donoghue, addressed the principles and modern methods available for data visualization, which serve as useful research tools for exploring the huge volume and complexity of current biological data.

Accepted submissions for oral talks were organized in six sessions around a common topic, named "Proteins" (eight talks), "Data" (seven), "Machine Learning" (seven), "Disease" (seven), "Genes" (six) and "Systems" (seven). These comprised presentations by local and international group leaders, early career scientists and PhD students, selected by senior researchers in the Scientific Committee under the guidance of its Chair, Dr. Morten Nielsen. Also, 221 posters were accepted and split in two afternoon sessions. The presenting authors of the top three posters as chosen by the votes of participants received a cash prize.

Besides the scientific program, there were a couple of social events in the evenings to maximize opportunities of networking. To close the first conference day Dr. Seán O'Donoghue took the stage of the cocktail party to premiere new animations from the VIZBiPlus project[5]. On the last day, the Farewell Party got people together for dinner, drinks and dancing.

Figure 2. "Come to the dry side", an invitation from the RSG-Argentina, printed on merchandise available at ISCB-LA 2016. Photo of the mug by Bart Cuypers.

Satellite activities

Satellite activities took place on 18-19 November 2016 at Universidad Nacional de San Martín, in Buenos Aires, Argentina. The first day offered the workshop "Data Visualisation Methods and Tools - A Practical Guide"[6]. Dr. Seán O'Donoghue covered general approaches and computational tools for Data Viz during the morning, followed by a hands-on introduction to the ggplot2 R library by Alan Bush and Gonzalo Corti Bielsa. On the following day, the second Latin America Student Council Symposium[7] was attended by 62 delegates who enjoyed 13 student talks and 35 poster presentations. Keynote lectures by Prof. Ruth Nussinov and Dr. Seán O'Donoghue presented their research topics with a special focus on career development advice for young students.

Conclusions

ISCB-LA 2016 continued the positive trend of support and adhesion it has enjoyed since its inception in 2010[8], highlighting the expanding relevance of Bioinformatics, Computational Biology and related fields in the region. More information about ISCB-LA 2016, including the final schedule and program, is available at: http://www.iscb.org/iscb-latinamerica2016.

Acknowledgements

We would like to recognize the dedication and support of other members of the local Organising Committee; the ISCB Steering Committee; the Chair and members of the Scientific Committee; the Boards of Directors from A²B²C, ISCB, ISCB-SC and RSG-Argentina; support staff at UCA; and students who acted as volunteers during the conference. We also want to acknowledge the great support we had from funding agencies and conference sponsors.

[5] http://vizbi.org/Plus/
[6] http://lascs2016.iscbsc.org/lascs2016-dataviz-workshop
[7] http://lascs.iscbsc.org
[8] http://www.iscb.org/archive/conferences/iscb/iscb-la2010.html

EMBnet.journal
Bioinformatics in Action

Romano (2017) EMBnet.journal 23, e884
http://dx.doi.org/10.14806/ej.23.0.884

Reports

Bioinformatics activities at the University Hospital San Martino IST in Genoa

Paolo Romano [1]

[1] IRCCS AOU San Martino IST, Genoa, Italy.

Abstract

This report summarises some of the bioinformatics activities that have been carried out since 1986 at the National Cancer Research Institute of Genoa, now University Hospital IRCCS San Martino IST. Two main interrelated research lines are highlighted: data management for biological resources, and automation of data retrieval and analysis. As developments in Information and Communication Technologies (ICTs) are fundamental to bioinformatics, the NETTAB workshops, which are devoted to the analysis of the impact of new ICTs on bioinformatics research, are also presented.

Introduction

Bioinformatics is still a relatively young discipline. In the '80s, biology and ICT were completely different from today. When the EMBL Nucleotide Sequence Data Library was first established, in 1980, the challenges mainly related to establishing the first databases and developing sequence comparison algorithms.

Many algorithms and software tools, which are now the basis for every molecular data analysis, have since been developed; meanwhile, data availability has been increasing at an impressive speed, thanks to the advent of high-throughput technologies and '*omics*' projects. At the same time, we have moved from local elaboration of data to remote data analysis.

These transformations have required the development and implementation of new tools for remote processing and data sharing. Hence, the focus nowadays has shifted to the integration and analysis of an unprecedented amount of information, aiming to build an interoperable, semantically aware, social and collaboratively-based network environment for bioinformatics.

In this short report, I summarise the main activities of the National Cancer Research Institute of Genoa, now IRCCS AOU San Martino IST, since the '80s, following, and sometimes anticipating, the evolution of bioinformatics tools and databases.

Biological resource data management

The term 'biological resource' is applied to living biological material collected and held in culture collections: bacterial and fungal cultures; animal, human and plant cells; viruses or isolated genetic material. A wealth of information about biological resources has been accumulated in Biological Resource Centres (BRCs) and, although dispersed, a large part of this information is still accessible. Various coordinated efforts have been put into making this information jointly available on-line. Many more improvements can be achieved by adopting innovative ICTs to deepen integration of this information in the bioinformatics network environment.

Automation of data retrieval and analysis

In biology, data integration is limited by the great number of available resources, their size and frequency of updates, their heterogeneity and distribution on different servers. Integration of these data can therefore be achieved only by adopting flexible and extensible tools. XML, Web Services (WSs) and Workflow Management Systems (WMSs) can support the creation and deployment of software able to automate data retrieval and analysis.
A WMS is able to design and create workflows, and to manage their execution. Its main components are i) a graphical interface for composing workflows, entering data and displaying different types of results; ii) an archive to store workflow descriptions, as well as results of executions and related traces; iii) a scheduler able to invoke services when needed; iv) a registry of available services; v) Application Programming Interfaces (APIs) for interoperating with services; and vi) a monitor tool to control workflow execution.

For this to happen, a 'technology-savvy' status must be achieved by providers and users. In this status, databases adhere to standards, and include semantic metadata; software is distributed on the network and can interoperate; and data-analysis procedures can be carried out on the network. A shared methodology for software development should also be adopted by developers and service providers. This could include

Article history
Received: 23 February 2017
Published: 10 April 2017

Romano (2017) EMBnet.journal 23, e884
http://dx.doi.org/10.14806/ej.23.0.884

i) XML schemas for creating shared data models; ii) XML-based languages for data storage and exchange; iii) WSs for software interoperability; iv) ontologies for WS discovery, selection and interoperation; and v) workflows for executing analysis processes.

The Interlab project

The Interlab Project was funded in 1989. One of its goals was to implement data-banks of biological resources. The Cell Line Data Base (CLDB), which collected and made available information on cell lines, the B Line Data Base (BLDB), which stored information relating to HLA typed B lymphoblastoid cell lines, and the Molecular Probe Data Base (MPDB), containing data on oligonucleotides, were built in that context. The databases were first made available on-line through packet-switching data networks (ITAPAC in Italy). Researchers could connect by means of personal computers equipped with standard modems. Dec VT100 terminal emulation was required in order to exploit the interface. Later, new interfaces were built using the Wide Area Information Servers (WAIS) technology, the Gopher system and, with the advent of the World-Wide Web concept, through Web servers. A new hypertext interface was developed for the CLDB – HyperCLDB – to allow effective indexing of its contents by search engines like Google and Yahoo (Romano et al., 2009). HyperCLDB[1] consists of many pages (currently, ~8,750), including detailed descriptions of cell lines and indexes of their features. In each page, hyperlinks are added to connected pages and facilitate navigation.

Common Access to Biological Resources and Information (CABRI)

The Common Access to Biological Resources and Information (CABRI) project was funded by the European Union from 1996 to 1999 (Romano *et al.*, 2005). Among its objectives, it aimed to ease access to information in biological resource catalogues. CABRI[2] is based on the Sequence Retrieval Software (SRS), a search engine designed for integrated queries of molecular biology databases. With SRS, data must reside locally and be stored in 'flat files' (text-only files) with pre-defined, shared syntaxes. Both explicit and implicit links between databases can be defined. At the time, SRS was a good option for making integrated searches of databases with similar contents and interlinks in local environments. CABRI catalogues were implemented in SRS by comparing the data structures of collections' databases, and then defining three shared data-sets for each material: the Minimum Data Set (MDS) includes information needed to uniquely identify a resource; the Recommended Data Set (RDS) includes information useful to achieve an improved description of the characteristics, functions and properties of a resource; and the Full Data Set (FDS) includes all available information related to a resource. Data-input procedures were defined for each item of the MDS and RDS: they provide a textual description of its contents and specify the input process for the corresponding values. CABRI currently includes 28 collections that can be searched either via a simplified interface or the standard SRS interface.

Microbial Resource Research Infrastructure (MIRRI)

The European Microbial Resource Research Infrastructure (MIRRI) project can be considered an evolution of CABRI. One of its main objectives is to provide access to information available in the European collections of microorganisms through a dynamic Information System. The MIRRI-IS should include a repository for BRC catalogues, a tight interconnection with domain information systems, a unique portal for catalogues and associated data, and an interoperable system based on APIs and workflows. Three demonstration systems were developed in the MIRRI preparatory phase: the BacDive demonstrator aims to extend the contents of catalogues with a greater number of well-defined data; the StrainInfo demonstrator is targeted towards a better integration among collection catalogues through the identification of common strains; and the USMI Galaxy demonstrator aims both to support data curation and to integrate catalogues with external resources. A five-year plan for the implementation of the MIRRI-IS has been defined (Romano *et al.*, 2017).

IST Bioinformatics Web Services (IBWS)

A suite of WSs – the IST Bioinformatics Web Services (IBWS) (Zappa *et al.*, 2010) – was developed to make databases available at the IRCCS San Martino IST accessible through standard APIs. IBWS has been developed by using standard tools, and should be easy to invoke by any compliant software, such as Taverna. The main advantage offered by IBWS relates to the possibility of accessing a set of unique archives, which otherwise could only be queried manually, through standard APIs.

Bioinformatics Web Enactment Portal (BioWEP)

The use of WMSs can be difficult for the majority of biologists. Web portals can allow users to enact useful workflows in a friendly environment. The Bioinformatics Web Enactment Portal (Biowep) is a Web application that allows selection and execution of pre-defined, annotated workflows (Romano *et al.*, 2007). It is based on a server-side implementation of the Taverna

[1] http://bioinformatics.hsanmartino.it/hypercldb/
[2] http://www.cabri.org/

enactor. Workflow annotation, which is achieved via an ontology of bioinformatics tasks and data-types, involves registration of the data-types for the main components of the workflow. Users can then select workflows of interest on the basis of their annotation.

NETTAB Workshops series

Continuous monitoring of technological developments, and of their impact on biological research, is needed in order to promote swift adoption of the most promising and innovative bioinformatics tools. This is the objective of the NETTAB Workshops.

NETTAB[3] Workshops are a series of International meetings on "Network Tools and Applications in Biology", held annually in Italy. They aim to introduce participants to the most innovative ICTs, and provide a unique forum for bringing together biologists and bioinformaticians with computer science experts. Workshops include sessions devoted to tools, systems, platforms and early applications of relevant technologies. Keynote lectures and selected presentations are included in the programme, alongside poster sessions and tutorials.

Because of the continuous technological evolution, the workshops focus each year on a different technology or domain. Since 2001, many themes have been discussed, including standardisation for data integration (Genoa, 2001), multi-agent systems (Bologna, 2002), scientific workflows (Naples, 2005), Web Services (Santa Margherita di Pula, 2006), Semantic Web (Pisa, 2007), collaborative research (Catania, 2009), wikis (Naples, 2010), social and mobile applications (Venice, 2013), and reproducibility (Rome, 2016).

[3] http://www.nettab.org/

References

1. Romano P, Kracht M, Manniello MA, Stegehuis G, Fritze D. (2005) The role of informatics in the coordinated management of biological resources collections, *Applied Bioinformatics*. 2005, **4**(3):175-86. http://dx.doi.org/10.2165/00822942-200594030-00002
2. Romano P, Bartocci E, Bertolini G, De Paoli F, Marra D et al. (2007) Biowep: a workflow enactment portal for bioinformatics applications. *BMC Bioinformatics* 2007, **8**(Suppl 1):S19. http://dx.doi.org/10.1186/1471-2105-8-S1-S19
3. Romano P, Manniello A, Aresu O, Armento M, Cesaro M et al. (2009) Cell Line Data Base: structure and recent improvements towards molecular authentication of human cell lines. *Nucleic Acids Research*, **37**(Database issue):D925-D932. http://dx.doi.org/10.1093/nar/gkn730
4. Romano P, Smith D, Bunk B, Vasilenko A, Glöckner FO. Designing the MIRRI information system. *PeerJ Preprints*. Submitted on February 17, 2017. https://peerj.com/preprints/2815/
5. Zappa A, Miele M, Romano P. (2010) IBWS: IST Bioinformatics Web Services. *Nucleic Acids Research*, **38**(Web Server issue):W712-W718. http://dx.doi.org/10.1093/nar/gkq416

EMBnet.journal
Bioinformatics in Action

Manzoor et al. (2017) EMBnet.journal 23, e891
http://dx.doi.org/10.14806/ej.23.0.891

Reports

A stepping-stone to developing bioinformatics in Pakistan

Shahid Manzoor[1], Adnan Niazi[2], Erik Bongcam-Rudloff[3]

[1]Department of Information Technology, University of the Punjab, Lahore, Pakistan; [2]Department of Immunology, Genetics and Pathology, Science for Life Laboratory Uppsala, Uppsala University, Uppsala, Sweden; [3]SLU Global Bioinformatics Centre, Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden.

Introduction

Bioinformatics employs a wide range of 'informatics' techniques to analyse and extract information associated with large-scale biological data. *In silico* tools and methods have been pivotal for DNA and protein analysis, including those for sequence translation and alignment, gene finding, gene annotation, protein structure prediction and phylogenetic reconstruction. Advances in computational and molecular biology research, and application of high-throughput next-generation sequencing technologies in areas such as genomics, transcriptomics and proteomics, has taken bioinformatics to an even higher level (Cole *et al.*, 2014); (Wang and Zhang, 2013). Academic groups, research consortia and industries worldwide are intensively using bioinformatics as a tool to address life science related questions. This field of study is relatively new in Pakistan: it was introduced in 2002, when Muhammad Ali Jinnah University, now Capital University of Science and Technology (CUST), took the initiative and started an undergraduate degree programme in bioinformatics for the first time. Later, a similar programme was introduced by the COMSATS Institute of Information Technology (CIIT), another renowned university in Pakistan and member of EMBnet since 2006. HEC and the aforementioned universities played a vital role in developing and promoting awareness of bioinformatics, and the importance of education and training in this field, among scientific communities in Pakistan. Consequently, several universities launched similar programmes at both graduate and postgraduate levels.

Importance of bioinformatics in Pakistan

Bioinformatics cannot be overlooked in a country like Pakistan because of its unique genetic resources, in terms of human population and biodiversity of crops and animal species. Pakistan produces a large variety of agricultural products, such as cotton, wheat, rice, sugarcane, fruit and vegetables, in addition to cattle and poultry. Furthermore, its geographical features, and the presence of various ethnicities with familial and social characteristics in a population of over 200 million, is a valuable resource for the study of genetic disorders, such as Down syndrome, Fragile X syndrome, intellectual disability, psoriasis, schizophrenia, deafness, Alzheimer's disease, albinism and epilepsy.

Bioinformatics is widely practiced within the pharmaceutical industry in the development of health-care products, and also in agriculture and environmental protection (Lyall, 1996); (Xue and Zhao, 2008). The growth of the pharmaceutical industry demands advanced tools and methods for the discovery of drug targets (Fagan and Swindells, 2000), for drug design (Kelly and Clark, 2003) and for the identification of new disease markers to improve early diagnosis and develop new therapeutic strategies. Pakistan is aiming to raise its research standards in agricultural, biotechnological and biomedical sectors by exploiting bioinformatics approaches in many life-science domains.

A step forward to improve bioinformatics in Pakistan

To meet the demand for bioinformatics-oriented research and education in Pakistan, a key step was taken by the HEC through the approval of an "*Overseas scholarship for MS/MPhil leading to PhD in bioinformatics*" programme in 2006. The HEC offered 50 scholarships in bioinformatics (Ilyas and Sadique, 2011), aiming to create a critical mass of highly qualified researchers. The inclusion of this specialised human resource in bioinformatics in major research organisations and institutes aimed to boost research activities of Pakistan, and to allow the development of new projects with great economic returns. To this end, a collaboration was established with our Swedish colleagues during the EMBnet Annual General Meeting in Torremolinos (ES) (Rahman and Chohan, 2010). Scholarships were assigned to talented graduate students in selected fields of science. Applicants were shortlisted and called for interview after a graduate-assessment test. In 2008, an expert panel (comprising Erik Bongcam-Rudloff, Shahid Nadeem Chohan, Raheel Qamar and a representative from the HEC) nominated

Article history
Received: 15 March 2017
Published: 11 May 2017

Manzoor et al. (2017) EMBnet.journal 23, e891
http://dx.doi.org/10.14806/ej.23.0.891

10 candidates for higher studies at SLU. This was the first time (2009) that SLU had set up a bioinformatics MSc course with students selected in this way. All students started PhD programmes after completing their MSc studies. These scholars were awarded PhD degrees after successful public defence of their doctoral theses; the research topics of these theses are listed in Table 1.

Table 1. List of PhD theses under the programme.

Title of the project	Main Supervisor	Co Supervisor(s)
Bioinformatics screening for candidate mutations underlying phenotypic traits in domestic animals. ISBN 978-91-576-8132-4	Prof. Göran Andersson	Prof. Leif Andersson, Prof. Kerstin Lindblad-Toh, Prof. Erik Bongcam-Rudloff
Bioinformatics analysis of bacterial pathogens from East African camels. ISBN 978-91-576-8242-0	Prof. Erik Bongcam-Rudloff	Dr. Etienne de Villiers & Dr. Richard Bishop (Kenya)
Computational and comparative investigations of syntrophic acetate-oxidising bacteria (SAOB). ISBN 978-91-576-8060-0	Prof. Erik Bongcam-Rudloff	Prof. Anna Schnürer, Dr. Bettina Müller
Bioinformatics studies on the mechanisms of gene regulation in vertebrates. ISBN 978-91-576-8112-6	Prof. Göran Andersson	Prof. Leif Andersson
In silico analysis of Treponema and Brachyspira genomes. ISBN 978-91-576-8240-6	Prof. Erik Bongcam-Rudloff	Dr. Anna Rosander, Dr. Desiree Jansson
Mapping and functional characterisation of candidate genes and mutations for chicken growth. ISBN 978-91-576-8046-4	Prof. Örjan Carlborg	Dr. Stefan Marklund, Dr. Anna Johansson
Genome-wide analyses of *Bacillus amyloliquefaciens* strains provide insights into their beneficial role on plants. ISBN 978-91-576-8080-8	Prof. Erik Bongcam-Rudloff	Prof. Johan Meijer, Dr. Sarosh Bejai
Bioinformatics analysis of whole genome sequencing data. ISBN 978-91-576-8064-8	Prof. Leif Andersson	Prof. Lars Rönnegård, Prof. Erik Bongcam-Rudloff
Towards High-Throughput Phenotypic and Systemic Profiling of *in vitro* Growing Cell Populations using Label-Free Microscopy and Spectroscopy: Applications in Cancer Pharmacology. ISBN: 978-91-554-9082-9	Prof. Mats Gustafsson	Dr. Mårten Fryknäs, Dr. Ulf Hammerling
Integrated Computational and Experimental Approaches for Accelerated Drug Combination Discovery and Development: Applications in Cancer Pharmacology. ISBN: 978-91-554-9177-2	Prof. Mats Gustafsson	Prof. Rolf Larsson, Dr. Claes Andersson

The scholars also published more than 40 research articles in peer-reviewed journals during their PhD programmes (Figure 1).

The success of this education scheme had a wide resonance in the country, and several institutes subsequently launched degree programmes in bioinformatics. Currently, around 20 universities and research institutes are offering bioinformatics programmes at undergraduate and postgraduate levels (see Table 2).

These educational programmes will provide Pakistan with specialised human resources able to improve its research capability and its competitiveness at an international level.

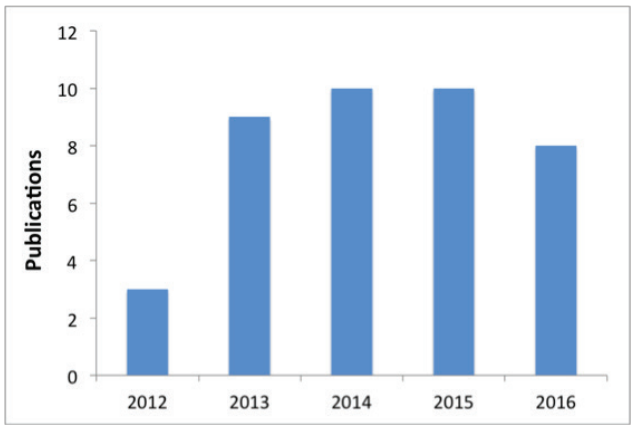


Figure 1. Number of articles published by scholars of the programme..

Table 2. Institutes offering education programmes in bioinformatics in Pakistan.

Sr.#	Institute	City	Programme
1	Baqai Medical University	Karachi	BS
2	Capital University of Science and Technology	Islamabad	BS/MS
3	Comsats Institute of Information Technology	Islamabad	BS/MS
4	Comsats Institute of Information Technology	Sahiwal	BS
5	Federal Institute of Health Sciences	Lahore	BS
6	Forman Christian College	Lahore	BS
7	Government College University	Faisalabad	BS/MS
8	Government Postgraduate College Mandian	Abbottabad	BS
9	Hazara University	Mansehra	BS/MS
10	International Islamic University	Islamabad	BS/MS
11	Khushal Khan Khattak University	Karak	BS
12	National University of Sciences and Technology	Islamabad	MS
13	Qarshi University	Lahore	BS
14	Quaid-e-azam University	Islamabad	BS/MPhil/PhD
15	Shaheed Benazir Bhutto Women University	Peshawar	BS/MPhil
16	Sir Syed University of Engineering & Technology	Karachi	BS
17	The Superior University	Lahore	BS
18	University of Agriculture	Faisalabad	BS
19	University of Sindh	Jamshoro	MPhil
20	Virtual University of Pakistan	Lahore	BS/MS

Conclusions

This joint venture between Pakistan's HEC and SLU (SE) was a stepping-stone to expand the scope of bioinformatics in educational and research centres in Pakistan. In addition, it provided important international networking opportunities within the field. Similar efforts in the future will allow Pakistan to augment the pool of skilled bioinformaticians to meet its challenging needs in different research fields. The Masters and PhD programmes developed via this initiative are also being used as a model for bioinformatics education in several countries in Asia and Africa.

Acknowledgments

This work was supported by the Higher Education Commission in Pakistan, by the University of the Punjab, Lahore, Pakistan, and by the Swedish University of Agricultural Sciences (SLU) in Sweden. We are grateful to the leadership at the Faculty of Veterinary Medicine and Animal Science, SLU (SE) for their great support. We also wish to express our gratitude to Nils-Einar Eriksson (UU), for his support and advice. The successful completion of all PhD theses would not have been possible without the commitment of all the main- and co-supervisors at SLU and Uppsala University.

