EMBnet.journal

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...

Editorial

This editorial marks 20 years of existence of EMBnet publications, first in the form of EMBnet. news and later, when the publication changed character to a peer reviewed journal, under the name EMBnet.journal.

EMBnet.news was one of the first publications in the field of bioinformatics, at a time when the term bioinformatics was barely known; nevertheless, the first articles were written by some of the earliest European bioinformaticians, and were distributed to the nascent bioinformatics community. EMBnet.news led the way in publishing articles, reviewing newly created bioinformatics software tools (EGCG 8.0, the Staden package, EMBOSS, W2H, the SRS Sequence Retrieval System, CINEMA and many others), and reviewing books in the field.

This year, EMBnet.journal has changed the way it publishes articles to a more agile model, to adapt to the rapidly transforming digital publishing arena. Once submitted via the Open Journal System, articles are sent for peer review and are published as soon as possible after acceptance. All articles are published online, and a complete volume is produced and archived at the end of the year.

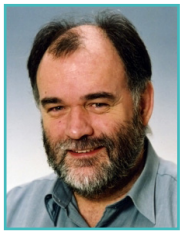
In this way, we strive to produce a journal that lives up to its motto, "Bioinformatics in Action". We will, during 2015, invite all our previous authors and encourage new ones to submit contributions to the journal: we accept full research and education articles, as well as technical reports and reviews of new software, databases and books. The Editorial Board welcomes our readers to a new decade with one of the oldest publications in the field of bioinformatics.

EMBnet.journal Editorial Board

Contents

Cover picture: DNA spiral, BMC, Uppsala, Sweden, 2002, © Erik Bongcam-Rudloff

Next generation sequencing and phylogenetic networks



David A. Morrison

Swedish University of Agricultural Sciences, Uppsala, Sweden

Received 20 March 2014; **Published** 26 March 2014

Morrison DA (2014) *EMBnet.journal* **20**, e760. http://dx.doi.org/10.14806/ej.20.0.760

Introduction

Next Generation Sequencing (NGS), or massively parallel sequencing, can potentially provide a fast and cost-effective means of generating multi-locus sequence data for phylogenetics, which is the field that tries to reconstruct the genealogical history of evolutionary change. Unfortunately, the cost for the number of samples typically employed in phylogenetics is currently still beyond the reach of most researchers. This will soon change, and phylogenetics will become phylogenomics.

Phylogeneticists therefore now need to think about the relationship between NGS and their current paradigms, in terms of both data analysis and interpretation. In particular, there has been recent interest among phylogeneticists in using phylogenetic networks rather than phylogenetic trees as the main paradigm for interpretation (Morrison, 2011; Bapteste *et al.*, 2013). Trees are intended only for the study of vertical evolutionary processes, directly from parent to offspring; but networks can accommodate horizontal processes as well, such as recombination, hybridisation, introgression and horizontal gene transfer, all of which are common in one taxonomic group or another. These horizontal processes are represented by reticulations in the network, which do not appear in a tree.

Most of the published discussions about NGS in relation to phylogenetics have focused on trees, rather than networks (Rannala and Yang, 2008; Whelan, 2011; McCormack *et al.*, 2013; Lemmon and Lemmon, 2013). Here, I raise some of the important issues that need to be addressed when using networks.

NGS and phylogenetics

NGS and phylogenetics have so far had only a brief association. McCormack *et al.* (2013) have commented on this:

"Despite this obvious potential, NGS has been slow to take root in phylogenetics compared to other fields like metagenomics and disease genetics. We suggest that this lag has been caused by four specific aspects of phylogeographic and phylogenetic research: the predominant focus on non-model organisms, the need for sequencing large numbers of samples per species, the lack of consensus regarding library preparation protocols for particular research questions, and the transitional state of the technology (whole-genome data are still neither cost-effective, nor even desirable for phylogenetics, but are paradoxically easier to collect).

Another issue is the historical importance of utilizing gene trees in phylogenetics. Gene trees are most robustly inferred from loci with high information content, for example, a non-recombining locus containing a series of linked SNPs. Individual SNPs, on the other hand, have low information content on a per-locus basis and have been used predominately with classification methods such as Structure and Principal components analysis ... While distance-based genealogies and phylogenies can be built from unlinked SNPs, this ignores models of molecular substitution and probabilistic tree-searching algorithms that have led to more robust phylogenetic inference in the last several decades."

Furthermore, no-one has yet shown that many of the questions currently being asked by phylogeneticists will actually benefit from genomic data. We may well be able to answer some new questions, but that is quite a different thing from NGS initiating a revolution, as it has done in other fields of biology. The essence here is that, in science, the questions must come first — collecting data for the sake of it is usually unproductive. So, we need a clear demonstration that genomics is actually needed in phylogenetics (as opposed to other disciplines, where it may indeed be very useful). If an increased volume of data will solve a phylogenetic problem, then that is good, but there is no necessary