References

1. Cole C, Krampis K, Karagiannis K, Almeida JS, Faison WJ *et al.* (2014) Non-synonymous variations in cancer and their effects on the human proteome: workflow for NGS data biocuration and proteome-wide analysis of TCGA data. *BMC Bioinformatics* **15**, 28. http://dx.doi.org/10.1186/1471-2105-15-28
2. Fagan R, Swindells M (2000) Bioinformatics, target discovery and the pharmaceutical/biotechnology industry. *Curr Opin Mol Ther* **2**, 655–661
3. Ilyas M, Sadique S, Masood K, Qamar R, Chohan SN (2011) The development of computational biology in Pakistan: still a long way to go. *PLoS Comput Biol* **7**, e1001135. http://dx.doi.org/10.1371/journal.pcbi.1001135
4. Kelly DE, Clark A (2003) Modern approaches to drug discovery and design: setting the scene. *Biochem Soc Trans* **31**, 428. http://dx.doi.org/10.1042/bst0310428
5. Lyall A (1996) Bioinformatics in the pharmaceutical industry. *Trends in Biotechnology* **14**, 308–312
6. Rahman N, Chohan SN (2010) Pakistan EMBnet node: progress report. *EMBnet.news* **15**(4), 35–36. http://journal.embnet.org/index.php/embnetnews/article/view/48
7. Wang X, Zhang B (2013) customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* **29**, 3235–3237. https://dx.doi.org/10.1093/bioinformatics/btt543
8. Xue J, Zhao S, Liang Y, Hou C, Wang J (2008) Bioinformatics and its Applications in Agriculture. In Li D (ed.), *Computer And Computing Technologies In Agriculture* **2**, 977–982. http://dx.doi.org/10.1007/978-0-387-77253-0_29

EMBnet.journal
Bioinformatics in Action

Aron et al. (2017) EMBnet.journal 23, e886
http://dx.doi.org/10.14806/ej.23.0.886

Reports

H3ABioNet: Developing Sustainable Bioinformatics Capacity in Africa

Shaun Aron[1,2], Kim Gurwitz[1,3], Sumir Panji[1,3], Nicola Mulder[1,3], H3ABioNet Education and Training working group* as member of the H3Africa Consortium

[1]H3ABioNet - a pan-African Bioinformatics Network; [2]Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, South Africa; [3]Computational Biology Division, Department of Integrative Biomedical Sciences, University of Cape Town, South Africa.

* all H3ABioNet Education and Training working group members are listed on the last page of this article.

Introduction to H3ABioNet

H3ABioNet[1] - a pan-African bioinformatics network (Mulder *et al.*, 2016) - was established in 2013 as part of the Human Heredity and Health in Africa[2] (H3Africa) initiative (H3Africa Consortium *et al.*, 2014). H3ABioNet comprises 32 research institutions across 14 African countries with 1 partner institution in the USA and 1 partner institution in the UK. In addition to infrastructure building endeavours, H3ABioNet has a significant training component, with the aim of training bioinformaticians as well as biologists, geneticists, and clinicians in the use of bioinformatics to facilitate data analysis for their H3Africa project. The initial need for bioinformatics training was focused on addressing the needs of both the H3Africa community as well as training requests from within the H3ABioNet network. Between May 2013 and October 2016, H3ABioNet has hosted or supported 49 training events across various African countries. These events range from extensive four-week bioinformatics postgraduate training courses, to specialised training workshops on various topics, such as: data management, introductory and advanced system administration, biostatistics, genome-wide association studies (GWAS), next generation sequencing data analysis, and metagenomics. Training events have also taken the form of internships, online training, and hackathons. In total, approximately 1036 individuals have received training from H3ABioNet training events over the above mentioned period. A few select training events are highlighted below. We have learnt several important lesson about planning successful training events and have attempted to adapt our approaches to address various challenges in developing bioinformatics capacity in Africa, which is discussed in more detail later in this report.

Select training activities to date

Face-to-face training

Due to the initial lack of expertise across specialised bioinformatics analysis areas, a significant amount of effort was put into developing and hosting formal training workshops. Over the period of the project, the network has attempted to cover as many specialised topics through the presentation of hands on, face-to-face workshops, ranging from a Train-the-Trainer workshop (covering various fundamental bioinformatics topics), to specialised training in GWAS, metagenomics, and NGS data analysis. Some of these earlier face-to-face workshops were presented by invited expert trainers from abroad, however, subsequent training was conducted by local H3ABioNet members who were identified and mentored to become competent local trainers. In addition to the formal hosting and funding of workshops, H3ABioNet has also contributed towards the planning and support of various workshops run locally at its nodes. H3ABioNet has also provided travel fellowships to H3Africa members wishing to attend training workshops provided by other H3Africa projects in order to promote interaction and knowledge transfer between these projects.

Internship program

Although the face-to-face training approach proved successful in providing participants with a good foundation to return to their institutes and further develop their skills with some assistance, the short term contact did not provide enough exposure to allow the participants to confidently and independently analyse their own data. To complement the more formal training workshops, H3ABioNet set up an internship program for both H3ABioNet and H3Africa consortium members to spend a dedicated period of time at a host laboratory in order to acquire bioinformatics skills relevant to their research interests. The host laboratories in the internship program were both local and international, with the inclusion of a dedicated internship opportunity at the Harvard node of H3ABioNet. There have been 16 internship placements of H3ABioNet students to date and interns have visited both local and international laboratories to learn specialised analysis skills in: GWAS, metabolic network and structural modelling, microbiome

Article history
Received: 06 April 2017
Published: 11 May 2017

[1] http://www.h3abionet.org
[2] http://h3africa.org/

data analysis, NGS data analysis, computational system administration, and big data transfer processes. The internship program has proven useful in providing young, enthusiastic students with the opportunity to not only develop specialised bioinformatics skills, but also to collaborate and interact with leading groups in their field of interest. Upon returning to their home institutes, the interns were encouraged to share the knowledge gained with their peers through local training events and were earmarked to be teaching assistants for the next workshop relevant to the specific skills they had obtained.

Online training

In an effort to further increase access to training across Africa, H3ABioNet has begun offering courses with an online component. In 2016, the network ran the first iteration of its Introduction to Bioinformatics course (IBT). The course was very popular with over 350 enrolled participants and over 70 volunteer staff in total, across 20 local classrooms spanning ten African countries. During this three month course, classrooms met each other and the trainer biweekly in a virtual classroom.

Hackathons

As the bioinformatics capacity across the network developed, a critical mass of expertise across a range of bioinformatics areas was developed. Together with the imminent arrival of genomic data from the H3Africa projects, the need arose for a shift from capacity development to data centric events and outcomes based hackathons. The *H3ABioNet Infrastructure and Research working groups* spearheaded this endeavour. The hackathons were aimed at gathering a group of people with a diverse set of skills - from computer programmers and bioinformaticians to biologists, clinicians, biostatisticians, and mathematicians - to develop and implement solutions to a particular bioinformatics problem. H3ABioNet has hosted two hackathons to date, both yielding very successful outcomes. The first was a *Cloud Computing Hackathon* aimed at developing reproducible workflows for anticipated H3Africa bioinformatics analysis pipelines available for use on heterogeneous computing environments as Docker containers, which can also be deployed on the cloud infrastructure. H3ABioNet's second hackathon was hosted in partnership with IBM Research Africa and the University of Notre Dame and was aimed at addressing a research question that forms part of a DREAM challenge on malaria drug resistance. As more data becomes available through the various H3Africa projects, and the availability of individuals with a range of skills within the network increases, the hackathon format for workshops will be implemented more regularly in future training events in the network.

Challenges

At the onset of the training programs developed and implemented by H3ABioNet, it was anticipated that there would be challenges to face; some common to other training programs and some unique to the African setting. Some of the more relevant challenges that had a direct effect on the sustainability of the training in Africa included: the high costs of airline tickets within the continent; acquiring visas on time; availability of suitable computational infrastructure at training sites and once participants returned home; socio-political instability; and over subscription to courses due to high demand. On a more positive note, the occurrence of these challenges encouraged a more flexible approach to training and led to the implementation of the different formats of training described above. In particular, the training with an online component has proven to be a cost efficient and sustainable format for training in Africa. Further, H3ABioNet funding enabled many centres to purchase equipment, and the network's *Infrastructure working group* has worked closely with these centres to set up their own computing infrastructure. In this way, many more training sites are available and individuals have greater access to resources after the training events have concluded.

Conclusions

H3ABioNet has aimed to create a sustainable approach to further the development of bioinformatics capacity in Africa in partnership with the H3Africa Consortium. Together with complementary training initiatives focused on capacity development in Africa, H3ABioNet has provided numerous young African scientists with access to bioinformatics training opportunities, better equipping them to pursue a career in the field. Planning and implementing training in Africa has also brought to light the many obstacles that are faced by African institutions in developing bioinformatics skills. Some of these obstacles can be addressed and overcome using alternative training methods to develop a sustainable approach to bioinformatics capacity development in Africa.

Acknowledgements

Research reported in this publication was supported by the National Human Genome Research Institute (NHGRI) and the Office of the Director (OD), National Institutes of Health under award number U41HG006941. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Mulder NJ, Adebiyi E, Alami R, Benkahla A, Brandful J (2016) H3ABioNet, a sustainable pan-African bioinformatics network for human heredity and health in Africa. *Genome Res.* **26**, 271-277. http://dx.doi.org//10.1101/gr.196295.115
2. H3Africa Consortium., Rotimi C, Abayomi A, Abimiku A, Adabayeri VM (2014) Research capacity. Enabling the genomic revolution in Africa. *Science* **344**, 1346-1348. http://dx.doi.org/10.1126/science.1251546

[3] https://www.ibm.com/blogs/research/2016/10/hacking-anti-malarial-drug-resistance

Aron et al. (2017) EMBnet.journal 23, e886
http://dx.doi.org/10.14806/ej.23.0.886

*H3ABioNet Education and Training working group members

Nicola Mulder[1], Shaun Aron[2], Sumir Panji[1], Kim Gurwitz[1], Judit Kumuthini[3], Shakuntala Baichoo[4], Segun Fatumo[5], Oyekanmi Nash[5], Jonathan Kayondo[6], Faisal M. Fadlelmola[7], Samar Kassim[8], Jean-Baka Domelevo Entfellner[9], Odile Ouwe Missi Oukem-Boyer[10], Samson Pandam Salifu[11], Winston Hide[12,13], Kais Ghedira[14], Amel Ghouila[14], Lerato Magosi[15], Alia Benkhala[14], Victoria Nembaware[1], Mary Piper[12], Radhika S. Khetani[12], Anne Fischer[16].

[1]Computational Biology Division, Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town, South Africa; [2]Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg, South Africa; [3]Centre for Proteomic and Genomic Research, Cape Town, South Africa; [4]University of Mauritius, Moka, Mauritius; [5]National Biotechnology Development Agency, Abuja, Nigeria; [6]Uganda Virus Research Institute, Entebbe, Uganda; [7]Centre for Bioinformatics and Systems Biology, Faculty of Science, University of Khartoum/ Future University of Sudan, Khartoum, Sudan; [8]Medical Biochemistry and Molecular Biology Department, Faculty of Medicine, Ain Shams University, Cairo, Egypt; [9]South African National Bioinformatics Institute/Medical Research Council of South Africa Bioinformatics Unit, University of the Western Cape, Cape Town, South Africa; [10]Centre de Recherche Medicale et Sanitaire, Niamey, Niger / Cameroon Bioethics Initiative (CAMBIN); [11]Kumasi Centre for Collaborative Research in Tropical Medicine/Kwame Nkrumah University of Science and Technology, Kumasi, Ghana; [12]Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA; [13]Sheffield Institute for Translational Neuroscience, Department of Neuroscience, University of Sheffield, Sheffield, United Kingdom; [14]Institute Pasteur of Tunis, Tunis, Tunisia; [15]Botswana Harvard AIDS Institute Partnership, Gaborone, Botswana/ University of Oxford, Oxford, United Kingdom; [16]Molecular Biology and Biotechnology Department, International Centre for Insect Physiology and Ecology, Nairobi, Kenya.

EMBnet.journal
Bioinformatics in Action

Pfeil et al. (2018) EMBnet.journal 23, e905
http://dx.doi.org/10.14806/ej.23.0.905

Reports

Report on the "Big Data Training School for Life Sciences", 18-22 September 2017, Uppsala, Sweden

Juliane Pfeil[1]✉, **Sabrina K. Schulze**[2], **Eftim Zdravevski**[3], **Yen Hoang**[4]

[1] Division Molecular Biotechnology and Functional Genomics, Technical University of Applied Sciences, Wildau, Germany

[2] Cell2Fab (Synthetic Biology, Faculty of Biochemistry and Biology), University of Potsdam, Potsdam, Germany

[3] Department of Information Systems, Faculty of Computer Science and Engineering, Sts. Cyril and Methodius University in Skopje, Skopje, Macedonia

[4] Department of Signal Transduction, German Rheumatism Research Center Berlin, A Leibniz Institute, Berlin, Germany

Competing interests: JP none; SKS none; EZ none; YH none

Abstract

In September 2017 a "Big Data Training School for Life Sciences" took place in Uppsala, Sweden, jointly organised by EMBnet and the COST Action CHARME (Harmonising standardisation strategies to increase efficiency and competitiveness of European life-science research - CA15100). The week programme was divided into hands-on sessions and lectures. In both cases, insights into dealing with big amounts of data were given. This paper describes our personal experience as students' by providing also some suggestions that we hope can help the organisers as well as other trainers to further increase the efficiency of such intensive courses for students with diverse backgrounds.

Course of the training school

The "Big Data Training School for Life Sciences"[1] was a joint initiative of the EU COST Action CHARME[2] and EMBnet[3]. The main objective of these organisations is to network scientists of different countries (within Europe and neighbouring areas) to serve, support and sustain the biological and biomedical research. The aim of the training school was to increase the efficiency of life-science research and interdisciplinary collaboration by training students and researchers who have to cope with the need to manage big data for their research activity. The school took place from 18th to 22nd September 2017 at the Campus Ultuna of the Swedish University of Agriculture[4] (SLU) in Uppsala, Sweden.

The programme included lectures accompanied by hands-on sessions in the first three days (September 18-20), and lectures (Lecture days, 21-22 September) open to a wider audience in the last two days.

The hands-on sessions, as well as the lectures, dealt with different topics in the field of Bioinformatics for big data management and analysis. In total 25 students (the max. number allowed) took part in the first three days.

Figure 1 shows a nice group picture of some trainers and of trainees inside the building of the Ultuna Campus.

The participants for the hands-on sessions applied for a stipend (CHARME grants) several months in advance. The scientific committee[5] selected them based on diverse criteria, including research background and motivation, while also attempting to ensure an equal distribution of CHARME grants on the basis of national, gender and ethical diversity.

Trainees selected had different scientific backgrounds: bioinformaticians with limited knowledge in Big Data technologies, computer scientists that moved to bioinformatics, and scientists with biological/medical backgrounds with some knowledge in bioinformatics.

The first day was opened by Erik Bongcam-Rudloff, Professor of Bioinformatics at SLU and vice-Chair of CHARME. He gave an overview and showed the impact of Big Data in different areas of life and its potential for the future. The day was closed by Dr Jim Dowling, a distributed systems researcher at RISE SICS[6] and Associate Professor at KTH ICT School[7]. He presented a lecture and a hands-on session about Hops Hadoop, an open-source platform for analysing Big Data in an uncomplicated way. The second day was opened with an introduction to machine learning by Gioele La

[1] http://astrocyte.com/COST-CHARME/COST-CHARME/Home.html

[2] http://www.cost-charme.eu

[3] https://www.embnet.org

[4] https://www.slu.se/en/

[5] http://astrocyte.com/COST-CHARME/COST-CHARME/About.html

[6] https://www.sics.se/

[7] https://www.kth.se/en/ict

Article history
Received: 19 December 2017
Published: 05 February 2018

Pfeil et al. (2018) EMBnet.journal 23, e905
http://dx.doi.org/10.14806/ej.23.0.905

Figure 1. Group picture of the participants and some of the lecturers of the first three days. @Photo by Erik Bongcam-Rudloff.

Manno (Karolinska Institutet) followed by a practical example in Apache Spark presented by Marco Capuccini (Uppsala University). In another hands-on session in the afternoon, Apurva Nandan (CSC - IT Center for Science, Finland) introduced Apache Spark SQL, a toolkit for easy parallelisation of computations. The day was finished by Dr Vaughan Wittorf from PetaGene[8], explaining the idea of how storing and transferring next-generation sequencing (NGS) data can be done more efficiently. On Wednesday, Dr Witold Rudnicki (Associate Professor at University of Białystok) started with a session about gene expression data and a possible way to extract informative changes with R and the classification method Random Forests. Kim Kultima, Payam Emami, Stephanie Herman, and Ola Spjuth from Uppsala University were the last speakers on this day. They introduced their method of handling metabolomics data produced by mass spectrometry with OpenMS and Pachyderm. Their work is dedicated to the two major projects Caramba.clinic[9] and Phenomenal[10].

The last two lecture days were opened by Erik Bongcam-Rudloff and Domenica D'Elia (Chair of EMBnet and of the Dissemination Working Group in CHARME). After the introduction of the objectives of EMBnet and CHARME, there was a presentation from Roxana Merino Martinez (Karolinska Institutet) that presented the B3Africa project[11]. During the day, there were presentations by the two speakers Gaurav Kaul and Erik Gullbring from the technology leaders Intel and Microsoft, respectively, that showed powerful frameworks for cloud computing and machine learning. The importance of the programming language R was highlighted by Seija Sirkiä (CSC). During this day, Prof. Witold Rudnicki presented more information about the work he shortly introduced during the hands-on sessions. Ola Spjuth gave a deeper insight into the project Phenomenal. Bjorn Lindell from SLU focused on the impact of cloud solutions. The day was closed with an open and fruitful discussion about the use of Big Data technologies in bioinformatics. The last day was introduced by Anders Herlin (SLU), who showed possibilities to improve the conditions of livestock farming using Big Data. Kim Kultima also gave more insights into the projects Caramba.clinic and Phenomenal. After that, Nataša Sladoje (Centre for Image Analysis, Uppsala University) showed her work and some possible training opportunities of the COST

[8]https://www.petagene.com/

[9]http://www.caramba.clinic

[10]http://phenomenal-h2020.eu

[11]http://www.b3africa.org/

Action NEUBIAS[12] (Network of European Bioimage Analysis). In this context, one participant of the training school was asked for a spontaneous presentation of the mobile microscopic system from the company Oculyze[13]. Possible applications, especially in the B3Africa project, were discussed. The last day was closed by Professor Dimitrios P. Vlachakis from the Biotech Department of the Agricultural University of Athens, Greece, who presented a long-term scientific project about NGS data of inhabitants in Cyprus.

Pros and Cons from students' point of view

As mentioned above, this was the first edition of this training school. The transmitted contents of the lessons fit very well into the area of Big Data and highlighted the topic from different perspectives. Despite the heterogeneous education background of the participants, everyone had certainly obtained helpful information about the way to improve the quality of their work, and the exchange of information and comments among participants was very rewarding. All the lecturers were keen in passing along their knowledge and their experience to the audience. When questions came up, they were immediately answered or their assistants came to solve the problems directly with the person who had asked help. Unfortunately, the available time was too short to satisfy the needs of all attendants, especially during the hands-on sessions.

After the school conclusion on Friday afternoon, the organisers invited us to discuss and report our personal opinion on the school to improve possible future editions. The discussion was open to all attendants and after a constructive exchange of comments and ideas we were invited by Domenica D'Elia and Erik Bongcam-Rudloff to submit this article for its publication in the EMBnet.journal. The main objective was to allow a wider dissemination of the challenges encountered during hands-on training of aspects related to Big Data issues. We hope that it will allow trainers to better programme the learning objectives of this type of schools to the actual possibility to transmit huge amounts of knowledge in the space and time that a one-week training school can have.

Our first impression was that it would have been more advantageous if the interesting lectures, which had taken place on Thursday and Friday, would have occurred before the hands-on sessions. Much of the basics could have been covered here instead of during the sessions. By doing so, the time of the hands-on sessions would not have been as limited to introduce the methods by the lecturer. More than once it was noticed that the lecturer had more to tell, but could not do because the time slots ended. Still, this introduction was necessary due to the diverse background of the attendees.

As highlighted earlier, the participants had different scientific backgrounds. This diversity made it a challenging task to adequately set the learning goals. The goals of the school ranged from introductions to machine learning and Big Data technologies, to ways of applying such technologies for addressing computational challenges in bioinformatics. The selection of more uniform groups of participants could have allowed trainers to focus much on trainees' specific needs. Nevertheless, this approach has a major drawback that it hinders interdisciplinary collaboration, one of the most important goals of this training school. Indeed, the advantage of the education background diversity was that the students were forced to interact more with each other by asking questions and offering support in a complementary way. By doing so, personal interactions were spontaneously fostered.

A way to somewhat balance out the different trainees' backgrounds could be achieved by "assigning some homework" before the training school. In this case, the trainee could have gotten a list of required knowledge and maybe texts/links to study by themselves beforehand. This would improve the effectiveness of training by allowing trainees to concentrate much more on the hands-on work. Indeed, during the hands-on sessions, participants faced for the first time with a large amount of material to deal with. It was necessary to understand the lesson and to do some practical work in parallel. Although this helped enhance the effects of the learning, in some cases (*e.g.*, students with a poorer bioinformatics background) it lowered its efficiency.

Another aspect that should be improved is related to the preparation of the computer environment for the training school that includes software and other materials. The downloads of this material during the training sessions took away precious time, which could have been used to better address the issues in Big Data. Moreover, because of the different computers' efficiency in downloading, not all machines were ready to work in due time. Therefore, the hands-on sessions were resumed after the first few successful downloads, as the lecturers had to respect the timetable for the explanation of methods and tools foreseen to be completed in the session. Many participants had to follow the meaningful input and monitor the installation in parallel. This was complicated and forced people to focus on one thing or the other.

Another suggestion we would like to provide is about the number of topics that the school should include. It was good for us to know the huge variety of bioinformatic issues related to Big Data and the variety in finding new and more effective solutions, but it would have been even better if the focus was on fewer topics, but with more intensity.

If bioinformaticians want to use Big Data technologies, it is not always clear where to start, as the information provided in a school can be overwhelming. To improve this aspect, the application for the school can include a survey about what common algorithms prospective students use or have used in the past. Then,

[12]http://eubias.org/NEUBIAS/

[13]www.oculyze.de

Pfeil et al. (2018) EMBnet.journal 23, e905
http://dx.doi.org/10.14806/ej.23.0.905

the results of such a carefully crafted survey can help in identifying some algorithms that could be presented during the course for a Big-Data approach and a traditional method. Comparison in terms of performance for some use-cases could further complement this. This can engage bioinformaticians more and incline them to push for Big Data approaches in their departments. This can ultimately lead to initiatives for reusable open source implementation of popular algorithms in Big Data technologies, which could draw people from computer science into bioinformatics.

Overall, we had an interesting week where we learnt a lot about dealing with Big Data. The organisers, as well as the lecturers, did a good job communicating present Big Data challenges and solutions. Our above remarks should be considered as mere constructive feedback and suggestions for improvement of other training schools, not only on Big Data but in general.

Conclusions

The “Big Data Training School for Life Sciences” was an excellent idea for learning trends in this field for young scientists and perfect for networking purposes. Realising the existence of immense computational challenges in bioinformatics is of great importance for interdisciplinary collaboration, and the school was particularly successful in this regard. The networking part was done during the breaks and in the evenings. One of these opportunities was during the “Ethno-Party” on Thursday. Every participant had been asked to bring some food and drinks from their home country to share.

Some people stayed in touch even after the end of the school. A follow-up symposium with the title “Bioinformatics meets Synthetic Biology” was organised. It took place on 20th October 2017 in Potsdam-Golm, Germany. Five participants of the original school and five other scientists presented their projects and work.

Acknowledgement

We would like to thank SLU for organising the school. A huge thanks to Erik Bongcam-Rudloff and his organisation team for their great work. Also, thanks to EMBnet and to COST Action CHARME for funding this school and the stipends so that we had the opportunity to attend. Another thanks to all the lecturers and speakers during this week who gave us some insight in their work and personal experiences. The "Big Data Training School for Life Sciences" 18-22 September 2017 Uppsala, Sweden, was financed by the COST action CA15110 supported by the EU framework program H2020.

EMBnet.journal
Bioinformatics in Action

Gill et al. (2017) EMBnet.journal 23, e892
http://dx.doi.org/10.14806/ej.23.0.892

Technical Notes

PolarFCS: A Multi-Parametric Data Visualisation Aid for Flow Cytometry Assessment

Pavandeep Gill 1, Joanne Luider 1, Etienne Mahe 1 ✉

[1]University of Calgary, Calgary, Alberta, Canada

Competing interests: PG none; JL none; EM none

Abstract

Currently available Flow-Cytometry Software (FCS) analysis platforms are computationally efficient and user-friendly, but may lack the functionality of single-plot, multi-parametric data visualisation. Methods to overcome this include gating techniques and/or dimensionality reduction. However, these strategies make Flow-Cytometry (FC) data analysis more time- and labour-intensive; profound errors can also result from incorrect FCS use. We have developed PolarFCS, a software tool capable of single-plot, multi-parametric data visualisation. Unlike traditional clinical FC plots, which typically operate directly on a data-set to produce single-parameter FC histograms or two-parameter orthogonal scatter plots, PolarFCS operates on the flow-parameter calculated centre of mass of each event in the data-set, and presents these as a dot-plot. We compare PolarFCS to our traditional clinical FCS workflow, using a selection of clinical plasma cell FC data. Multiple flow plots and gating strategies are required in the traditional software to isolate neoplastic populations. In PolarFCS, however, positional re-arrangement and scaling of the poles can be used to quickly isolate a population of interest. We also compare both approaches in a case of Minimal Residual Disease (MRD) assessment, and again, the versatility of the polar adjustment and parameter scaling allowable with PolarFCS is demonstrated. PolarFCS employs strategies that allow more accurate, standardised and detailed FC data analysis compared to traditional FCS platforms. Visualisation of multiple parameters in a single plot is an effective and invaluable feature that many other platforms currently do not offer.

Availability: PolarFCS can be downloaded at **https://github.com/etiennemahe/PolarFCS**

Introduction

Flow cytometry (FC) is a powerful technique through which multiple physical and chemical characteristics (e.g., size, granularity, viability and immunophenotype) of several thousands of cells can be rapidly analysed. In analytical FC, cells in solution are individually passed through an optical flow cell, whereafter scattered incident light can be used to interpolate cellular characteristics. When cells are pre-treated with fluorescently-tagged antibodies, a highly sensitive immunophenotypic signature can also be obtained, exploiting variable incident excitation frequencies and variable emission spectra. Of note, the intensity of light emitted by each antibody-linked fluorophore can be related to antigen expression, and panels containing a number of antibody-linked fluorophores can be easily employed (Verbsky and Routes, 2017). The resulting digitised data are amenable to analysis by FC Software (FCS).

Several FCS platforms are currently available, both as freeware and as commercially available software. While many FCS platforms are computationally efficient and user-friendly, current standard platforms are subject to limitations in functionality. Specifically, the standard means of FCS visualisation involve presentation and analysis of FC data as either single-parameter histograms or two-parameter-correlated plots (either dot plot, density or contour plots) (Ormerod, 2008). As such, these approaches make it impossible to directly visualise data in multi-parameter form (*i.e.*, more than two simultaneous parameters in a single plot). To overcome this, gating techniques are frequently used to analyse subsets of FC data (*e.g.*, using forward scatter and side scatter to analyse only lymphocytes), followed by separate analyses of additional parameters (*e.g.*, lymphocyte CD3 or CD20 intensity) (Ormerod, 2008). These data can then be gated again to produce histograms of even further sub-populations. Gating typically requires the manual isolation of FC-event data in order to highlight cell sub-populations of interest, and requires expert knowledge of the underlying cell populations. Gated FC analysis is a time- and labour-intensive process, and incorrect gating strategies can result in profoundly erroneous FC analysis results, especially when very minor sub-populations are considered. Additionally, a limited two-parameter plot makes it difficult to visualise important multi-parameter relationships that cannot be seen in just two dimensions.

Article history
Received: 17 March 2017
Accepted: 8 May 2017
Published: 11 October 2017

Gill et al. (2017) EMBnet.journal 23, e892
http://dx.doi.org/10.14806/ej.23.0.892

To assist users in overcoming the limitations of traditional FC data analysis, we have developed PolarFCS, a freely-available software tool capable of single-plot, multi-parametric data visualisation. We tested PolarFCS on a selected series of bone marrow specimens submitted for plasma cell FC, and use these exemplars to highlight some of the features that a multi-parametric FC analysis might afford the user.

Materials, Methodologies and Techniques

Technical Summary: PolarFCS

PolarFCS provides a means of single-plot, multi-parametric FC analysis. Unlike traditional FC plots, which operate directly on a data-set, producing either single-parameter FC data histograms or two-parameter orthogonal scatter plots, PolarFCS operates on the flow-parameter calculated centre of mass of each event in the data-set. To do this, PolarFCS begins by considering each FC parameter as a polar axis in two-dimensional space. For each event, the signal intensity of each parameter is plotted along the length of each parameter's corresponding polar axis. The result, when these coordinates are subtended by lines, is a polygon in two-dimensional space. A simple algorithm[1] is then applied to determine the polygon's centre of mass. The centres of mass for each event are then presented as a dot-plot.

Akin to other FC data-analysis platforms, PolarFCS also allows users to interact with the data-set to permit more detailed data analysis. PolarFCS is able to incorporate previously determined gating strategies by way of dot-plot colour variation; it also allows users to adjust the orientation of the polar axes. Such adjustments will have the effect of re-positioning certain data subsets relative to others. Such adjustments can also be used to amplify or dampen the contribution of a given parameter relative to the others. Finally, an adjustable counting field allows users to accurately determine the number of events within a given area relative to the total event count. This latter tool aims to assist users in comparing event fractions between studies, or to compare the relative abundance of a given subset of events.

PolarFCS is written in MATLAB[2] and consists of two primary functions. The `masthead` PolarFCS function offers users a basic graphical-user-interface to select data files for input (either as tabular '.txt' or '.csv' data, or as FCS-formatted '.lmd' List Mode Data files[3]), and to select and rename FC data parameters as desired. The second, `makefcspolarscatter`, function can be invoked directly by MATLAB users, requiring a user-defined uncompressed data-set, parameter list, plot title and colour matrix. The colour matrix allows users to encode a specific pre-defined gating strategy in the form of colour variation in the final plot. The PolarFCS Windows- and Mac OSX-compiled stand-alone versions, as well as the source code for the `makefcspolarscatter` function, are available for download at https://github.com/etiennemahe/PolarFCS. Other compilation formats and additional source code can be made available by specific request to the corresponding author.

[1]http://www.seas.upenn.edu/~sys502/extra_materials/Polygon%20Area%20and%20Centroid.pdf

[2]https://www.mathworks.com/products/matlab.html

[3]More detailed formatting descriptions are provided at https://github.com/etiennemahe/PolarFCS

Results

Application Example: PolarFCS as a visual aid in plasma cell FC

The current haematopathology work-up of plasma cell neoplasia relies heavily on FC. Plasma cell FC provides

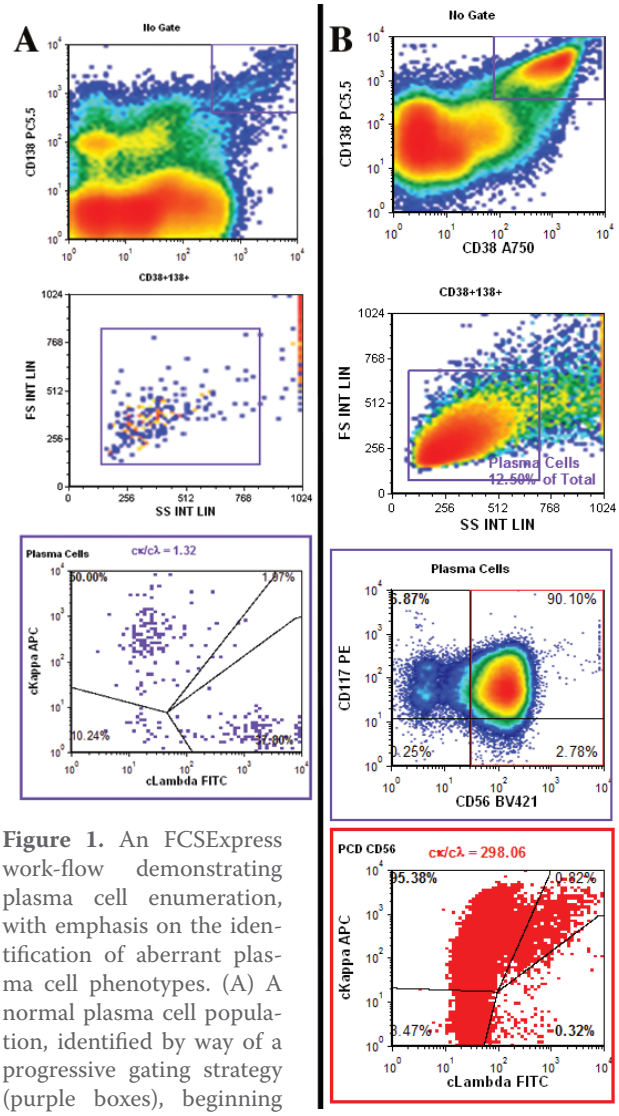


Figure 1. An FCSExpress work-flow demonstrating plasma cell enumeration, with emphasis on the identification of aberrant plasma cell phenotypes. (A) A normal plasma cell population, identified by way of a progressive gating strategy (purple boxes), beginning from CD38/CD138-bright events (which are notably few in number), followed by gating around a region of typical plasma cell forward and side-scatter, followed finally by assessment of cytoplasmic kappa/lambda ratios (which are within the normal limits in this case); (B) A neoplastic plasma cell population, identified by way of a comparable progressive gating strategy, with aberrant dual CD56/CD117 expression (red-highlighted gate), and an associated extreme increase in the cytoplasmic kappa/lambda ratio.

Gill et al. (2017) EMBnet.journal 23, e892
http://dx.doi.org/10.14806/ej.23.0.892

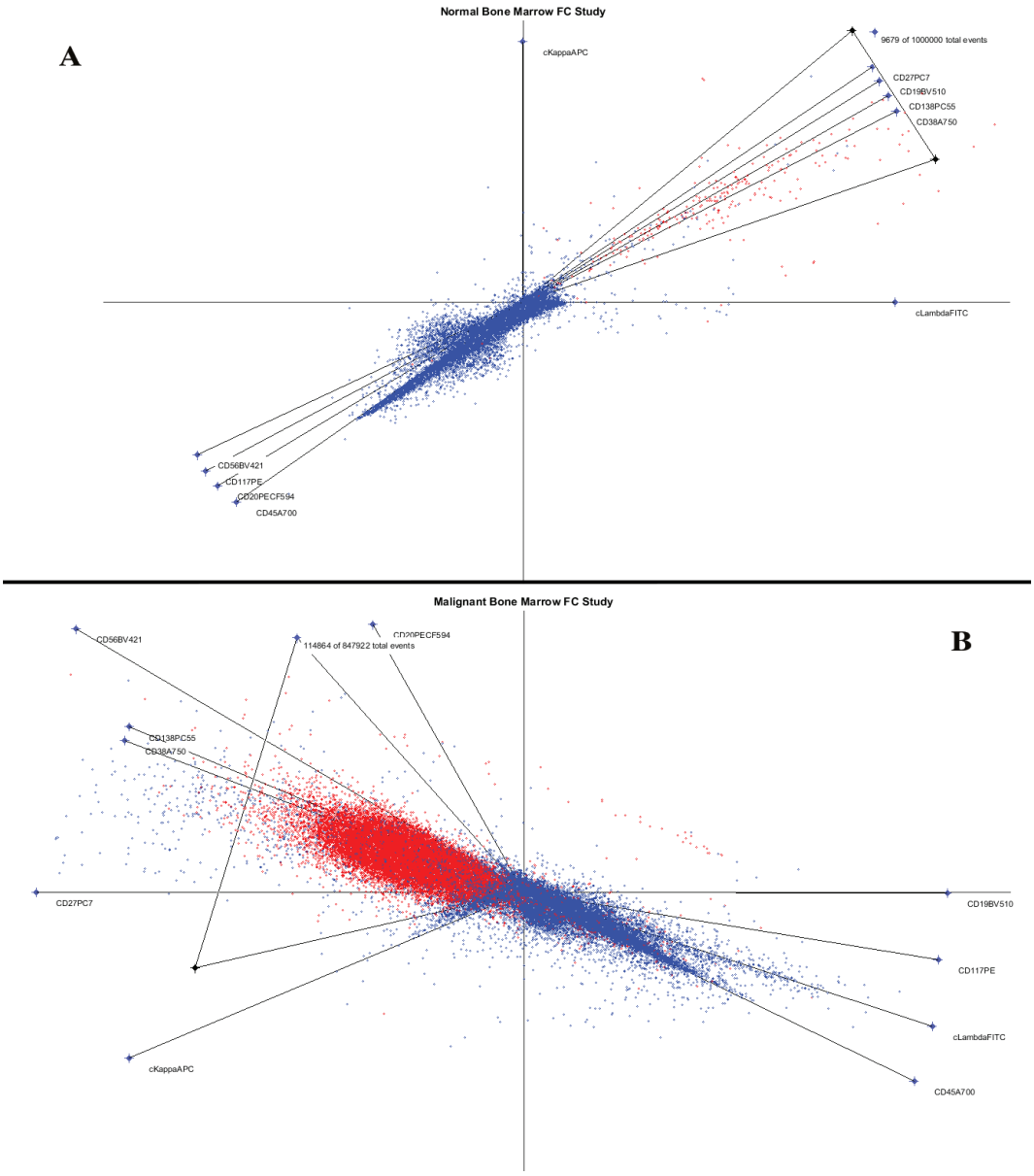


Figure 2. PolarFCS demonstrating plasma cell enumeration, with the plasma cells identified by the gating strategies in figure 1 highlighted red. (A) Normal plasma cell population; (B) Neoplastic plasma cell population. Note that positional re-arrangement and scaling of the PolarFCS poles can be used to isolate populations of interest.

a rapid and sensitive quantitation of the plasma cell fraction from bone marrow specimens, and allows rapid and reproducible clonal assessment. FC can be used to identify aberrant immunophenotypes suggestive of neoplasia, and plasma cell FC permits highly sensitive Minimal Residual Disease (MRD) assessment.

To explore how PolarFCS might be applied in these contexts, we identified example clinical cases to compare our standard FC informatic workflow (FCSExpress5, IVD, DeNovo Software, Glendale, CA[4]) with one employing PolarFCS[5]. For demonstration purposes, we exported the gated populations of interest, as well as the overall non-gated study data, as raw text-formatted data, and subsequently employed MATLAB's set algebra functions to define a custom colour profile for input into the `makefcspolarscatter` function. In our jurisdiction, the type of exploratory work outlined in this manuscript qualifies as a 'quality assurance or evaluation project' (ARECCI Ethics Screening Tool - 2005, revised 2010[6]), and therefore does not require direct institutional ethics board review. Nevertheless, the tenets of the Declaration of Helsinki, as further required by Canadian Tri-Council Research Ethics Panel (Canadian Institutes of Health Research, 2014[7]), were adhered to rigorously.

[4]https://www.denovosoftware.com/

[5]Some of the de-identified data-sets used in our analyses are available for download at https://github.com/etiennemahe/PolarFCS

[6]http://www.aihealthsolutions.ca/media/Sept-2013-Paper-Version-ScreeningTool.pdf

[7]http://www.pre.ethics.gc.ca/pdf/eng/tcps2-2014/TCPS_2_FINAL_Web.pdf

Gill et al. (2017) EMBnet.journal 23, e892
http://dx.doi.org/10.14806/ej.23.0.892

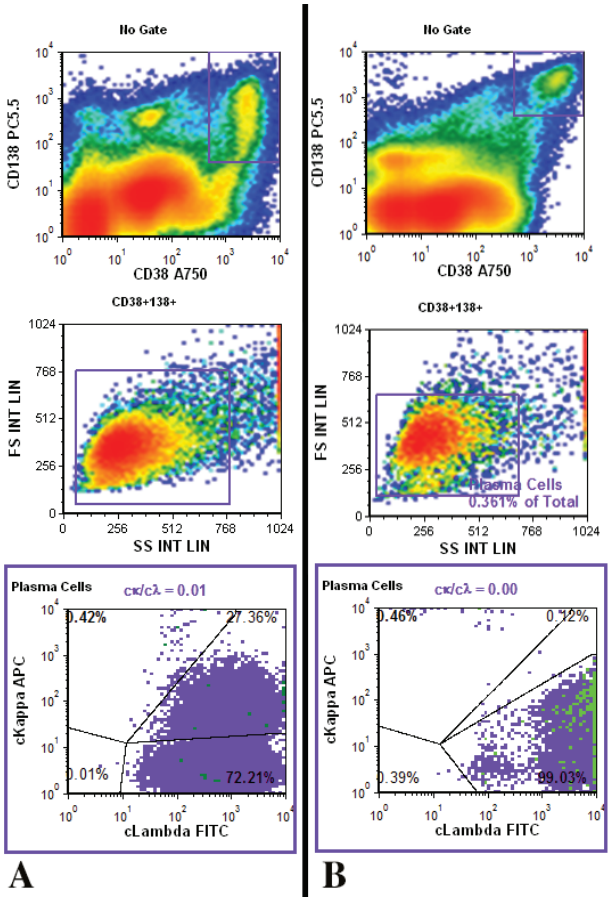


Figure 3. FCSExpress workflow highlight a minimal residual disease work-up. (A) Plasma cell myeloma at diagnosis; (B) Follow-up status post autologous bone marrow transplant.

Plasma cell enumeration, with emphasis on the identification of aberrant plasma cell phenotypes, is shown in Figures 1-2. Figure 1A delineates the orthogonal flow plots required to gate out a small phenotypically normal plasma cell population, whereas Figure 1B highlights an abundant neoplastic (and immunophenotypically aberrant) plasma cell population. In contrast, the appertaining PolarFCS plots are shown in Figure 2 ('Normal' case in Figure 2A and 'Plasma Cell Myeloma' in Figure 2B). This contrast serves to highlight the clear difference between 'normal' and 'malignant' PolarFCS signatures, as well as the versatility of pole adjustment and parameter scaling allowable in PolarFCS. Of note, PolarFCS allows poles corresponding to expected 'normal' plasma cell phenotypic features to be subtended in opposition to 'abnormal' phenotypes; in doing so, PolarFCS allows clear separation of these distinct sub-populations, without an otherwise required multi-step gating strategy.

MRD assessment is highlighted in Figures 3-4. Figure 3 delineates the orthogonal flow plots of a case of plasma cell myeloma, at diagnosis (Figure 3A) and at follow-up status, post autologous bone marrow transplant (Figure 3B). In contrast, the appertaining PolarFCS plots are shown in Figure 4 (at diagnosis in Figure 4A, and at follow-up in Figure 4B). This contrast serves to highlight the persistent disease population identifiable in both studies, facilitated by positioning of the PolarFCS poles in comparable orientations. The PolarFCS plot in Figure 4B also demonstrates a characteristic 'wall-artefact', partly highlighted in green in the lowest flow plot of Figure 3B. Arguably, this artefact is more readily apparent in the PolarFCS plot. When the original and corresponding PolarFCS gated-event counts are compared, it becomes apparent that the reference flow plot likely over-estimates the gated-event count.

Discussion

The application of FC to the work-up of haematological malignancies is invaluable. Each year, our laboratory receives several thousand specimens for FC analysis; and, from among these, several hundred cancer cases are diagnosed (Canadian Cancer Society, 2016[8]). Thus, in order to optimise laboratory processes, an efficient FC workflow requires both sensitive and efficient FCS. Current clinical FCS systems might be amenable to additional process improvements, including dimensionality reduction techniques, for visualisation of multiple parameters in a single plot.

Examples of currently available tools for dimensionality reduction include Spanning-tree Progression Analysis of Density-normalised Events (SPADE), Principle Component Analysis (PCA), and viSNE. SPADE works by clustering cells into sub-populations, measuring the average of each cluster, and then presenting the data in a branched-tree structure showing the connections between the different clusters (Qiu *et al.*, 2011). PCA tries to maintain single-event data through linear transformations of relationships between putatively correlated variables; however, this limits PCA in its application to non-linear relationships (Amir *et al.*, 2013). viSNE is a dimensionality-reduction software using the t-distributed Stochastic Neighbour Embedding (t-SNE) algorithm, and has the capacity for single-event data, and visualisation of non-linear relationships (Amir *et al.*, 2013).

With these (and other) dimensionality-reduction tools, there are limitations to how much of the high-dimensional data can be represented in a low-dimensional map. Detail relating to less dominant parameters can be lost using these approaches. In addition, there may be limited user control over which parameters are included or excluded. In contrast, PolarFCS allows users to select the inclusion of parameters, as well as their scaling, by way of polar adjustment. The number of events that can be analysed using viSNE also has an upper limit; as such, viSNE may need to be run on a well-defined, pre-filtered subset of data (possibly requiring additional pre-analytical analysis algorithms). No such data limits are inherent to PolarFCS.

It bears noting that PolarFCS may also be subject to limitations in its current version. For example,

[8]http://www.cancer.ca/~/media/cancer.ca/CW/cancer%20information/cancer%20101/Canadian%20cancer%20statistics/Canadian-Cancer-Statistics-2016-EN.pdf?la=en

Gill et al. (2017) EMBnet.journal 23, e892
http://dx.doi.org/10.14806/ej.23.0.892

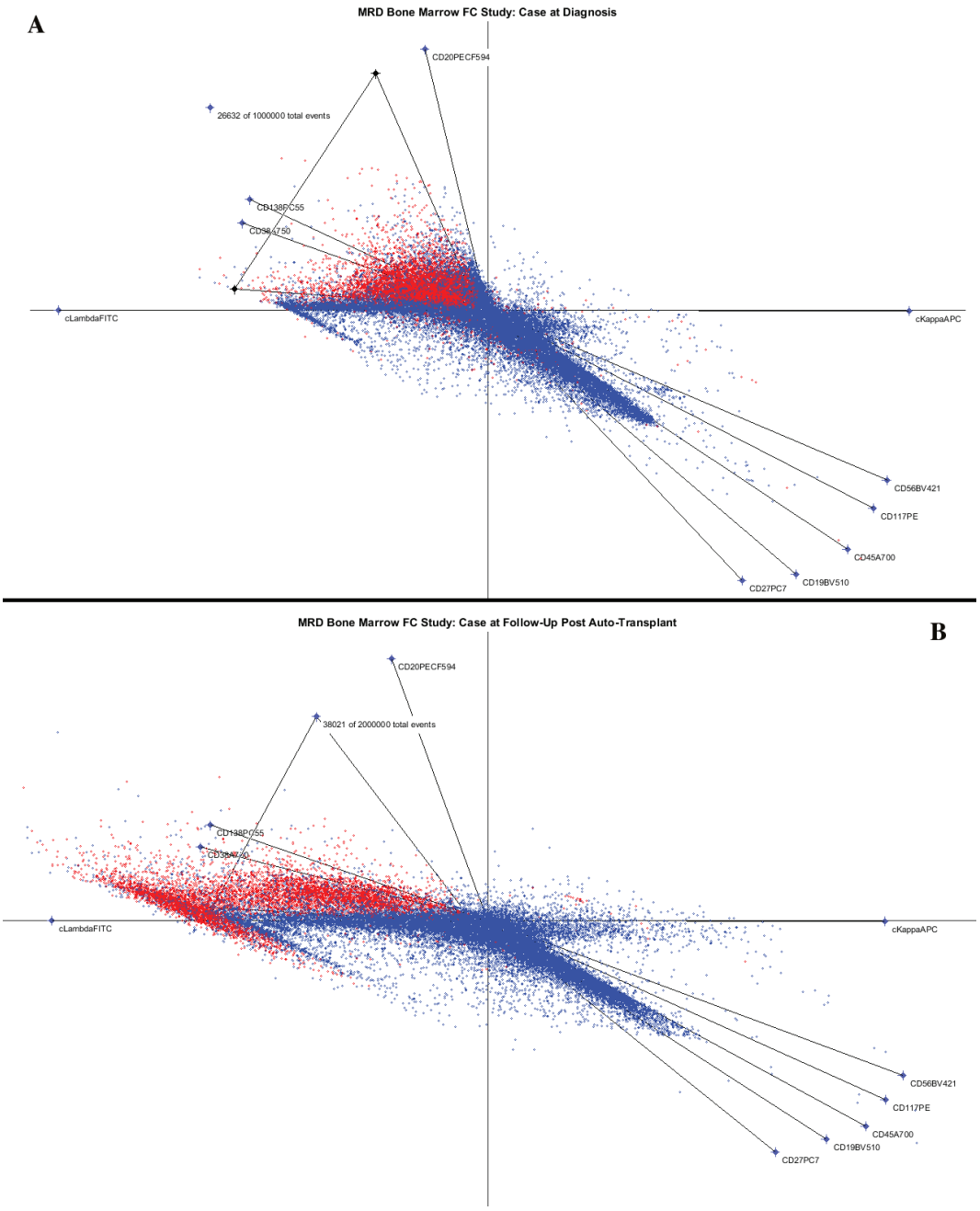


Figure 4. PolarFCS applied to the cases highlighted in Figure 3. (A) Plasma cell myeloma at diagnosis; (B) Follow-up status post autologous bone marrow transplant. Note that PolarFCS polar axes can be arranged in comparable orientations between experiments, to permit visual (and numerical) comparison of FC changes over time. Also highlighted in this case is the very obvious "wall artefact" which has likely unknowingly skewed the FCSExpress results, but which can be easily highlighted in PolarFCS.

the current PolarFCS software lacks the capacity to incorporate successive gating strategies, a limitation also found with viSNE. However, as has been emphasised, such gating strategies can be time-consuming and error-prone, and may not be necessary in multi-parametric analyses. Also, PolarFCS has not yet been tested across the broader array of clinical scenarios mandating FC analysis. That said, in the limited settings in which we have employed it, as outlined herein, PolarFCS appears to provide a visually satisfying adjunct to current FCS solutions.

In conclusion, we proffer PolarFCS as a novel and simple FC data-analysis system that offers single-plot, multi-parametric FC data visualisation and analysis. PolarFCS has been compiled to be compatible with both Windows and MacOSX operating systems, which we hope will allow wide availability.

Gill et al. (2017) EMBnet.journal 23, e892
http://dx.doi.org/10.14806/ej.23.0.892

Key Points

- PolarFCS is an FCS program that allows easy single-plot, multi-parametric data visualisation, a feature that traditional FC data-analysis programs currently do not offer.
- We compare PolarFCS to a traditional FCS platform in the analysis of haematological malignancies, and show how the visualisation functionalities available in PolarFCS allow more accurate, standardised and detailed data analysis.
- PolarFCS has highly valuable characteristics, especially as most hospital laboratories receive thousands of specimens for FC each year, and traditional FC analysis is a time- and labour-intensive process.

Acknowledgements

The authors acknowledge the ongoing contributions of all staff and technologists of the "Flow Cytometry Laboratory of the Division of Hematology & Transfusion Medicine of the Calgary Lab Services".

References

1. Amir ED, Davis KL, Tadmor MD, Simonds EF, Levine JH *et al.* (2013) viSNE enables visualization of high dimensional single cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31**(6), 545-552. http://dx.doi.org/10.1038/nbt.2594
2. Ormerod MG (2008) Flow Cytometry – A Basic Introduction. 1st Ed. De Novo Software. http://flowbook.denovosoftware.com (accessed on January 31, 2017).
3. Qiu P, Simonds EF, Bendall SC, Gibbs KD, Bruggner RV *et al.* (2011) Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.* **29**, 886-891. http://dx.doi.org/10.1038/nbt.1991
4. Verbsky J and Routes JM (2017) Flow cytometry for the diagnosis of primary immunodeficiencies. In: UpToDate, Notarangelo LD (Ed), UpToDate, Waltham, MA. Updated February 2017: http://www.uptodate.com/contents/flow-cytometry-for-the-diagnosis-of-primary-immunodeficiencies (accessed on February 28, 2017).

EMBnet.journal
Bioinformatics in Action

Korfiati et al. (2017) EMBnet.journal 23, e882
http://dx.doi.org/10.14806/ej.23.0.882

Technical Notes

InSyBio ncRNASeq: a Web tool for analysing non-coding RNAs

Aigli Korfiati[1]✉, Konstantinos Theofilatos[1], Christos Alexakos[1], Seferina Mavroudi[1,2]

[1] InSyBio Ltd, United Kingdom

[2] UK & Department of Social Work, School of Sciences of Health and Care, Technological Educational Institute of Western Greece, Koukouli, Patra, Greece

Competing interests: AK, KT, CA: employer of InSyBio Ltd. InSyBio Ltd participates in the NBG Business Seeds program by the National Bank of Greece and financed this study. However, InSyBio and NBG Business Seeds program were not involved in the study design, collection, analysis & interpretation of data and writing of the paper. They only participated in the process of selecting the suitable journal for submitting the paper. SM: external collaborator of InSyBio Ltd and an employer in the Department of Social Work, School of Sciences of Health and Care, Technological Educational Institute of Western Greece. InSyBio Ltd participates in the NBG Business Seeds program by the National Bank of Greece and financed this study. However, InSyBio and NBG Business Seeds program were not involved in the study design, collection, analysis & interpretation of data and writing of the paper. They only participated in the process of selecting the suitable journal for submitting the paper.

Abstract

Non-coding RNA (ncRNA) genes encode non-protein-coding RNAs, which are classified into infrastructural and regulatory ncRNAs, depending on their role. Regulatory ncRNAs are involved in a variety of key cellular processes, and are hence associated with several diseases. For this reason, their accurate and efficient identification, and the identification of their targets, is a promising research area and an open topic for the bioinformatics community. We present InSyBio ncRNASeq, a cloud-based Web platform that assists users to analyse ncRNA sequences, and predict whether they are miRNAs, pseudo-hairpins or whether they belong to another ncRNA category. Additionally, InSyBio ncRNASeq offers a unique miRNA target-prediction pipeline, which results in scored miRNA target sites in 3'-untranslated regions (3'-UTR) of messenger RNAs (mRNAs). We show this tool to have the highest accuracy metrics compared with other state-of-the-art methods and tools.

Introduction

Traditionally, most RNA molecules were regarded as information-carrying intermediates from the gene to the translation machinery, with exceptions being transfer RNA (tRNA) and ribosomal RNA (rRNA), both of which are infrastructural RNAs that are central to the protein-synthesis process. However, since the late 1990s, evidence (Liu *et al.*, 2005) suggests that other types of non-protein-coding RNA molecules are present in organisms ranging from bacteria to mammals, and are involved in a variety of cellular processes, including plasmid replication, phage development, bacterial virulence, chromosome structure, DNA transcription, RNA processing and modification, development control, regulation of gene expression and others. In humans, it is estimated that about 98% of the genome can be transcribed, out of which only ~2% code for proteins, suggesting the possibility that a large percentage of the genome may encode ncRNAs. ncRNAs have been reported as biomarkers for disease and response to treatment in cancer, liver and cardiovascular diseases, and central nervous system disorders, among many others (Lopez *et al.*, 2015).

Although the importance of ncRNAs in cellular activities is well known, and new members and classes of ncRNA are continuously being reported, our current knowledge about the collection of all ncRNAs present in a particular genome is very limited because of the lack of effective computational or experimental methods. Indeed, their computational identification represents one of the most important and challenging issues in computational biology. Existing methods for ncRNA-gene prediction rely mostly on homology information, thus limiting their applications to ncRNA genes that have homologues.

Evidence linking ncRNAs with diseases seems to be attracting the attention of the pharmaceutical and biotechnology industries. Companies and institutions are developing ncRNA-based strategies against cancer, cardiovascular, neurological and muscular diseases. Thus, continued investigation of ncRNA biogenesis and function will provide a new framework for a more comprehensive understanding of human diseases.

InSyBio ncRNASeq enables users to analyse ncRNAs. Users can search and analyse both a specific RNA sequence of interest and a full-sequence data-set

Article history
Received: 11 December 2016
Accepted: 7 April 2017
Published: 09 October 2017

Korfiati et al. (2017) EMBnet.journal 23, e882
http://dx.doi.org/10.14806/ej.23.0.882

derived from available online databases or produced by experimental or computational procedures. With InSyBio ncRNASeq, users can predict and analyse ncRNAs and miRNA target genes. InSyBio ncRNASeq also allows result storage in its knowledge-base, equipped with information retrieval tools, to allow users to produce and extract their own data-sets.

Methodologies

RNA characterisation and miRNA prediction

InSyBio ncRNASeq enables users to characterise ncRNA sequences. A set of 58 informative features is calculated by supplying the sequence in FASTA format. The features include sequence, thermodynamic and structural properties of the RNA sequences; they were first introduced in (Kleftogiannis *et al.*, 2015), and are presented in Supplementary Table 1. Batch calculations of many sequences are allowed. This enables users to analyse data-sets derived from online databases, experimental sequencing or computational techniques. The results are presented in a browseable table in the Web interface, and can also be downloaded as a tab-delimited text file. For each ncRNA, its sequence and values of its 58 features are presented.

An additional functionality provided by InSyBio ncRNASeq is the prediction of pre-miRNAs, discriminating them from pseudo-hairpins and other molecules. The prediction of pre-miRNAs and pseudo-hairpins is accomplished through the application of a novel methodology, which combines a Genetic Algorithm (GA) (Holland, 1995) with epsilon-Support Vector Regression-SVR (Smola and Schölkopf, 2004) techniques. GAs, a state-of-the-art meta-heuristic optimisation technique, are used to optimise the feature subset to be used as input for the classification model and the regularisation parameters (C), the width of the Gaussian distribution (sigma) of the Radial Basis Function, and the epsilon threshold of e-SVR models. The accuracy of this technique in predicting pre-miRNAs is 95%. A sequence is predicted as 'other' if the minimum free energy is more than -15 kcal/mol, or the number of base pairs is less than 18. The supported input file comprises RNA sequences in FASTA format; batch calculations of many sequences are also allowed. For each ncRNA, the system provides the sequence, its calculated confidence score, the prediction whether it is an miRNA, a pseudo-hairpin or 'other', and its 58 features.

miRNAs and transcripts knowledge-base

The InSyBio ncRNASeq knowledge-base includes stem-loop and mature miRNAs. For each stem-loop, InSyBio ncRNASeq provides the following information: accession ID, name, species, length, description and comments (if any). For the sequences, the FASTA format is downloadable, and the sequence, its description and secondary structure (in dot-bracket notation) are presented. Visualisation of the secondary structure is performed with FornaContainer[1], which shows the Minimum Free Energy (MFE) structure. Stem-loops are linked to their corresponding mature miRNAs. For each mature miRNA, its accession ID, name and sequence are presented. The miRNA sequence is downloadable in FASTA format. Additional information about each mature miRNA is evidence, which may be experimental, similarity to another stem-loop structure, or found in the literature. References for the miRNA of interest are also available, as well as external links to other databases, such as MIRBASE (Griffiths-Jones *et al.*, 2006), ENTREZGENE (Maglott *et al.*, 2007), HGNC (Bruford *et al.*, 2008), RFAM (Griffiths-Jones *et al.*, 2005), and to publications.

The InSyBio ncRNASeq knowledge-base also includes genes and transcripts. For each gene, the information reported is the name, source, description, location, biotype, annotation method and version, as well as the list of its transcripts and links to each transcript. For each transcript, the information provided is the name, source, description, corresponding gene, corresponding protein, location, transcription start site, length, biotype, annotation method, version and 3'-UTR sequence. The 3'-UTR sequence is visualised with FornaContainer, and available for download in FASTA format.

miRNA target analysis and prediction

InSyBio ncRNASeq enables users to calculate 124 features for every miRNA and potential target-site pair within an mRNA. These features include sequence, thermodynamic and structural properties of the miRNA:mRNA pair; a detailed description of them can be found at (Korfiati *et al.*, 2015). Users can supply a file of miRNA sequences and a file of the respective mRNA binding-site sequences, both in FASTA format. Feature calculation of many miRNA:mRNA pairs is allowed in batch. The results are presented in a browseable table in the Web interface, and can also be downloaded as a tab-delimited text file. For each miRNA:mRNA pair, the miRNA sequence, the mRNA binding-site sequence and the 124 miRNA:mRNA pair features are provided. A description of the supported features for the characterisation of miRNA:mRNA pairs is available in Supplementary Table 2.

With InSyBio ncRNASeq, users can computationally validate miRNA target sites with an intelligent computational technique (hybrid combination of GAs and e-SVRs) and 124 informative features, based on the method presented in (Korfiati *et al.*, 2015), by supplying a file of miRNA sequences and a file of the respective mRNA binding-site sequences in FASTA format. The ncRNASeq tool then creates the miRNA:mRNA pairs by combining all the mRNA target sites from the first file with all the miRNAs from the second file. For each

[1]https://github.com/pkerpedjiev/fornac

Korfiati et al. (2017) EMBnet.journal 23, e882
http://dx.doi.org/10.14806/ej.23.0.882

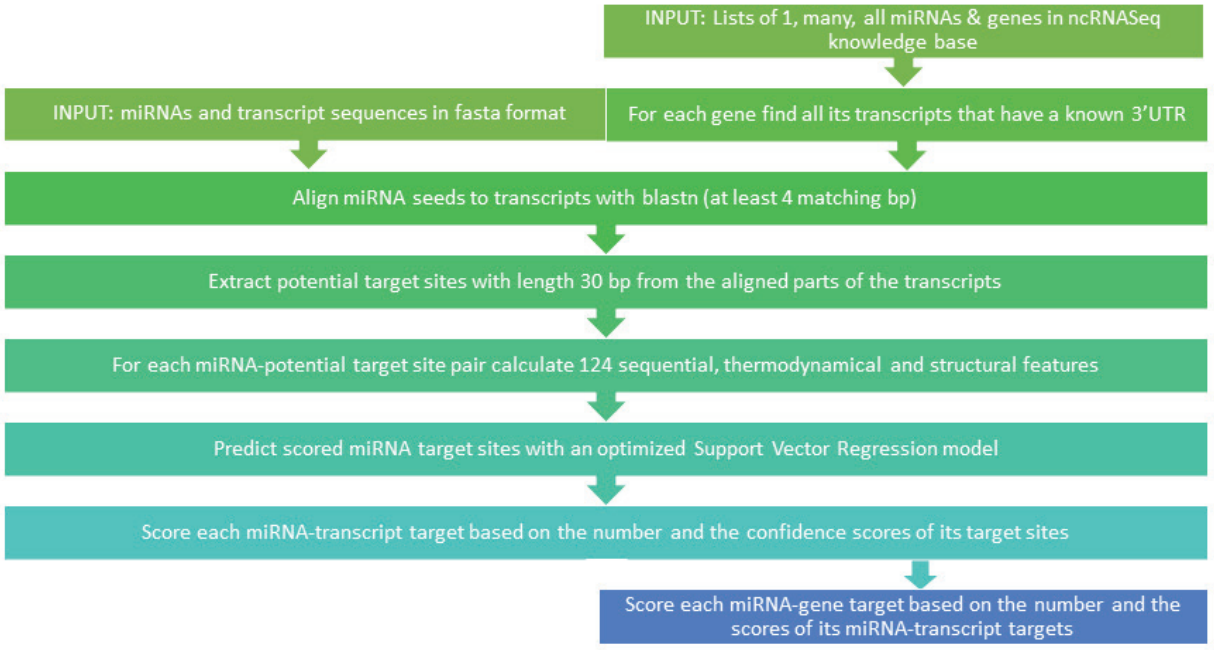


Figure 1. InSyBio ncRNASeq miRNA target-prediction workflow.

miRNA:mRNA pair, InSyBio ncRNASeq provides information about the miRNA sequence, the mRNA binding-site sequence, whether the miRNA:mRNA pair shares a targeting relation or not, the confidence score of the prediction, and the 124 miRNA:mRNA pair features grouped in categories (Thermodynamic, Positional, Motif and Structural). A link from the given miRNA to a protein in the InSyBio Interact database is also presented, indicating that the specific miRNA regulates the expression of the respective mRNA. InSyBio Interact contains information on protein functions, clusters, protein-protein interactions, etc.

Apart from miRNA target-site prediction, InSyBio ncRNASeq enables users to perform miRNA target prediction, either selecting genes or transcripts and miRNAs from the InSyBio ncRNASeq knowledge-base or providing their own miRNAs and transcript sequences in FASTA format. miRNA targets can be predicted in the 3'-UTR sequence of all transcripts of a gene for which a confidence score for the miRNA-gene interaction and additional scores for the miRNA-transcript interactions are provided. Figure 1 shows the miRNA target-prediction workflow.

Results and Conclusions

As a case study, InSyBio ncRNASeq was used to predict the miRNAs that target the 24 biomarkers reported in (Theofilatos *et al.*, 2016) for Parkinson's Disease (PD). For these biomarkers, InSyBio ncRNASeq yielded 18,359 interactions from 1,594 miRNAs (Table 1). miRTarBase (Chou *et al.*, 2016), the experimentally validated miRNA-target interaction database, contains 585 interactions for the same genes with 489 miRNAs. The results of InSyBio ncRNASeq overlap those of MiRTarBase in 190 interactions (32.5%). The miRNA target prediction tool miRTar (Hsu *et al.*, 2011), which utilises four of the most well-known target-prediction algorithms ((TargetScanS (Lewis *et al.*, 2005), miRanda (Bino *et al.*, 2004), RNAhybrid (Krüger and Rehmsmeier, 2006) and PITA (Kertesz *et al.*, 2007), match only 16 interactions (2.7%) of the experimentally validated interactions in miRTarBase. Figure 2 shows the interaction network among the highest-scoring interactions (10%) predicted by InSyBio ncRNASeq.

These results demonstrate that InSyBio ncRNASeq outperforms other tools in predicting miRNA targets. The predicted interactions are scored, which helps biologists to better design their experiments. Additional

Table 1. miRNA target-prediction results for 24 biomarkers of Parkinson's Disease obtained using InSyBio ncRNASeq and miRTar, compared with experimentally validated interactions in miRTarBase

	Interactions	miRNAs	Genes	Matching results in miRTarBase	Matching results in miRTarBase / miRTarBase total interactions	Matching results in miRTarBase / total predicted interactions
miRTarBase	585	489	24	-	-	-
miRTar	1,430	683	24	16	0.027	~0.011
ncRNASeq	18,359	1,594	24	190	0.325	~0.010

Korfiati et al. (2017) EMBnet.journal 23, e882
http://dx.doi.org/10.14806/ej.23.0.882

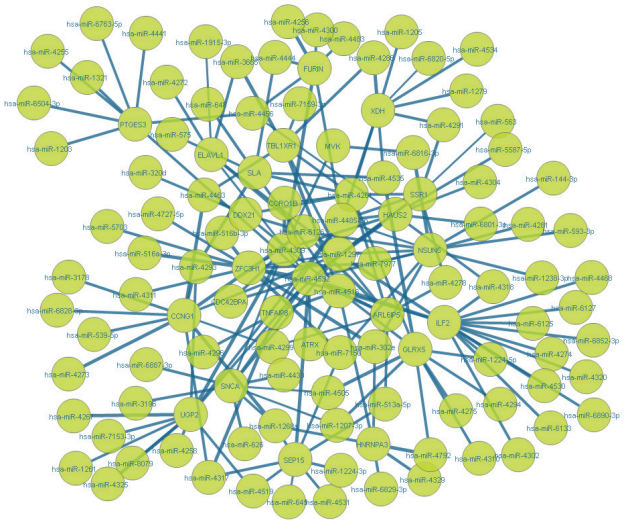


Figure 2. Interaction network of the highest-scoring miRNA target interactions (10%) predicted by InSyBio ncRNASeq for the 24 biomarkers of PD.

advantages of InSyBio ncRNASeq are that it predicts scored target sites in each mRNA, and computes sequence, thermodynamic and structural properties of the predicted miRNA:mRNA pair. Additionally, InSyBio ncRNASeq offers users the possibility to run batch searches with numerous miRNAs and mRNAs, with all the possible pairs being considered for potential miRNA:mRNA interactions scored according to the number of target sites and their confidence score. InSyBio ncRNASeq also enables users to interpret their results thanks to the integration of structural and functional information collected from external sources for miRNAs, genes and encoded proteins. Finally, the possibility offered to users for the upload of their own data-sets (transcript sequences) assists researchers searching for miRNA targets in altered transcripts, transcripts with mutations, etc.

Availability

InSyBio ncRNASeq is one of the tools included in the InSyBio Suite. A demo version of InSyBio ncRNASeq is freely available at http://demo.insybio.com. A free evaluation version includes a one-month free licence that can be obtained by emailing info@insybio.com. To purchase the commercial version of InSyBio ncRNASeq, users can contact sales@insybio.com for a detailed quote and information. InSyBio is registered with the Information Commissioner's Office (ICO) under registration reference number ZA182885 for data-processing issues.

Key Points

- InSyBio ncRNASeq is a cloud-based Web platform that assists users to analyse ncRNA sequences and predict pre-miRNAs, discriminating them from pseudo-hairpins and other molecules.
- With InSyBio ncRNASeq, users can also predict and analyse miRNA targets. It predicts scored target sites in each mRNA, and computes sequence, thermodynamic and structural properties of the predicted miRNA:mRNA pair.
- InSyBio ncRNASeq outperforms other tools in predicting miRNA targets. The predicted interactions are scored, which helps biologists to better design their experiments. It also enables users to interpret their results thanks to the integration of structural and functional information collected from external sources for miRNAs, genes and encoded proteins.
- The possibility offered to users for the upload of their own data-sets (transcript sequences) assists researchers searching for miRNA targets in altered transcripts, transcripts with mutations, etc.
- The InSyBio ncRNASeq knowledge-base includes stem-loop and mature miRNAs, genes and transcripts.

Acknowledgements

InSyBio participates in the National Bank of Greece (NBG) Business Seeds programme. The NBG does not retain any copyright for this work.

References

1. Bino J, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS (2004) Human microRNA targets. *PLoS Biol* **2**(11), e363. http://dx.doi.org/10.1371/journal.pbio.0020363
2. Bruford EA, Lush MJ, Wright MW, Sneddon TP, Povey S, Birney E (2008) The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res.* **36**(suppl 1), D445-D448. http://dx.doi.org/10.1093/nar/gkm881
3. Chou CH, Chang NW, Shrestha S, Hsu SD, Lin YL, *et al.* (2016). miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.* **44**(D1), D239-D247. http://dx.doi.org/10.1093/nar/gkv1258
4. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR and Bateman A (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research* **33**(suppl.1), D121-D124. http://dx.doi.org/10.1093/nar/gki081
5. Griffiths-Jones S, Grocock RJ, Van Dongen S, Bateman A and Enright AJ (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34**(suppl.1), D140-D144. http://dx.doi.org/10.1093/nar/gkj112
6. Holland J (1992) Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. Cambridge, MA, USA:MIT Press.
7. Hsu JBK, Chiu CM, Hsu SD, Huang WY, Chien CH *et al.* (2011) miRTar: an integrated system for identifying miRNA-target interactions in human. *BMC Bioinformatics* **12**, 300 http://dx.doi.org/10.1186/1471-2105-12-300
8. Kertesz M, Iovino N, Unnerstall U, Gaul U and Segal E (2007) The role of site accessibility in microRNA target recognition. *Nature Genetics* **39**(10), 1278-1284. http://dx.doi.org/10.1038/ng2135

9. Kleftogiannis D, Theofilatos K, Likothanassis S, Mavroudi S (2015) YamiPred: A novel evolutionary method for predicting pre-miRNAs and selecting relevant features. *IEEE/ACM Trans Comput Biol Bioinform.* **12**(5), 1183-92. http://dx.doi.org/10.1109/TCBB.2014.2388227
10. Korfiati A, Theofilatos K, Kleftogiannis D, Alexakos C, Likothanassis S and Mavroudi S (2015) Predicting human miRNA target genes using a novel computational intelligent framework. *Information Sciences* **294**, 576-585. http://dx.doi.org/10.1016/j.ins.2014.09.016
11. Krüger J and Rehmsmeier M (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Research* **34**(Suppl. 2), W451-W454. http://dx.doi.org/10.1093/nar/gkl243
12. Liu C, Bai B, Skogerbø G, Cai L, Deng W, Zhang Y *et al.* (2005) NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res.* **33**(suppl.1), D112-D115. http://dx.doi.org/10.1093/nar/gki041
13. Lewis BP, Burge CB and Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**(1), 15-20. http://dx.doi.org/10.1016/j.cell.2004.12.035
14. Lopez JP. Cruceanu C, Fiori LM, Laboissiere S, Guillet I *et al.* (2015) Biomarker discovery: quantification of microRNAs and other small non-coding RNAs using next generation sequencing. *BMC Med Genomics.* **8**, 35. http://dx.doi.org/10.1186/s12920-015-0109-x
15. Maglott D, Ostell J, Pruitt KD and Tatusova T (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research* **35** (suppl_1), D26-D31. https://doi.org/10.1093/nar/gkl993
16. Smola AJ and Schölkopf B (2004) A tutorial on support vector regression. *Statistics and Computing* **14**, 199. http://dx.doi.org/10.1023/B:STCO.0000035301.49549.88
17. Theofilatos K, Dimitrakopoulos C, Alexakos C, Korfiati A, Likothanassis S and Mavroudi S. (2016) InSyBio BioNets: A new tool for analyzing biological networks and its application to biomarker discovery, *EMBnet Journal* **22**, e871. http://dx.doi.org/10.14806/ej.22.0.871

EMBnet.journal
Bioinformatics in Action

Klingström et al. (2017) EMBnet.journal 23, e897
http://dx.doi.org/10.14806/ej.23.0.897

Galaksio, a user friendly workflow-centric front end for Galaxy

Tomas Klingström 1*✉, Rafael Hernández-de-Diego 1*, Théo Collard 1, Erik Bongcam-Rudloff 1

[1]SLU-Global Bioinformatics Centre, Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Sweden

*Both authors contributed equally (TK and RHD)

Competing interests: TK none; RHD none; TC none; EBR none

Abstract

There is a severe shortage of statisticians and bioinformaticians available in research. As universities fail to cover the increasing need of graduates with the necessary skills, ad hoc training and workshops have become commonplace but are insufficient to cover the needs. Technical solutions that distribute the workload more efficiently between researchers with a different education background (*e.g.*, computer scientists and biologists) are therefore necessary to cover some of this shortage.

Galaksio provides a workflow-centric graphical user interface for the Galaxy Workflow Management system that is easy to use for biologists and medical researchers who need to run routine tasks in bioinformatics. Combined with back end tools such as BioBlend, CloudMan and Pulsar, Galaksio provides a novel, layered approach to Galaxy making it easier to divide research tasks to researchers depending on their skills in interdisciplinary subjects such as bioinformatics and computational science.

Galaksio is developed by the B3Africa project for the eB3Kit but can easily be installed independently using docker and configured to provide access to workflows on any Galaxy server using the Galaxy API. Galaksio can be downloaded at: **https://github.com/fikipollo/galaksio.**

Introduction

Galaxy is a widely supported workflow management system used in bioinformatics (Goecks *et al.*, 2010; Leipzig, 2016; Tastan Bishop *et al.*, 2015; Atwood *et al.*, 2015) to facilitate accessible and reproducible research. One of the main aims of Galaxy is to provide access to bioinformatic analysis tools for experimentalists with limited expertise in programming (Atwood *et al.*, 2015; Blankenberg *et al.*, 2010). Nevertheless, our experience with Galaxy, gained by implementing it in the eBiokit (Hernández-de-Diego *et al.*, 2017) and by using Galaxy in several training and capacity building projects (Fuxelius *et al.*, 2010; Atwood *et al.*, 2015; Mulder *et al.*, 2016) has shown us that many potential Galaxy users find themselves in a bit of a conundrum when trying to use Galaxy. Researchers skilled enough in bioinformatics to install and configure tools prefer command line tools, whereas less advanced users are left on their own struggling to find and combine tools using the user interface provided by Galaxy. Therefore, many research groups remain reliant on in-house scripts maintained by a small number of bioinformaticians spending significant time on providing ad hoc support to other researchers in the group. To provide an attractive technology platform for researchers it was therefore deemed necessary to provide a more simplified, workflow-centric model of operations. In the workflow-centric model researchers with limited bioinformatics training are provided with prepared workflows and default input parameters, while more advanced users can create and modify workflows using the normal Galaxy GUI. This allows research teams to work in a more efficient way. Trained bioinformaticians can adapt and develop tools and then provide the finished workflows for routine analysis to lab researchers.

In standard Galaxy all users rely on the same GUI, despite significantly different education background and expertise. Trained bioinformaticians often rely on a set of skills dependent on education decisions taken by students several years ahead of enrolling at a university (Wightman and Hark, 2012) while other researchers may have little or no formal training. Given the complexities of training needs, influential stakeholders such as the US National Research Council has therefore concluded that bioinformatics research is likely to be carried out by two disparate groups of researchers: quantitative biologists, who work at the interface of mathematical/computer science and biology, and research biologists, who need familiarity with a range of mathematical and computational concepts without necessarily being an

Article history
Received: 18 May 2017
Accepted: 3 October 2017
Published: 9 November 2017

Klingström et al. (2017) EMBnet.journal 23, e897
http://dx.doi.org/10.14806/ej.23.0.897

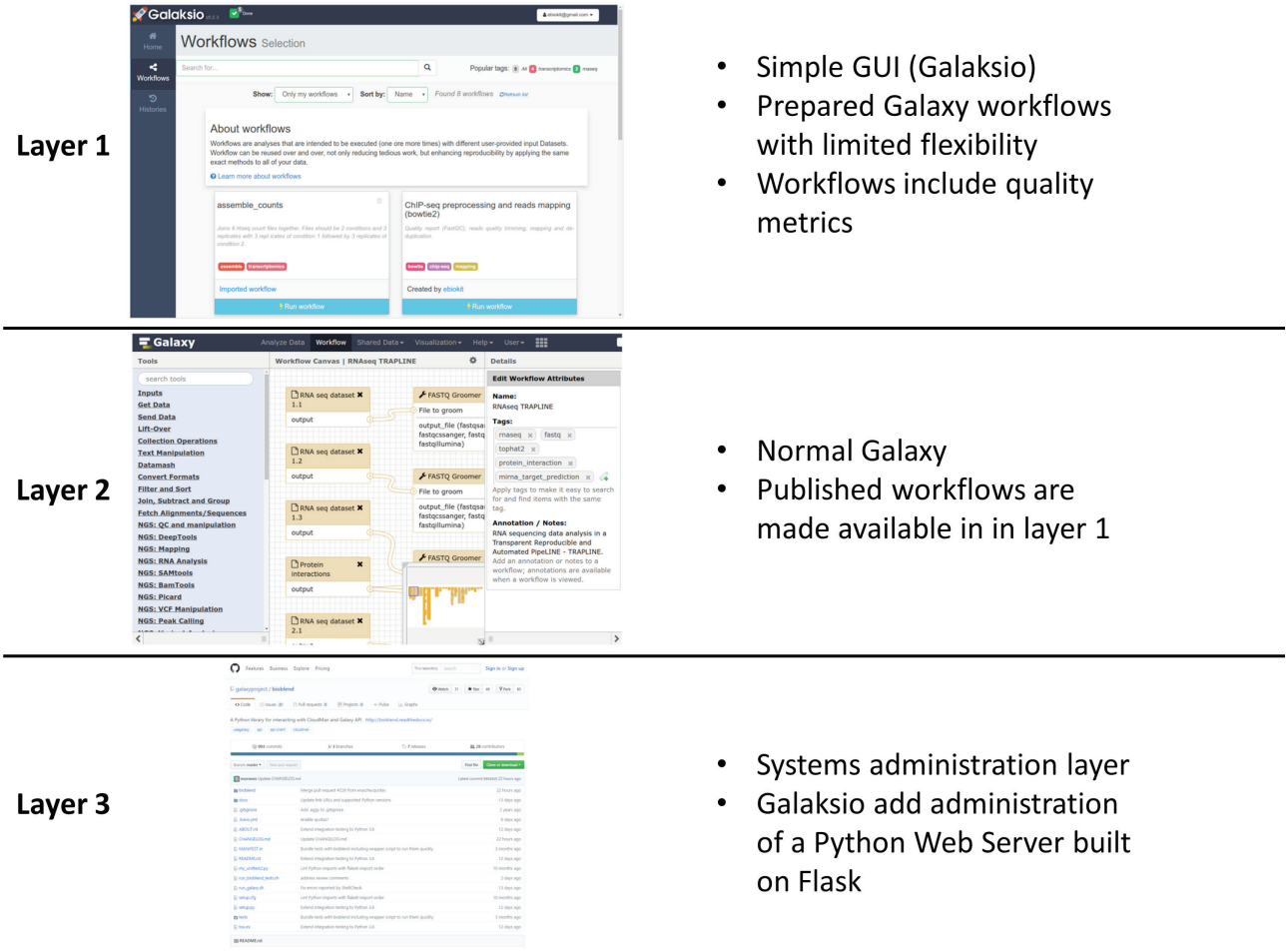


Figure 1. This figure shows the layered approach used by Galaksio and implemented in the eB3Kit to divide labour more efficiently between researchers with different background.

expert (National Research Council (US) Committee on Undergraduate Biology Education to Prepare Research Scientists for the 21st Century, 2003).

We therefore present Galaksio, a solution based on the Galaxy API and a Python web server, that we have developed to provide a layered access to Galaxy functions that facilitate the work of research biologists through an easy-to-use web interface, while the default Galaxy interface is used by bioinformaticians to create new workflows and systems administration tasks that are facilitated by packages created by other researchers such as BioBlend (Sloggett *et al.*, 2013), CloudMan (Afgan *et al.*, 2010) and Pulsar (Afgan *et al.*, 2015). With Galaksio, all data is managed within the normal Galaxy workflow management system and user credentials are passed on to the Galaxy server to manage user privileges, meaning that Galaksio can be used to access all workflows created on a normal Galaxy server using the command line tools implemented on the server.

Thanks to Galaksio, the Galaxy user's experience can be managed at three different levels: 1) a layer suited to research biologists (*i.e.*, users using tools); 2) a layer suited to bioinformaticians (*i.e.*, users developing tools); 3) a layer suited to computer scientists (*i.e.*, users developing the environment tools work in) (Figure 1).

This approach is currently being implemented in the B3Africa project using the eB3Kit which includes Galaksio and relies on these resources to connect the relatively light weight Mac Pro Server, commonly hosting the eB3Kit, to external computing resources (Klingstrom *et al.*, 2016).

Materials, Methodologies and Techniques

Galaksio has been designed as a multiuser web application and is divided in two components: the server side application and the web interface for users.

The server side, which is built on Python Flask server[1], is responsible for accessing the Galaxy data using the tools provided by the Galaxy application programming interface (API) (Blankenberg *et al.*, 2010; Goecks *et al.*, 2010). The Galaksio web interface has been developed using AngularJS[2] and Bootstrap[3], both popular HTML, CSS, and JavaScript cross-browser frameworks for developing responsive and user-friendly

[1]http://flask.pocoo.org/
[2]https://angular.io
[3]http://getbootstrap.com

Klingström et al. (2017) EMBnet.journal 23, e897
http://dx.doi.org/10.14806/ej.23.0.897

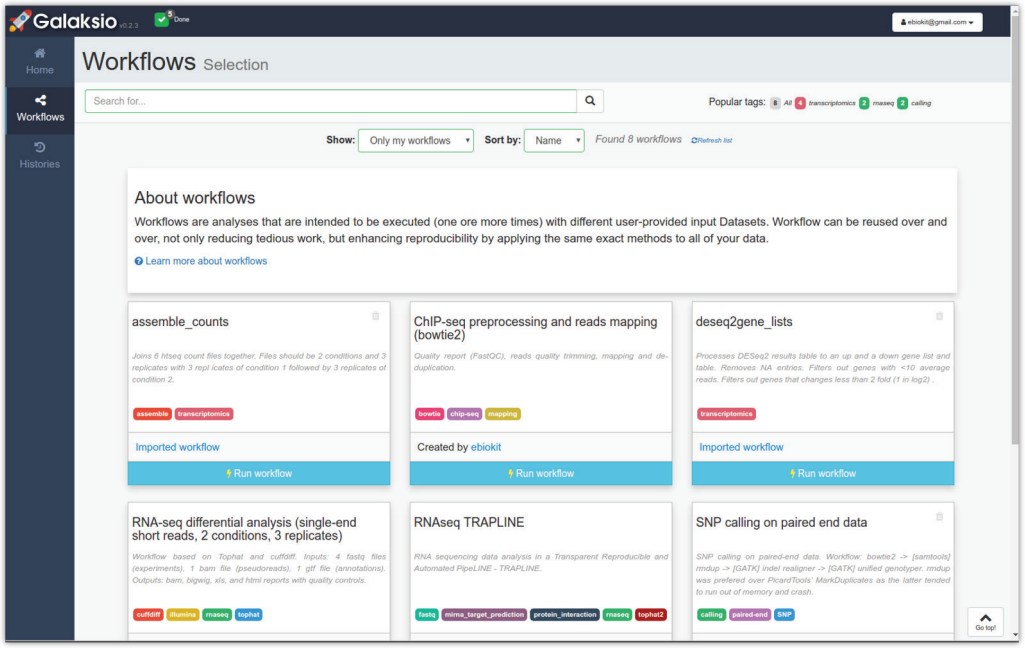


Figure 2. The figure shows the graphical interface for the workflow selection in Galaksio.

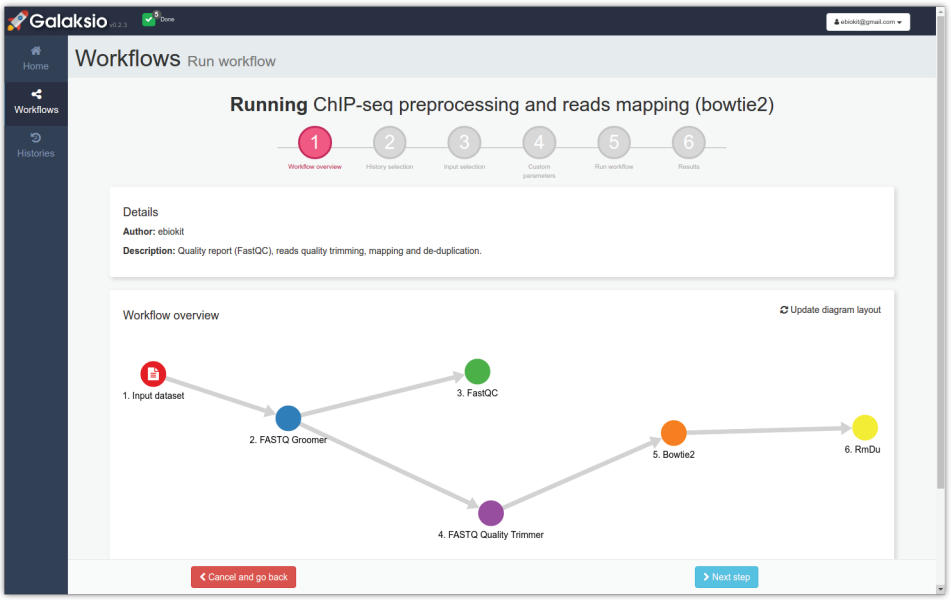


Figure 3. The figure shows the Galaksio web interface that is presented to the user after the selection of a workflow.

web applications. The exchange of the data between clients and the server is handled using asynchronous JavaScript and XML (AJAX) communication.

Results

Galaksio is free to use and is distributed under the GNU General Public License, Version 3. A public copy of the application is hosted at the SLU facilities as part of the eBioKit platform[4] and source code is available at GitHub[5], allowing other laboratories to browse, propose code reviews, and download the code in order to set up their own instance of the application. Additionally, Galaksio can easily be installed using Docker[6], an open-source virtualisation software that provides a lightweight, stand-alone, portable, and ready-to-execute package that includes the software and all the dependencies necessary to run the application independently of the operating system installed on the server. Documentation for the project can be found at the ReadTheDocs platform[7].

Figure 2 shows the Galaksio's GUI for biologists. Using this interface users can run any workflow implemented in the associated Galaxy instance in just

[4]http://ebiokit.eu/

[5]https://github.com/fikipollo/galaksio

[6]https://www.docker.com

[7]https://galaksio.readthedocs.io

Klingström et al. (2017) EMBnet.journal 23, e897
http://dx.doi.org/10.14806/ej.23.0.897

a few clicks and get a clear image of the analysis steps included in the selected workflow (Figure 3). The user interface allows the user to customise the execution of pre-selected tools, the uploading of the necessary files, the downloading of the results, and the execution of several workflows simultaneously in the background.

Table 1 provides an overview of all the developed features in the current Galaksio version. As all interactions with Galaxy are managed through the Galaxy API, the Galaksio implementation can be hosted independently as a separate server sending commands to any available Galaxy server. This includes public servers such as the popular usegalaxy.org website. Information on the connected server is provided when logging in via the Galaksio interface. It should however be noted that Galaksio, while light-weight in itself, is completely dependent on the speed of the Galaxy server when returning workflows and any user restrictions defined by the Galaxy server such as the amount of storage available.

Use case

Due to delays in achieving approval for tool wrappers created by the Galaksio team, an alternative use case has been created with much appreciated support from Marius van den Beek at the Institut Curie, Paris, France. The test dataset is available from the Zenodo data repository (Freeberg and Heydarian, 2016) but all data can also be imported from usegalaxy.org.

History containing dataset collections: https://usegalaxy.org/u/tomkl/h/galaksio-use-case-mouse-chip-seq-data.

Main workflow: https://usegalaxy.org/u/tomkl/w/copy-of-imported-parent-workflow-chipseq

Subworkflow: https://usegalaxy.org/u/tomkl/w/copy-of-imported-chipseqtutorialchild1

The workflows can be imported inside Galaksio by any users logged into a Galaksio server connected to usegalaxy.org. Other use cases will be added with the addition of "Galaksio use case" in the name of the workflow to make them easy to be identified in the Galaksio's repository. Issues are tracked using the Galaksio repository on GitHub[8] and external contributions are welcome.

Discussion

Compared to the clearly defined classes of "research biologist" and "quantitative biologist", proposed by the US National Research Council, bioinformatics has developed into a field where its practitioners share a number of characteristics, but none of which are essential enough to characterise what a bioinformatician truly is (Vincent and Charette, 2015). Many people may therefore be highly skilled and productive researchers in bioinformatics, despite very limited skills in one or more of the core competencies associated with being a bioinformatician (Smith, 2015). Due to the shortage

Table 1. Implemented and planned features for Galaksio.

Feature	Category	Implemented	Planned
User sign-in/out	Users	X	
User sign-up	Users	X	
Workflow listing	Workflows	X	
Workflow importing	Workflows	X	
Workflow execution	Workflows	X	
Workflow creation	Workflows		X
Simultaneous execution of workflows	Workflows	X	
Recovering previous executions	Workflows	X	
Help and description for tools in workflow	Workflows	X	
Input selection and parameter configuration	Workflows	X	
History selection	History	X	
History creation	History		X
History deletion	History		X
Dataset uploading	Dataset manipulation	X	
Dataset downloading	Dataset manipulation	X	
Dataset deletion	Dataset manipulation	X	
Dataset collection creation	Dataset manipulation	X	
Dataset collection deletion	Dataset manipulation		X
Tool execution	Tools		X

[8]https://github.com/fikipollo/galaksio/issues

Klingström et al. (2017) EMBnet.journal 23, e897
http://dx.doi.org/10.14806/ej.23.0.897

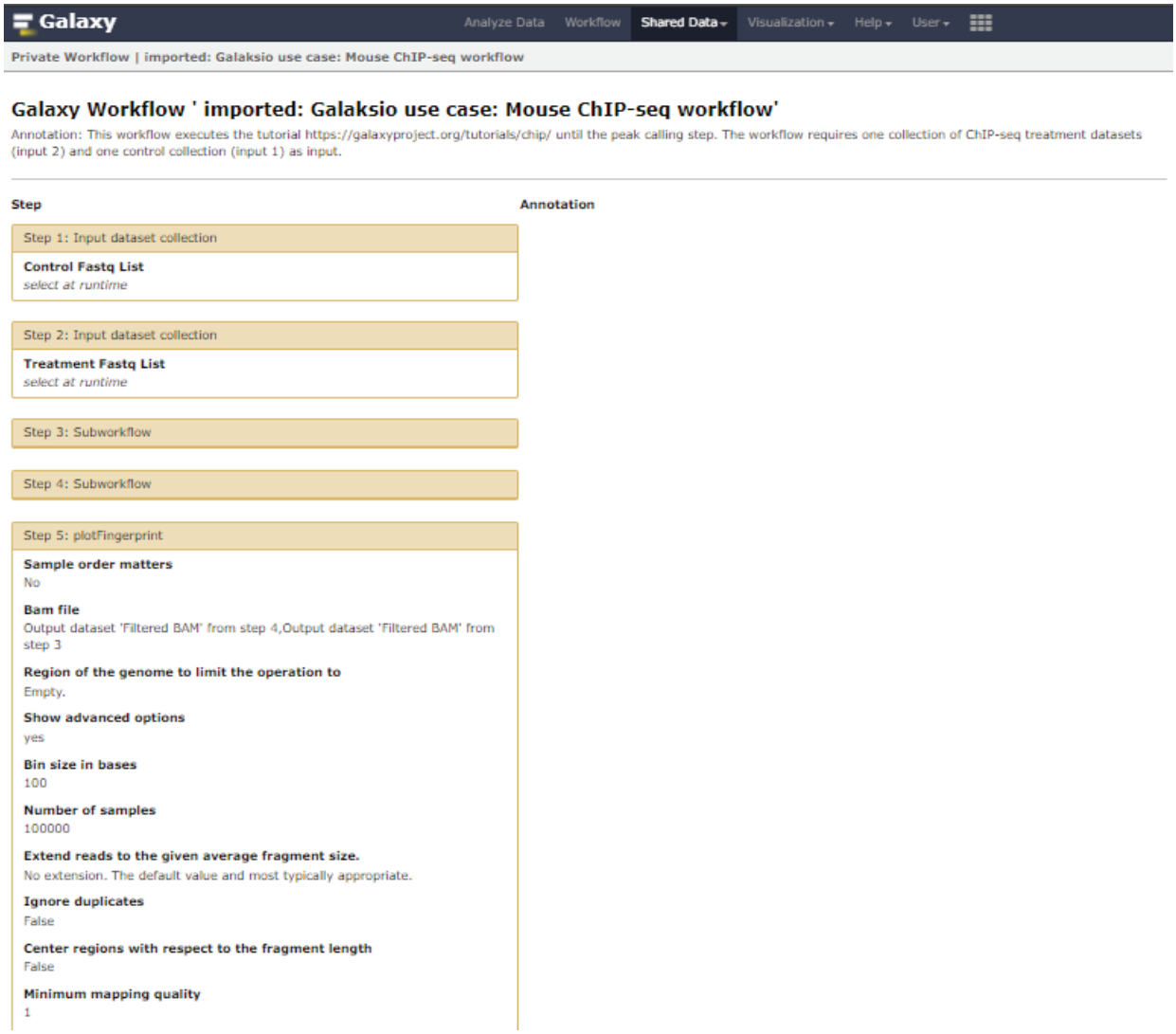


Figure 4. The figure displays a report generated by Galaxy by exporting a workflow after running a ChIP-seq use case.

of comprehensive university programmes in the field (Williams and Teal, 2017; Atwood *et al.*, 2015), most researchers currently active in bioinformatics have participated in a number of courses, workshops and self-learning sessions that, step by step, has taken them to a skill level where they may be considered qualified bioinformaticians or quantitative biologists. Such a self-organised curriculum encourages bioinformaticians to obtain exactly the skills necessary to complete their own projects but with limited consideration for auxiliary skills such as code documentation and a deeper understanding of computer science.

As a result of this self-motivated style of learning, significant delays occur when new technologies emerge if they require significant retraining of practitioners before becoming fully competitive with the new solution. This is perhaps most evident in the slow adoption of distributed computing systems such as Hadoop[9]. While significant investments in large Hadoop infrastructures has been made, the production of bioinformatics tools to use them has been delayed as bioinformatics tools are developed by bioinformaticians focused on high-level languages which, until recently, had limited support for Hadoop. Thereby delaying the adoption of distributed computing in bioinformatics (Oliphant, 2016).

The Galaksio interface itself is tailored towards enhancing user friendliness for biologists and medical researchers with limited IT-skills. The implementation of such a tool is a necessary step towards a multi-layered approach to Galaxy which allows distribution of labour not only between biologists and bioinformaticians, but also between "scripting" bioinformaticians and bioinformaticians with a strong background in computer science. Enabling researchers with the latter form of education background to provide access to more advanced computation tools by creating tools such as BioBlend (Sloggett *et al.*, 2013), CloudMan (Afgan *et al.*, 2010) and Pulsar (Afgan *et al.*, 2015) connect the Galaxy workflow management system to more powerful computation resources.

A common objection to user-friendly and automated systems such as Galaksio is the fear that automation can increase the error rate or can reduce the willingness of

[9]http://hadoop.apache.org/

researchers to learn bioinformatics properly. Automation is however one of the core concepts of advanced research ever since the introduction of the automated sequencing (Smith *et al.*, 1986). Indeed, without the automation of routine tasks even the sequencing and analysis of a single genome would be an impossible task (Ewing *et al.*, 1998). The relevance of automation within specific research tasks is perhaps best demonstrated by the common reliance on FASTQ files, with automatically assigned phred-quality scores, rather than the more expansive sequence read format (SRF) when working with large volumes of data (Clarke *et al.*, 2012; Van der Auwera *et al.*, 2013). With Galaksio automation is moved from a per-tool basis to a per-workflow basis and it is therefore appropriate to not only look at the risks that a further automation of tasks can bring, but also to evaluate how the current state of automation is facilitated in bioinformatics and other IT heavy fields. As an example, in healthcare the data management is seen as a way to reduce error rates and three key factors to success have been proposed for automation to be beneficial (Nolan, 2000):

- the system should prevent errors;
- procedures must be transparent so that they may be intercepted;
- procedures should be designed to mitigate the adverse effects of errors when they are not detected and intercepted.

Current practices in research are far from optimal when considering these three criteria for automation of bioinformatics. When dealing with bioinformatics tasks beyond their expertise, biologists may prefer commercial software that provides a more comprehensive, but also expensive platform with a dependency on proprietary software (Pabinger *et al.*, 2014; Smith, 2015b). As an alternative they may rely on outsourcing computing tasks to collaborators. Other biologists take the course of establishing their own curriculum of training as previously discussed. Some of these researchers may, over time, become proficient bioinformaticians but even in the best case scenario researchers are likely to produce a number of papers based on ad-hoc scripting with low transparency and potentially serious errors, unlikely to be caught by reviewers. In comparison, prepared workflows accessed in Galaxy or Galaksio limits the time spent on ad-hoc scripting and provide a comprehensive file history with source data and the individual steps used to generate the final results that greatly improve the reproducibility of the results (see Figure 4).

The downside of Galaksio is that it does not provide a natural exposure to the command line environment. However, Galaksio provides a comprehensive overview of any workflow available in the Galaxy system. If used properly Galaksio can therefore also serve as a training tool to explain theoretical concepts prior to coding exercises and function as a road map for researchers aiming to improve their skills in bioinformatics and build their own workflows step-by-step using the command line.

Conclusions

Galaksio does not replace the role of trained bioinformaticians in a research environment. It does however allow bioinformaticians to automate routine tasks and promote transparency in research as researchers with limited, or no, bioinformatics training can run best practice procedures and automatically generate the data necessary for others to evaluate their work. Such automation of routine tasks have contributed positively to the productivity and to the reduction of error rates in other information heavy fields (Horsfall, 1992; Leek and Peng, 2015; Nolan, 2000). Automation can thereby reduce the work load of expert bioinformaticians and provide them with the freedom to target more challenging tasks as well as to develop a curriculum for the evaluation and training of colleagues with basic or intermediate training (Peng, 2015).

Key Points

- Galaksio is built to provide a more layered approach to Galaxy, providing a simplified user interface based on workflows.
- Galaksio reduces the workload of bioinformaticians as routine tasks can be performed with minimal training. The presentation of workflows also provides a comprehensive overview of necessary input data as well as methodological changes to the end user.
- Galaksio can be used to rapidly deploy new services. Public Galaxy servers are a powerful tool to support collaborative research and Galaksio provides a more lightweight user interface for researchers who wish to make a specific project or workflow available.

Acknowledgements

This work was financed by the B3Africa project. B3Africa is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 654404.

We would also like to acknowledge the much appreciated help provided by Marius van den Beek at the Institut Curie, Paris, France and other active users in the Galaxy chat.

References

1. Afgan E, Baker D, Coraor N, Chapman B, Nekrutenko A, *et al.* (2010) Galaxy CloudMan: delivering cloud compute clusters. BMC Bioinformatics **11** (Suppl 12), S4. http://dx.doi.org/10.1186/1471-2105-11-S12-S4
2. Afgan E, Coraor N, Chilton J, Baker D, Taylor J, *et al.* (2015) Enabling cloud bursting for life sciences within Galaxy: Enabling Cloud Bursting for Life Sciences within Galaxy. Concurr. Comput.

Pract. Exp. **27** (16), 4330–4343. http://dx.doi.org/10.1002/cpe.3536
3. Atwood TK, Bongcam-Rudloff E, Brazas ME, Corpas M, Gaudet P, *et al.* (2015) GOBLET: The Global Organisation for Bioinformatics Learning, Education and Training. PLOS Comput. Biol. **11** (4), e1004143. http://dx.doi.org/10.1371/journal.pcbi.1004143
4. Blankenberg D, Kuster GV, Coraor N, Ananda G, Lazarus R, *et al.* (2010) Galaxy: A Web-Based Genome Analysis Tool for Experimentalists. Current Protocols in Molecular Biology. John Wiley & Sons, Inc., Hoboken, NJ, USA, Hoboken, NJ, USA,
5. Freeberg M and Heydarian M (2016) Training Material For Chip-Seq Analysis. http://dx.doi.org/10.5281/zenodo.197100
6. Fuxelius H, Bongcam E, and Jaufeerally Y (2010) The contribution of the eBioKit to Bioinformatics Education in Southern Africa. EMBnet.journal **16** (1), 29. http://dx.doi.org/10.14806/ej.16.1.173
7. Goecks J, Nekrutenko A, Taylor J, and Galaxy Team T (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. **11** (8), R86. http://dx.doi.org/10.1186/gb-2010-11-8-r86
8. Hernández-de-Diego R, de Villiers EP, Klingström T, Gourlé H, Conesa A, *et al.* (2017) The eBioKit, a stand-alone educational platform for bioinformatics. PLOS Comput. Biol. **13** (9), e1005616. http://dx.doi.org/10.1371/journal.pcbi.1005616
9. Horsfall K (1992) The human impact of library automation University of South Australia Library,.
10. Klingstrom T, Mendy M, Meunier D, Berger A, Reichel J, *et al.* (2016) Supporting the development of biobanks in low and medium income countries. IEEE, pp. 1–10
11. Leek JT and Peng RD (2015) Opinion: Reproducible research can still be wrong: Adopting a prevention approach: Fig. 1. Proc. Natl. Acad. Sci. **112** (6), 1645–1646. http://dx.doi.org/10.1073/pnas.1421412111
12. Leipzig J (2016) A review of bioinformatic pipeline frameworks. Brief. Bioinform. **18** (3), 530–536. http://dx.doi.org/10.1093/bib/bbw020
13. Mulder NJ, Adebiyi E, Alami R, Benkahla A, Brandful J, *et al.* (2016) H3ABioNet, a sustainable pan-African bioinformatics network for human heredity and health in Africa. Genome Res. **26** (2), 271–277. http://dx.doi.org/10.1101/gr.196295.115
14. National Research Council (US) Committee on Undergraduate Biology Education to Prepare Research Scientists for the 21st Century (2003) Bio2010: Transforming Undergraduate Education for Future Research Biologists National Academies Press (US), Washington (DC),.
15. Nolan TW (2000) System changes to improve patient safety. BMJ **320** (7237), 771–773.
16. Oliphant T (2016) Anaconda and Hadoop --- a story of the journey and where we are now. http://technicaldiscovery.blogspot.se/2016/03/anaconda-and-hadoop-story-of-journey.html (accessed 7 April 2017).
17. Peng R (2015) The reproducibility crisis in science: A statistical counterattack. Significance **12** (3), 30–32. http://dx.doi.org/10.1111/j.1740-9713.2015.00827.x
18. Sloggett C, Goonasekera N, and Afgan E (2013) BioBlend: automating pipeline analyses within Galaxy and CloudMan. Bioinformatics **29** (13), 1685–1686. http://dx.doi.org/10.1093/bioinformatics/btt199
19. Smith DR (2015) Broadening the definition of a bioinformatician. Front. Genet. **6**, 258. http://dx.doi.org/10.3389/fgene.2015.00258
20. Tastan Bishop O, Adebiyi EF, Alzohairy AM, Everett D, Ghedira K, *et al.* (2015) Bioinformatics Education--Perspectives and Challenges out of Africa. Brief. Bioinform. **16** (2), 355–364. http://dx.doi.org/10.1093/bib/bbu022
21. Vincent AT and Charette SJ (2015) Who qualifies to be a bioinformatician? Front. Genet. **6**, 164. http://dx.doi.org/10.3389/fgene.2015.00164
22. Wightman B and Hark AT (2012) Integration of bioinformatics into an undergraduate biology curriculum and the impact on development of mathematical skills. Biochem. Mol. Biol. Educ. **40** (5), 310–319. http://dx.doi.org/10.1002/bmb.20637
23. Williams JJ and Teal TK (2017) A vision for collaborative training infrastructure for bioinformatics: Training infrastructure for bioinformatics. Ann. N. Y. Acad. Sci. **1387** (1), 54–60. http://dx.doi.org/10.1111/nyas.13207

EMBnet.journal
Bioinformatics in Action

Rafael et al. (2018) EMBnet.journal 23, e903
http://dx.doi.org/10.14806/ej.23.0.903

Technical Notes

Establishment of "The South African Bioinformatics Student Council" and Activity Highlights

Candice Nancy Rafael[1]✉, Jon Ambler[2], Antoinette Niehaus[3], James Ross[4], Ozlem Tastan Bishop[1]

[1]Research Unit in Bioinformatics (RUBi), Department of Biochemistry and Microbiology, Rhodes University, South Africa

[2]Computational Biology Group, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, South Africa

[3]Centre of Excellence for Biomedical Tuberculosis Research/SA MRC Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, South Africa

[4]Faculty of Natural and Agricultural Sciences, Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, South Africa

Competing interests: CNR none; JA none; AN none; JR none; OTB none

Abstract

The South African Bioinformatics Student Council (SASBiSC) in bioinformatics has been set up to increase the visibility of bioinformatics as well as to filter information to students within the field regarding job, funding and workshop opportunities as they arise. This is a short description of the process of setting up a national Student Council for Bioinformatics in South Africa, affiliating to the International Society for Computational Biology (ISCB). We also report two examples of activities that were carried out over the last two years that are: 1) participation in the SciFest Africa; and 2) the organisation of the first Bioinformatics Student Symposium. We hope that our experience and methods for the creation of SASBiSC and of collaborative communities can be useful to others who might want to do the same.

Establishment of the Student Council

The South African Society for Bioinformatics[1] (SASBi) was officially formed in September 2012 during a joint Congress with the South African Genetics Society (SAGS). Prior to this there was no official body to represent bioinformatic researchers and students in the country. The establishment of SASBi also led to the establishment of the Student Society as a platform for students to meet and discuss their research activities, but also to socialise and broaden their network of knowledge and friendships. A small group of students joined as volunteers to pioneer and set up a SASBi Student Council (SASBiSC). As a first step, one representative, selected from the attendees present at the first Joint Congress of SASBi and SAGS, was elected to the main SASBi Council.

The SASBiSC was then established during the biennial SASBi-SAGS Congress in September 2014, in Pretoria, with the election of the President (Candice Rafael), the Secretary (Antoinette Colic), the Media Officer (Jon Ambler), the Development Officer (James Ross), and of a Faculty Advisor/Mentor (Prof. Özlem Tastan Bishop; first SASBi President). Students elected were voted onto the SASBiSC on the basis of their willingness to promote bioinformatics within the country. The geographical location of the students was also taken into account in a bid to get a fair distribution and representation of SASBiSC within the various institutions belonging to different countries present at the congress. The following provinces were represented within the SASBiSC: Gauteng, the Eastern Cape and the Western Cape.

The SASBiSC was modelled using the framework of the International Society for Computational Biology Student Council (ISCBSC) because one of our main goals was just the affiliation to ISCB and to become a Regional Student Group (RSG) (Macintyre *et. al*; 2013; Shanmugam and Macintyre; 2014). For a Student Council to be recognised as an ISCBSC, RSG requires the submission of various documents. First of all the composition of the SASBiSC committee, that would become the RSG committee, including the Chairperson, the Secretary and the Faculty Advisor/Mentor; the missions, objectives and goals of the RGS; the activity plan for the first year and the Constitution to govern and run the Council. The application has to be submitted to the ISCBSC for review; if the review is positive the

[1]http://sasbi.weebly.com/

Article history
Received: 18 September 2017
Accepted: 08 January 2018
Published: 05 February 2018

Rafael et al. (2018) EMBnet.journal 23, e903
http://dx.doi.org/10.14806/ej.23.0.903

application is submitted to the main ISCB Council for approval. The SASBiSC was officially recognised as RSG for South Africa in 2015. Two conditions are mandatory: the ISCB membership of at least two committee members and of the Faculty Advisor, and the creation and adoption of a SC's Constitution[2].

Activities of the Student Council

The initial goal of the SASBiSC was to improve the communication between South African students interested in bioinformatics (Figure 1). This has been done through various communication tools such as the SASBiSC's website[3], Facebook[4] and Twitter[5] accounts and a mailing list. These platforms have also been used for to give students the opportunity to make education and training experiences in other Institutes throughout funding and job opportunities circulated by the main SASBi Council as well as any other dissemination channel. The mailing list has been maintained and managed by the Secretary. The website and social media accounts have been managed mainly by the President and the Media Officer. However, all committee members have the possibility to access these media to increase the richness and visibility of the notices sent out. The SASBiSC has undertaken several projects, two of which are detailed below.

SCIFEST, AFRICA

SASBiSC was present at the national Science Expo (Scifest Africa) in Grahamstown in 2015 and 2016 with an exhibition desk. Scifest Africa, established in 1996, brings all levels of school pupils and teachers together with the aim of promoting public awareness, understanding of science, technology and innovation. This event is attended by thousands of learners from various schools around the country, looking for inspiration and potential career options, as well as representatives from various scientific organisations. The SASBiSC's aim at the event was to engage with the community for to increase awareness about bioinformatics as a research field and as a career choice for learners and teachers. Additionally, we aimed to interact with members of other scientific communities to grow relationships and encourage interdisciplinary communication and potential collaboration. These aims are in line with the directives outlined in the SASBiSC Constitution, as well as that of the ISCB Student Council.

Community Engagement

The SASBiSC exhibition desk included introductory movies about Bioinformatics and what it entails for the life sciences research, in several different languages: these included English, isiZulu and isiXhosa[6]. With 68,000 visitors to the event in 2015 and 56,000 in 2016, we were able to engage with a large number of learners and teachers from many different areas and schools. Many of the visitors we talked to were very interested in learning about Bioinformatics, as it is a discipline that few of them had ever heard of before. Also educators from local secondary school were interested to learn about Bioinformatics, as most were not familiar with the field, and were excited about this new possible avenue they could bring to their students exploring different career opportunities.

Media Coverage and Engagement with other members of the scientific community

In addition to direct community engagement, SASBiSC was able to discuss and promote Bioinformatics on various other media platforms, including interviews and radio 'blurbs' on WITS Radio Academy, the local Grahamstown radio station, and a snippet in a local newspaper.

Further, the SC met and established relationships with representatives from the Square Kilometre Array[7] (SKA), SA Innovation summit, Iziko, International Astronomical Union, and Southern African Association of Science and Technology Centres[8] (SAASTEC) members, involved in the Science Centres. Many people expressed an interest to collaborate in future events for the promotion of Bioinformatics and the use of bioinformatic techniques in their own fields.

FIRST STUDENT SYMPOSIUM

The student symposium was a one-day meeting held alongside the joint Congress of SASBi and SAGS[9] in Durban, South Africa, on the 20th September 2016. This was the first meeting of this type, and is planned to become a biennial or annual event. The main aim of this first symposium was to introduce students to bioinformatics research and job opportunities. The 35 student attendees at the symposium, mainly Honours and PhD students, were from bioinformatics and computational biology research groups in various fields and specialisations. For many of them this was the first time that they attended a conference. Nevertheless, they were able to interact with peers outside their own research labs. The symposium was funded with the support of generous funders including: the Bioinformatics Support Network[10] (BSP), The Scientific Group[11], International Society for Computational Biology (ISCB) Student

[2]http://sasbi.weebly.com/documents.html
[3]http://sasbistudents.weebly.com/
[4]https://www.facebook.com/SASBi-Students-446089572128356/
[5]https://twitter.com/SASBI_students
[6]http://bit.ly/2kK521O
[7]https://skatelescope.org/
[8]https://saastec.co.za
[9]http://sasbi.weebly.com/congress.html
[10]http://bio.chpc.ac.za
[11]http://www.scientificgroup.com

Rafael et al. (2018) EMBnet.journal 23, e903
http://dx.doi.org/10.14806/ej.23.0.903

Figure 1. Student's Symposium group photo.

Council[12], The South African Society for Bioinformatics[13] (SASBi) and a small fee from the participating students. The organisation was handled by the SASBiSC with the help of the organisers of the main conference, Anita Williams[14], Nicolette Crozier and Robyn Jacob.

Keynotes

The first keynote speaker was Prof. Fourie Joubert from the University of Pretoria who was the second President of SASBi from 2014 to 2016. The second keynote speaker was Prof. Nicola Mulder from the University of Cape Town. They presented and discussed their current research activities and future plans, as well as some works that had inspired them in the field. Prof. Nicola Mulder ended the day with a brief presentation on potential job opportunities in academia and in industry, which was an inspiring way to wrap up the day's talks.

Student Presentations

In total, five students were selected for an oral presentation, whereas others were invited to present their work as posters. Among students selected for an oral presentation, some were visiting students from Germany. It was a privilege for us to invite them to present their research activities at the SC Symposium. Indeed, this was a great opportunity for the country's students to learn about the research carried out outside their own labs, but also to learn about the scope of abroad research.

Award Winner

Each student's presentation was voted by attendees on the basis of different evaluation criteria on a marking rubric (Figure 2). Tobias Luttermann's talk was marked as the winning presentation and received a monetary prize for his excellent delivery and explanation of the research he had made and was hoping to do. The prize money was the remainder of the money secured from sponsors as well as the contribution from attending students once all the costs were covered.

Activities

The SASBiSC wanted to make the event more inclusive and not focused solely on research presentations, as the rest of the conference would centre on talks. We therefore incorporated a few smaller activities including an ice breaker, a mini hackathon, the Annual General Meeting (AGM) and a closing social event.

Icebreaker

Following the first keynote lecture and the first student presentations, a short icebreaker was held, just before the coffee break. This allowed students to feel more

[12]http://www.iscbsc.org

[13]http://sasbi.weebly.com

[14]http://www.conferencesandevents.co.za

Rafael et al. (2018) EMBnet.journal 23, e903
http://dx.doi.org/10.14806/ej.23.0.903

SASBi Student Council Symposium 20 September 2016

Presenter Evaluation Marking Sheet

Evaluator: _____

Name	1	2	3	4	5	6	7	Total
Sebastian Spaenig								
Patrick Blumenkamp								
Gugu P. Mahlangu								
Arnold Amusengeri								
Tobias Luttermann								

Criterion	Weight
1. Concise, accurate & up-to-date literature review	15
2. Knowledge gap and/or problem clearly identified and stated	15
3. Clear research hypothesis & objectives; Concise description of approach and methods	15
4. Results and discussion: interpretation of results and critical analysis of their meaning and impact	25
5. Summary of findings and future plans	5
6. Time management, visual media and speaker – audience contact	10
7. Ability of speaker to answer questions in a clear & meaningful manner.	15

Figure 2. Marking rubric used by each attendee to score oral presentations.

comfortable with each other, encouraging them to continue conversations during coffee and other breaks throughout the day. The icebreaker was modelled after others used before at ISCBSC Symposiums and termed "scientific speed-dating" (Grynberg *et al.*, 2011). This consisted of approximately 3-minute slots, during which students moved around the room and interacted with different students, learning about their research, areas of interest etc.

Mini Hack-a-thon

After lunch Jon Ambler presented the concept of the "hackathon". Owing to the limited amount of time he was only able to introduce the idea and process and hold a very short interactive time, hence it becoming a "mini" hackathon. The main concept behind the hackathon is centered around collaborative research in short sprints towards a common goal. Groups of around five attendees were formed within the venue, with students using their own personal laptops and collaborating using a Trello board setup[15]. Each member within the group worked on a specific aspect such as reading the data in, converting the data, visualising the test set and reporting back to the main group. The attendees responded very well to the concept and many were excited to return to their respective labs to try out the concept. Many had to be pulled away for the coffee break as they wanted to continue longer working on the test project. Overall, it was very successful and will hopefully lead to many collaborations in the future.

Annual General Meeting

Following the final keynote lecture, we moved straight on to the AGM with all the students present as they would form the main body the SASBiSC would ultimately serve. This was the first AGM since the SASBiSC was officially formed in Pretoria on September 2014. Here the SASBiSC reported on the activities it had undertaken during the past two years, including becoming an official RSG under the ISCBSC, hosting a stall at a local science festival, developing an online and social media presence, holding a film competition and finally organising the student symposium. Attendees gave ideas and suggestions for the future years of the SASBiSC which were reported by the Secretary for consideration and implementation where necessary by the new incoming Committee. Finally, the new Council was voted in and the Symposium officially closed with thanks to the sponsors, keynotes, organisers and volunteers, and attendees.

Closing Social Event

A closing social event was generously sponsored by The Scientific Group. This was hosted at The Market Outdoor Restaurant and gave the students a chance to interact in a relaxed, social environment, and network with student attendees from around the country and from abroad. It was a great end to a successful day that was the result of many months of work.

[15]https://trello.com/

Difficulties and challenges experienced

The main difficulties faced by the SASBiSC were communication and funders. Because Bioinformatics is still an emerging field in South Africa, many students and researchers never heard about it yet. This leads people to be reluctant in to get involved with events surrounding it. On the one other hand, funders like to see a track record and history of previous successful events which is difficult to show for a new organisation like SASBiSC that is just trying to establish itself. We overcame this difficulty by showing the capabilities of the Committee and the work it had made outside of the SASBiSC before its establishment. A well set up and a clear proposal for the event including the budget helped us to convince sponsors to fund our initiatives. Hosting a successful SciFest event and sending follow up reports including registers and photographs helped secure funders for the event.

Many students and educators admitted they had never heard about Bioinformatics and this made the recruitment of new students in the SC even more difficult. Communication between research groups within the country is limited and therefore the distribution of information is often difficult. Many announcements for funding opportunities, academic or job positions and workshops often do not reach the wider community that would benefit from such notices. The creation of the SASBiSC web portal and social media platforms will hopefully help with this in the future. However, there is still the challenge of acquiring the notices and calls to place on these sites.

Conclusions

A substantial amount of work was accomplished in these two years of the SASBiSC's existence. The framework is now in place, and on it future activities will be built. The activities of the SC will be continued by the new committee and perhaps enlarged by introducing new events and activities. The first Symposium was new to many students who had not attended any similar event before and therefore did not know what to expect. Through advertising and repeated posting via SASBi and University mailing lists as well as social media announcements, we were able to reach many students and to build a database of contacts in the field within the country. Suggested future projects are the organisation and hosting of workshops focused on bioinformatics techniques; as well as collaborations with larger bodies including BSP and the Centre for High Performance Computing[16] (CHPC) in order to bring forward to them needs in terms of training. In future events or symposiums, it could be beneficial to team with fellow RSGs within the African continent to host a collaborative international symposium. The members of the SC gained experience during their time on the Council through the various tasks and events undertaken. The members got to experience aspects involved in; large scale event organisation, communication (not only with fellow academics but also with the local and national community), marketing and proposal writing in the attempt to acquire funders and sponsors, and finally in the form of community building and networking for future collaborative work. We would also like to encourage the establishment of similar student groups that have the benefit of fostering skills that are otherwise often considered vocational, including interpersonal skills, communication, marketing, project management and community building. Within the academic environment these skills are often under-recognised, and yet can be exceptionally beneficial for effective and efficient establishment of bioinformatics.

Key Points

- Establishment of the first Student Council for Bioinformatics within South Africa.
- Affiliation of the South African Society for Bioinformatics Student Council (SASBiSC) to the International Society for Computational Biology (ISCB).
- Organisation of the first Student Symposium for Bioinformatics and Computational Biology.
- Presentation of bioinformatics as a field of study and as possible career avenue to young learners and educators.
- Emphasise the benefits of student run and motivated organisation within a scientific field.

[16]https://www.chpc.ac.za

Acknowledgements

The Student Council would like to acknowledge the following bodies for funding our various initiatives: Bioinformatics Service Platform (BSP), The Scientific Group, International Society for Computational Biology (ISCB) Student Council, South African Society for Bioinformatics (SASBi). Our acknowledgments also go to Anita Williams, Nicolette Crozier and Robyn Jacob for all their help in the organisation of the Symposium.

We would also like to thank the ISCB SC for the help and advice leading up to the events, setting up the RSG and advice regarding general Council running queries. Thanks also go to SASBi for providing a travel fellowship to assist the Council in attending the Symposium.

References

1. Grynberg P, Abeel T, Lopes P, Macintyre G, Pantano Rubiño L. (2011) Highlights from the Student Council Symposium 2011 at the International Conference on Intelligent Systems for Molecular Biology and European Conference on Computational Biology. BMC Bioinformatics **12**: A1.
2. Macintyre G, Michaut M, Abeel T (2013) The Regional Student Group Program of the ISCB Student Council: Stories from the Road. PLoS Comput Biol **9**(9): e1003241. https://dx.doi.org/10.1371/journal.pcbi.1003241
3. Shanmugam A, Macintyre G (2014) Establishing and Managing a Global Student Network. PLoS Comput Biol **10**(10): e1003920. https://dx.doi.org/10.1371/journal.pcbi.1003920

EMBnet.journal
Bioinformatics in Action

Pfeil et al. (2018) EMBnet.journal 23, e904
http://dx.doi.org/10.14806/ej.23.0.904

Mobile microscopy for the examination of blood samples

Juliane Pfeil[1], **Marcus Frohme**[1]✉, **Katja Schulze**[2]

[1]Division Molecular Biotechnology and Functional Genomics, Technical University of Applied Sciences, Wildau, Germany

[2]CSO and Computer Vision Expert, Oculyze GmbH, Wildau, Germany

Competing interests: JP none; MF none; KS The submitted work was performed in cooperation with the company Oculyze GmbH, Wildau Germany. Oculyze works in the field of mobile microscopy and computer vision. Therefore, Oculyze has a vested interest in the success of this field.

Abstract

The analysis of blood is one of the best possibilities to diagnose and control diseases and deficiency symptoms. Common blood tests that are performed in medical laboratories are time-consuming and work-intensive. In under-developed areas, there is often also a lack of specialised staff and facilities. The development of a mobile microscopic system that contains an automated image analysis and that can be used via a smartphone, could represent a valuable help to improve the diagnostic care, especially in those areas. It aims to enable a very fast, cheap, location- and knowledge-independent application for many use cases.

Introduction

Microscopy was already developed in the late 16th century and was initially more for wealthy peoples leisure than science. In modern days, it is still one of the most important tools in diagnostics and process monitoring. In most cases, a microscopic analysis requires expensive devices and trained specialists. In addition, investigations can often only be carried out in laboratories and they are very complex. The documentation is done via computer-assisted camera-systems, but automatic image analysis is rarely present.

As "Molecular Biotechnology and Functional Genomics" research group on automated microscopy and image analysis, we developed an automated recognition system of phytoplankton species for the evaluation of freshwaters trophic levels (Schulze *et al.*, 2013). The requirement of a mobile and affordable microscopic system that enables image analysis "in the field" was an obvious observation from this project.

Based on these results the company Oculyze[1] was founded with the objective to develop such a system and enable everyone to be a microscopy specialist with the support of machine learning analysis algorithms. The company reached that goal by providing a mobile microscopy solution and custom image recognition software for different biological samples.

The recently developed smartphone microscope (Figure 1) was used as a technical and economical "proof of concept" for the analysis of yeast in beer breweries. With this system, the control of concentration and viability of yeast cells is possible which is helpful to safeguard good and maintain a constant quality of beer (White and Zainasheff, 2010).

For the analysis of blood, there are several microscopic approaches commonly used as diagnostic tools. A specialist can count red and white blood cells, differentiate cell groups, investigate morphological changes and identify or quantify parasites *etc.* For

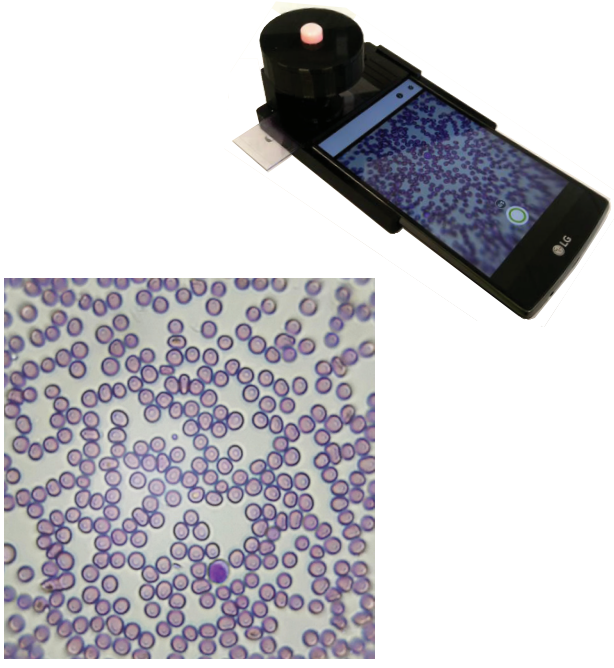


Figure 1. Recent Oculyze smartphone microscope. A stained blood sample is shown with a magnification of 400x. The used Hemacolor® chemicals are staining red blood cells bright purple and white blood cells dark purple.

[1]http://www.oculyze.de

Article history
Received: 16 November 2017
Accepted: 17 December 2018
Published: 05 February 2018

Pfeil et al. (2018) EMBnet.journal 23, e904
http://dx.doi.org/10.14806/ej.23.0.904

example, lower levels of red blood cell counts are an indicator of anaemia. In most cases, this disease is easy to cure because it is mainly caused by iron deficiency. Unfortunately, it is an under-diagnosed illness because of the elaborate application of a microscopic analysis (WHO, 2008). A fast and reliable test could improve the welfare of a large percentage of the population.

There are also many less developed countries in tropical regions that have an endemic problem of malaria infections. Still, so far, the gold standard of diagnosis is the microscopic validation of these parasites. The lack of an adequate infrastructure (laboratories, specialised employees) hinders a fast and reliable diagnosis and treatment (Malaria Atlas Project, 2017).

A mobile microscopic system can be used as a point-of-care diagnostic (POCD) tool directly at the patient bed or in remote areas with poor medical supply.

Materials, Methodologies and Techniques

Oculyze has developed a mobile microscopy system that can be used without expert knowledge (Figure 1). A combination of an optical module, a corresponding smartphone and automated image analysis, enables different cell counting approaches. The whole system (containing hard- and software) is cheaper than a cell counter and does not require trained personnel compared to analysing samples with a laboratory microscope. The optical module enables a 400x magnification with a high resolution of samples. With the connected smartphone, the user can record and analyse microscopic images automatically. Documentation is stored in the cloud, readily available from any internet connected device. The software currently comprises an Android application and cloud-based image analysis algorithms. The captured images are transmitted via mobile network to the cloud server and the results are displayed on the screen after a few seconds.

An uncomplicated adaption of the image analysis software to novel scientific issues is possible. For more challenging tasks, *e.g.* blood analysis, further developments for an advanced mobile microscope catching the following requirements are necessary (Figure 2):

- development of a mobile microscope that can be used with any smartphone or mobile device;
- automatic image recording and adjustable sample holder;
- fast and reliable image analysis for blood samples based on neural networks and deep learning.

Results

At first, requirements for a blood analysis system were evaluated. There are several indicators for the objective to develop a POCD tool:

- simple and fast sample preparation;
- easy to use;
- highly accurate results for diagnostics.

Herein, the sample handling should preferably be carried out by the patient itself - also to avoid the risk of infections. A drop of blood from the fingertip is enough to investigate diagnostic target parameters. It is placed on a microscope slide and white blood cells, thrombocytes and possibly parasites are stained with a DNA-sensitive dye. Operability by user-friendly application needs to be achieved. Yet, the most important point is the diagnostic reliability, which will be achieved with the development of deep learning algorithms and neural networks. The accuracy of this approach is very promising because the results of artificial intelligence programs are the superior

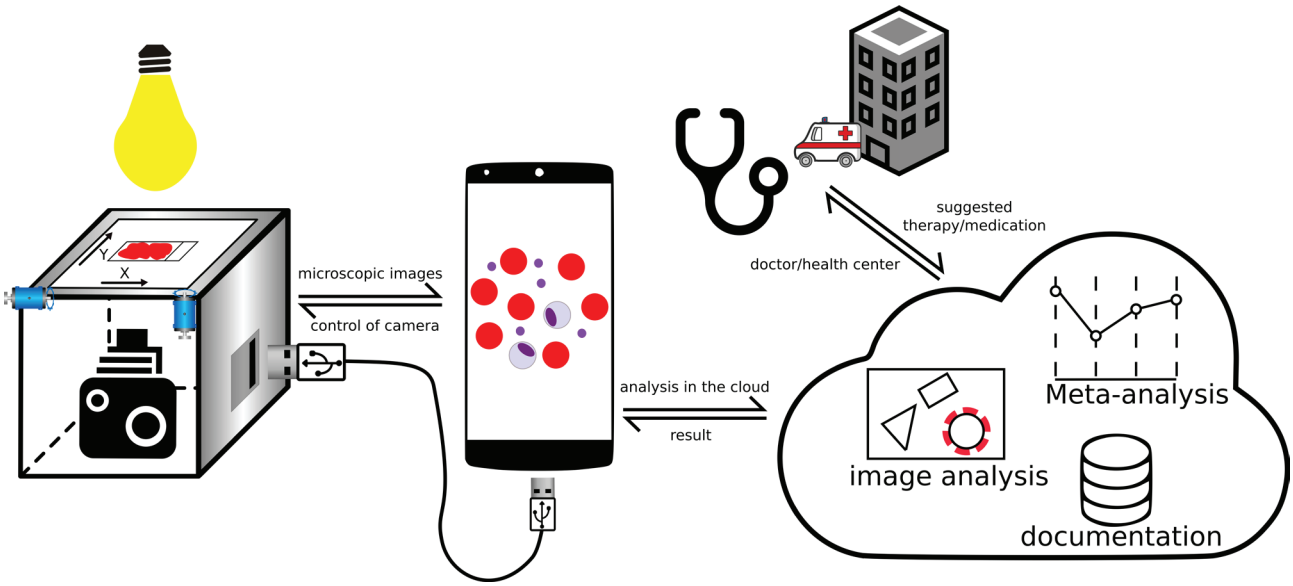


Figure 2. The planned, mobile microscope, consisting of a high-resolution camera, external illumination and a motorised sample holder, is connected via USB to a smartphone or another mobile device (laptop, tablet). An application on this device controls the external camera and transmits the captured images encrypted to a cloud platform. There the image analysis and recognition are carried out, the results are transferred back to the smartphone or to a doctor/health centre.

Pfeil et al. (2018) EMBnet.journal 23, e904
http://dx.doi.org/10.14806/ej.23.0.904

in image recognition compared to other algorithmic models (LeCun *et al.*, 2015).

Conclusions

At the present state, a fast and simple preparation and microscopy of blood samples with the recent system can be achieved. Nevertheless, the current magnification of 400x is sub-optimal for the analysis of small blood components (thrombocytes, parasites). Also for the reliability of diagnostic results, many images of one sample are essential. Therefore further development of the system is imperative. Furthermore, it can be said that nearly every application in medicine/diagnostics, environmental and food technology is imaginable, where mobile microscopy and automated analysis is advantageous. This opportunity will also enable many researchers to investigate biological samples of their area in a more efficient way.

Key Points

- A mobile microscope for the examination of blood samples is beneficial for several
- Applications: cell counting and differentiation, identification/ quantification of parasites
- It enables a fast and reliable diagnosis of blood-related diseases
- Automated image analysis is on the basis of neural networks and deep learning
- Experts on site and laboratories are not needed
- Utilisation as a POCD tool in remote-rural areas and endemic regions is a goal

Acknowledgements

The authors thank Tobias Hagemann for his efforts to investigate the feasibility of malaria detection in Africa. Work for this manuscript was financed by the Ministry of Science, Research and Culture of the federal state of Brandenburg, Germany in the HealthCampus initiative “Digital and analogue companions for an ageing population (DigiLog)” under grant no. GeCa: H228-05/002/004.

References

1. Benoist BD, McLean E, Egll I, Cogswell M (2008) World Health Organization. Worldwide prevalence of anaemia 1993-2005: WHO global database on anaemia (2008). Edited by Bruno de Benoist, Erin McLean, Ines Egli and Mary Cogswell. ISBN: 978 92 4 159665 7
2. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature **521**:436–444 http://dx.doi.org/10.1038/nature14539
3. Malaria Atlas Project: http://www.map.ox.ac.uk/. Version: 2017
4. Schulze K, Tillich UM, Dandekar T, Frohme M (2013) PlanktoVision- an automated analysis system for the identification of phytoplankton. BMC Bioinformatics 2013 **14**:115. http://dx.doi.org/10.1186/1471-2105-14-115
5. White C, Zainasheff J (2010) Yeast: the practical guide to beer fermentation. Brewers Publications. ISBN-13: 978-0937381960

sweet poison

Vivienne Baillie Gerritsen

Where there is a will, there is a way. We have all faced situations that seem hopeless yet, given time and determination, we end up finding a solution. Nature tackles apparent dead ends by using the forces that drive evolution. In this way, features that could appear to be disabling - when considering predation, reproduction or self-defence for instance - are lifted by using well-chosen tactics. Take a lack of mobility, for example. Flowers cannot fly so, to reproduce, they entice insects with intricate scents, nectar and colours who unsuspectingly collect pollen in the process. Marine cone snails are able to move but, like all snails, are sluggish. Unable to pounce on their prey, they release a cloud of toxin - known as the nirvana cabal - that stuns their victim thus giving the snail time to catch it. Toxins that make up the nirvana cabal are of great interest to pharmacologists because they are diverse, rapid and specific. Recently, a form of insulin - Con-Ins G1 - was discovered in certain cone snails that use it to immobilize their prey by causing an insulin shock.

Portrait of Jeanne Hebuterne, 1918

by Amadeo Modigliani (1184-1920)

Cone snails live in warm and tropical seas around the globe. They are carnivorous, and hunt marine worms, small fish, molluscs and even other cone snails. Out of 800 different species, over 100 cone snails are venomous and, when hunting, discharge a cocktail of toxins that smothers and immobilizes their catch or, in some instances, kills it. The toxins involved, and known as conotoxins, usually target the prey's nervous system and elicit a range of reactions: from sedation and paralysis, to sensory overload. Other toxins, like the newly-discovered insulin Con-Ins G1, head for the neuroendocrine system. Two different types of Con-Ins G1 have been identified: 1) molluscan-like insulins that are found in snail and worm hunters, and 2) fish-like insulins that are limited to cone-snail fish hunters.

Not only are toxins widely used in Nature, but they have also been perverted to serve chemical warfare and terrorism, or have starred in famous murder cases. In 1978 for instance, an umbrella was used to fire a ricin-loaded platinum pellet into the thigh of Georgi Markov, a Bulgarian dissident and journalist for the BBC, as he stood waiting for a bus in London. He died three days later. Insulin too has been used as a poison. A famous case of insulin-poisoning is known as the "Von Bülow" case. Sunny von Bülow, an American socialite and heiress, fell into an irreversible coma in 1980, at the age of 48. Tests revealed that the insulin level in her blood was unusually high. Her husband was accused, twice, of attempted murder but on both occasions was found not guilty. His wife died 28 years later.

Venom toxins, such as those used by cone snails, are small peptides, extremely stable

outside cells and highly specific to their physiological targets. Millions of years of evolution have fine-tuned them and given them very precise roles in hunting, self-defence and deterrence. They are remarkably potent and diverse, and their targets are - for the vast majority - receptors, ion channels and transporters located in the victim's nervous system. Besides conotoxins, the fish-hunting cone snail Conus geographus also synthesizes a specialized insulin - Con-Ins G1 - which actually constitutes the greater part of its venom. C.geographus attacks its prey in two different steps. It begins by releasing the nirvana cabal close to the fish it wants to catch, thus significantly reducing its locomotion. It then engulfs it in a sort of distended mouth while injecting a further shot of venom.

Con-Ins G1 is, perhaps unsurprisingly, very similar to fish insulin - implying that it can bind to fish insulin receptors. It consists of four regions: an N-terminal signal peptide, a B chain, one or more C peptides, and an A chain. After proteolytic processing, the A and B chains connected by disulphide bonds form the mature insulin molecule. An unusual feature: Con-Ins G1 undergoes posttranslational modifications that are characteristic to conotoxins, and no doubt play a role in stabilizing the protein's structure. This particular 43 amino-acid insulin is one of the shortest known to date and very similar to vertebrate insulin, especially with respect to the A chain (90% similarity). The human insulin B chain mediates receptor binding as well as the assembly of insulin into its storage form (hexameric). Though this segment is lacking in C.geographus insulin, it can still bind to human insulin receptor but cannot, however, assemble into the storage form - thus making it, like all toxins, readily available.

Conotoxins per se have a wide array of pharmacological targets, which make them potential drug candidates. They are reliable, effective, produce - it is said - no side effects and can, for example, reduce heart rate or pain instantly. They are also being considered in the treatment of Alzheimer's disease, Parkinson's disease, depression, epilepsy and even nerve injury. Scientists have demonstrated that cone snail insulins can mimic human insulin. Vertebrate insulin regulates carbohydrate and fat metabolism while, in the brain, it modulates energy homeostasis and has a role in memory and cognition. Con-Ins G1could therefore be engineered to become an ultra-rapid therapeutic insulin. Before this, however, more research will be needed to understand the molecular subtlety of cone snail venoms, and in particular Con-Ins G1 which could turn out to be an ideal candidate for insulin-related afflictions.

Cross-references to UniProt

Con-Ins G1, *Conus geographus* (Geography cone): A0A0B5AC95

References

1. Menting J.G., Gajewiak J., MacRaild C.A., Hung-Chieh Chou D.., Disotuar M.M., Smith N.A., Miller C., Erchegyi J., Rivier J.E., Olivera B.M., Forbes B.E., Smith B.J., Norton R.S., Safavi-Hemami H., Lawrence M.C.
A minimized human insulin receptor-binding motif revealed in a Conus geographus venom insulin
Nature Structural & Molecular Biology 23:916-920(2016)
PMID: 27617429

2. Safavi-Hemami H., Gajewiak J., Karanth S., Robinson S.D., Ueberheide B., Douglass A.D., Schlegel A., Imperial J.S., Watkins M., Bandyopadhyay P.K., Yandell M., Li Q., Purcell A.W., Norton R.S., Ellgaard L., Olivera B.M.
Specialized insulin is used for chemical warfare by fish-hunting cone snails
PNAS 112:1743-1748(2015)
PMID: 25605914

proteinspotlight

Protein Spotlight (ISSN 1424-4721) is a montly review written by the **Swiss-Prot** team of the **SIB Swiss Institute of Bioinformatics**. Spotlight articles describe a specific protein or family of proteins on an informal tone. http://web.expasy.org/spotlight/

a touch of warmth

Vivienne Baillie Gerritsen

We need heat. All warm-blooded animals know this instinctively because when life leaves us, the cold creeps in fast. Heat is produced in different ways inside us, and not only to keep our body temperature at a healthy level but also to keep it stable. After the fashion of small mobile furnaces, we carry adipose tissues that are full of stored fat waiting to be burnt down to release heat – a process termed thermogenesis. Researchers are becoming more and more interested in thermogenesis, especially adaptive thermogenesis which is the capacity an organism has to adjust its energy needs according to the environment, i.e. the amount of food that is available and the surrounding climate. Because where there is talk of food, there is talk of obesity and its direct cousin diabetes, two afflictions from which millions of people currently suffer worldwide. For some time already, scientists have known that molecules known as N-acyl amino acids, are important in biological processes such as thermogenesis, but they knew little more. Until they discovered an enzyme that is secreted by fat cells in adipose tissues, and that knows how to make them: peptidase M20 domain containing 1, or PM20D1.

Autumn Warmth, by Mary Pym

Courtesy of the artist

It is all very well to be told we are warm-blooded animals and that we need heat – ca. 37° Celsius – to keep us going. But why can our bodies not work at lower – or for that matter – higher temperatures without suffering from hypo- or hyperthermia? Well in theory they could. But to do so, over the millennia they would have needed to invent all sorts of systems to protect themselves from overheating, or freezing for that matter. Many creatures have developed such systems. There are varieties of fish, for instance, that are able to survive in freezing water, and bacteria that thrive in the vicinity of hydrothermal vents. It just so happens that 37° Celsius, or thereabouts, is the ideal temperature for the molecules that make us to function in an optimum way – like the hosts of metabolic pathways they are involved in. Take a protein: at 37°C, its 3D conformation is not denatured and therefore its activity unhindered. Organisms that live in extreme conditions have devised ways of protecting the smooth running of their metabolism. Two examples: antifreeze proteins and heat shock proteins.

One surprising fact is that humans are not comfortable at temperatures above 30°C, let alone 37°C. Our body prefers outside temperatures that are lower than the one it needs inside – and this is why we are condemned to create heat on an almost continuous basis. Some of the heat, we can access upon need, is stored in the form of fat in our adipose tissues, which are made up of adipocyte cells. There are two types of adipocyte cells: those that are bulging

with fat, and those that have less fat but are bulging with mitochondria and are known as brown fat cells. Mitochondria are a cell's powerhouse and are equipped to produce the currency of biological power known as ATP. Many different proteins are active in mitochondria. In those of brown fat cells, there is a protein which depends on PM20D1 and is intimately involved in thermogenesis. Its name: brown fat uncoupling protein 1, or UCP1.

UCP1 feasts on N-acyl amino acids, a process which generates heat. And PM20D1 synthesize N-acyl amino acids, thus providing UCP1 with its substrate. This, however, is a relatively recent finding. N-acyl amino acids are found in many different tissues, and for some time had been described by researchers and shown to have roles in many different biological pathways – from cell migration, cardiovascular function, memory and cognition to inflammation, pain and pathologies such as cancer, neurodegenerative diseases, diabetes and obesity. But how these metabolites are actually synthesized remained a mystery – until PM20D1 was discovered.

PM20D1 is a 500 amino-acid long enzyme. It is synthesized in fat cells in adipose tissue and subsequently secreted – thus able to influence neighbouring cells that are not necessarily specialized in making heat. PM20D1 is expressed upon exposure to cold, and catalyses the condensation of fatty acids and amino acids to form N-acyl amino acids. These N-acyl amino acids are then processed by UCP1 to create heat through the dissipation of chemical energy. Besides condensation, it turns out that PM20D1 is also able to hydrolyse N-acyl amino acids into their fatty acid and amino acid parts. This implies that PM20D1 probably has a role in regulating the levels of fatty acids and amino acids, and hence the production of heat. It could also be that the reaction leading to thermogenesis is actually driven by differences in the levels of fatty acids, or indeed N-acyl amino acids.

So PM20D1 has a role in regulating our body heat, and therefore in energy homeostasis. These are intriguing properties, because where energy homoeostasis is involved so are fat and sugar. It is a fact that when researchers administered N-acyl amino acids to mice, they not only increased energy expenditure – i.e. fat was burned to produce heat – but also improved glucose homeostasis. In the same way, there is a fair chance that a "therapeutic" increase in the expression of PM20D1 would also cause the level of N-acyl amino acids to rise. In a society where millions of people suffer from obesity and diabetes, these metabolites could be of great therapeutic value.

But things are not so straightforward. In order to lose weight, one might – and rightly so – imagine that if one eats less, the body will burn some of its own fat to keep itself going. The thing is, our body has this unique faculty of adapting to many situations, and it can very rapidly adjust its energy output by functioning on less energy per day. This is what is known as adaptive thermogenesis, which is not helpful in the fight against obesity for instance. What is more, PM20D1 and UCP1 are probably not the only proteins involved in thermogenesis and its regulation. Certainly, there seems to be therapeutic hope in N-acyl amino acids. And delving further into the molecular ways of PM20D1 will help pave the way.

Cross-references to UniProt

N-fatty-acyl-amino acid synthase/hydrolase PM20D1, *Homo sapiens* (Human) : Q6GTS8
Mitochondrial brown fat uncoupling protein 1 (UCP1), *Homo sapiens* (Human) : P25874

References

1. Long J.Z., Svensson K.J., Bateman L.A., Griffin P.R., Nomura D.K., Spiegelman B.M.
The secreted enzyme PM20D1 regulates lapidated amino acid uncouplers of mitochondria
Cell 166:424-435(2016)
PMID: 27374330

proteinspotlight

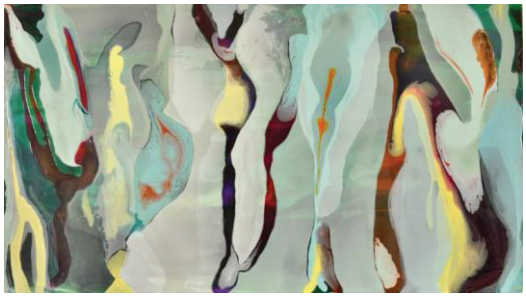
SIB Swiss Institute of Bioinformatics

Protein Spotlight (ISSN 1424-4721) is a montly review written by the **Swiss-Prot** team of the **SIB Swiss Institute of Bioinformatics**. Spotlight articles describe a specific protein or family of proteins on an informal tone. http://web.expasy.org/spotlight/

seeking past shelter

Vivienne Baillie Gerritsen

Nothing can survive without the means to defend itself. If bacteria are unable to protect themselves from freezing temperatures, they perish. If we cannot fight off the flu virus, we pass away. If plants cannot ward off toxic fungi, they wilt and die. In fact, we all spend a lot of time shunning "stresses", of either biological (biotic) origin, or non-biological (abiotic) origin. The good part is that when an organism has managed to check an infection or deal with harsh conditions once, it does not forget and will react all the faster if the same thing occurs again. In other words, somehow and somewhere, memories are engraved in an organism. This is precisely how a vaccination works in humans. Needless to say, scientists have also found ways to prepare a plant's resistance mechanisms in advance by treating it with certain substances or presenting it with stressful environmental conditions. This is called plant defence priming. Researchers also observed that this acquired state of a plant can also be inherited, which is like passing down a form of instinct: that of knowing how to deal with the enemy. One protein is known to be involved in the priming process, and has no doubt a role in preserving this protective memory. It has been named protein Impaired in BABA-Induced Sterility 1, or IBS1.



Aquatic Layers, by Laurie Levin

Courtesy of the artist

When an organism, such as a plant, is infected by a virus, bacteria or any other microorganism which may cause the plant harm, a whole series of molecular pathways are triggered off to defeat the pathogen. The same occurs when a plant has to deal with surroundings that are temporarily trying, such as periods of drought or high salinity for example. Once the biotic or abiotic stress has been overcome, the plant has ways of "remembering" what was unpleasant, and if the situation is to occur again, it will be faster to respond. Both a timesaver and a lifesaver, such acquired resistance has no doubt been used by living beings since the dawn of time. It is also the fodder of evolution.

Acquired resistance to pathogens can also be induced artificially. We all remember when the doctor came to school for our vaccinations, and we were told to join the long silent queue, with a smell of disinfectant in the air and our shirt sleeve rolled up as far as it would go, awaiting our dreaded turn as we watched those before us receiving their jag. Vaccinations are merely a way of engraving in our system the memory of a given pathogen, and if it happens to cross our path again, our bodies are able to react swiftly thereby lessening the pathogen's chance of doing too much harm. The same can be done by inoculating different substances into plants, and it is called priming. One such substance, which is able to induce resistance to a broad spectrum of biotic and abiotic stresses, is known as ß-aminobutyric acid, or BABA.

How can one sole substance prepare a plant to fight against such a wide range of pathogenic circumstances? Though plant defence priming has been known for decades, the molecular mechanisms underlying it are still poorly understood. It could be that BABA encourages the accumulation of signalling proteins – or perhaps their post-translational modification – which remain inactive for as long as the plant hasn't encountered some form of stress. When this happens, the signalling proteins would then be pulled out of their torpor to fire off signal transduction pathways involved in defence. Another hypothesis involves the accumulation

of pre-expressed transcription factors which would then react to stress by stimulating a set of defence genes. More recently, the importance of epigenetic mechanisms, such as histone modification or DNA methylation for instance, has been suggested – two processes known to be essential in the regulation of gene expression.

Scientist discovered the existence of at least one protein directly involved in the BABA priming of *Arabidopsis thaliana*. It has been given the unattractive name of Impaired in BABA-induced sterility 1 (or ISB1), because although a low concentration of BABA promotes priming, too high a concentration brings about sterility. Though, to date, ISB1 has revealed very little about itself, we do know that it regulates BABA priming and is involved in distinct signal transduction pathways. It bears sequence resemblance to kinases involved in the control of signal transduction pathways which regulate gene transcription, and is also similar to kinases that have a role in stress-related responses in plants. It also carries membrane-binding properties that are characteristic of many signal transduction proteins. Signal transduction, gene transcription, stress-related responses… three biological processes that are central to an organism's defence.

One surprising observation: priming seems to be hereditary. In other words, if a plant has acquired induced resistance to stress, this induced resistance – or its memory – is passed down to the plant's progeny. This means that the direct progeny of a primed plant does not need to be treated further with BABA to be primed itself. What is more, if the primed group of seedlings is also treated with BABA, it reacts fast, as though it were "primed to be primed". However, if BABA treatment is not repeated, the primed state seemingly dilutes in the succeeding generations to fade away altogether – although the passing of the primed state also depends on pathogen severity and the priming agent. How is the primed state transferred to the next generation from a molecular point of view? It could well be the doings of epigenetic mechanisms: DNA methylation, for instance, which is an ideal candidate because of its stability.

Plant priming has obvious implications for sustainable agriculture and its economics, as well as for generating crop varieties. Plants can be prepared to react rapidly to stress, and their offspring are either already prepared or can be readily primed by using lower concentrations of the priming agent. Furthermore, a primed plant is kept in a state of alert and not in an energy-consuming situation where, for example, several signal transduction pathways have kicked off but are subsequently frozen as they await the stress signal. The notion that a memory – such as resistance – can be passed down generations without it being actually inscribed in genes is very intriguing, and echoes the importance epigenetic mechanisms may have in the inheritance of states caused by traumas experienced by our forefathers, such as melancholy or depression.

Cross-references to UniProt

Protein Impaired in BABA-Induced Sterility 1, *Arabidopsis thaliana* (Mouse-ear cress): F4ICB6

References

1. Ana Slaughter, Daniel X., Flors V., Luna E., Hohn B., Mauch-Mani B.
Descendants of primed Arabidopsis plants exhibit resistance to biotic stress
Plant Physiology 158:835-843(2012)
PMID: 22209872

2. Conrath U., Beckers G.J.M., Flors V., García-Austín P., Jakab G., Mauch F., Newman Mari-Anne, Pieterse C.M., Poinssot B., Pozo M.J., Pugin A., Schaffrath U., Ton J., Wendehenne D., Zimmerli L, Mauch-Mani B.
Priming: Getting ready for battle
Molecular plant-microbe interactions 19:1062-1067(2006)
PMID: 17022170

proteinspotlight

Protein Spotlight (ISSN 1424-4721) is a montly review written by the **Swiss-Prot** team of the **SIB Swiss Institute of Bioinformatics**. Spotlight articles describe a specific protein or family of proteins on an informal tone. http://web.expasy.org/spotlight/

whispers

Vivienne Baillie Gerritsen

We all depend on cues. Without them, the notion of community would not exist. Cues are the cement of society, and their nature can be very diverse. Birds whistle. Hogs grunt. Plants give off scents. Fish use bioluminescence. Slugs release pheromones. Humans talk. Many species have more than one way of flinging cues to one another: while capable of emitting sounds, they can also discharge smells, touch each other and make gestures. Humans, for example, have brought signalling to a peak by adding clothes, tattoos, piercing, makeup, jewellery and all forms of bodily transformations to their repertoire to add refinement – and perhaps a touch of egocentricity – to their means of exchange. But though it may seem that individualism is, paradoxically, what drives communication these days, every signal is a manifestation of the belonging to a part – however small – of society. Many other animals have also evolved intricate means of communication. Ants, in particular. Over time, these insects have acquired an advanced form of social behaviour driven by these mysterious invisible cues called pheromones whose effects depend highly on a protein known as odorant receptor co-receptor, or Orco.

Synaesthetics, by Margaret Mannion Kallen

Courtesy of the artist

Ants and their intricate social behaviour have been under scientific scrutiny since the second half of the 19th century. The Swiss psychiatrist Auguste Forel (1848-1931) was one of the first scholars to consider them. In 1874, an initial account of his observations was made in a long treatise that was applauded by Charles Darwin, followed 50 years later by a five-volume opus on the social habits of ants: *Le monde social des fourmis*. Though the parallels he made between ant behaviour and human social and political behaviour remained controversial, Forel's contribution to the field was fundamental. So much so that, in 1979, Switzerland issued a 1,000 franc bank note bearing his portrait on one side, and drawings of ants on the other. At about the same period as Forel, the American entomologist William Morton Wheeler (1867-1937) was also studying ant behaviour, and saw each ant colony as an organism *per se*, thus founding the notion of superorganism. Towards the middle of the 20th century, the social behaviour of ants could be studied in the light of genetics. This gave birth to the field of sociobiology, pioneered by the American biologist and theorist E.O.Wilson.

If you have spent time watching ants, you will have noticed that they tend to follow each other along the same paths, which either lead them away from or towards their home: the ant hill. These are the older ants that are sent out to find materials for nest building or food for their kin. The younger ants stay inside, looking after the even younger ants, while yet other ants are building nests, nursing, foraging or taking care of the queen. Each ant knows what it has to do. And it does it without words and books of regulations. So what is it that guides them? What is sending out the orders? Pheromones. Pheromones are molecules that no one can see or smell – a sort of biological whisper that has the disturbing power of acting upon an organism's behaviour. In the

world of ants, pheromones are capable of giving shape to a whole society. The queen releases pheromones that prevent her progeny from being reproductive, while other pheromones keep ants on the same trail, stimulate them to socialize, groom eggs, nurse, forage, elicit alarm responses, build nests…

Pheromones influence behaviour by triggering off a reaction sensed by the ants' antennae, and which is relayed to the brain. They do this by binding to specific receptors – odorant receptors, or ORs – that are lodged in the membranes of antennal odorant receptor neurons in the antennae and whose axons plunge into the antennal lobe, itself composed of a dense network of globule-shaped nerve fibres, or glomeruli. When a pheromone activates an OR, a series of transduction pathways are set in motion, from the antennae down to the lobes and then to the ant's central nervous system. None of this is new; humans taste food and smell scents in much the same way for instance. What is surprising in ants is the co-receptor Orco: this protein is involved not only in binding ORs but in a number of other events too.

Orco is a highly conserved transmembrane olfactory co-receptor that forms a heterodimer with all insect ORs. Upon pheromone-binding, the Orco/OR heterodimer acts as a ligand-gated ion channel thus activating the odorant receptor neurons and transmitting the message down to the antennal lobe. When Orco is deficient in ants, some are found wandering on their own or unable to forage successfully, while others are seen to twitch their antennae abnormally or lay only few eggs and not tend to them. Such abnormal behaviour would be expected if the pheromone transduction pathway is tampered with. What scientists also discovered, however, is that not only does Orco show signs of being far more strongly involved in the olfactory pathway than it is in that of other insects but it also seems to have a direct role in 1) locating and maintaining ORs in the neuronal membranes as well as in 2) regulating the number of glomeruli in the antennal lobes – which would imply that it has a role in the ant's neural development. With respect to other insects, the ant olfactory system is particularly intricate. As an order of magnitude – and perhaps complexity – the repertoire of OR genes in *Drosophila* is encoded by 60 genes, whereas it is encoded by over 300 genes in the ant genome! As for the *Drosophila* antennal lobes, they contain about 42 glomeruli while the ant lobes have over 400...

Though Orco seems to have many skills, no one knows how the co-receptor fulfils them on the molecular level. The ant olfactory system probably still depends on an ancestral olfactory mechanism since, even when Orco is deficient, some ants are able to carry out typical social behaviours such as eliciting alarm responses or grooming eggs. Orco certainly seems to have an important role in the development of the ant's olfactory system, and it is crucial in the study of social behaviour that depends on chemicals. Though the brains of humans and ants cannot be compared, the wandering and antennae-twitching anti-social ants echo certain psychiatric traits in humans, which we know can be rectified with chemistry. The ant olfactory system is also a source of inspiration for network engineers. How do ants avoid congestion? How do they optimize their movements? It is very intriguing ground, and leaves us with that slightly troubled feeling that we are perhaps not the complete masters of our actions.

Cross-references to UniProt

Odorant receptor co-receptor, *Harpegnathos saltator* (Jerdon's jumping ant): E2BJ30
Odorant receptor co-receptor, *Ooceraea biroi (Clonal raider ant)*: A0A026W182

References

1. Trible W., Olivos-Cisneros L., McKenzie S.K., *et al.*
Orco mutagenesis causes loss of antennal lobe glomeruli and impaired social behavior in ants
Cell 170:727-735(2017)
PMID: 28802042

2. Yan H., Opachaloemphan C., Mancini G., *et al.*
An engineered *orco* mutation produces aberrant social behaviour and defective neural development in ants
Cell 170:736-747(2017)
PMID: 28802043

proteinspotlight

Swiss Institute of Bioinformatics

Protein Spotlight (ISSN 1424-4721) is a montly review written by the **Swiss-Prot** team of the **SIB Swiss Institute of Bioinformatics**. Spotlight articles describe a specific protein or family of proteins on an informal tone. http://web.expasy.org/spotlight/

EMBnet.journal
Bioinformatics in Action

ISSN 1023-4144

DEAR READER,

If you have any comments or suggestions regarding this journal we would be very glad to hear from you. If you have a tip you feel we can publish then please let us know. Before submitting your contribution read the "Authors guidelines"[1] and send your manuscript and supplementary fi les using our on-line submission system[2].

Past issues are available as PDF fi les from the web archive[3].

Visit EMBnet website for more information: www.journal.embnet.org

EMBNET.JOURNAL EXECUTIVE EDITORIAL BOARD

[1]http://journal.embnet.org/index.php/embnetjournal/about/submissions#authorGuidelines
[2]http://journal.embnet.org/index.php/embnetjournal/author/submit
[3]http://journal.embnet.org/index.php/embnetjournal/issue/archive

PUBLISHER

EMBnet Stichting p/a
CMBI Radboud University
Nijmegen Medical Centre
6581 GB Nijmegen
Th e Netherlands

Email: erik.bongcam@slu.se
Tel: +46-18-67 21 21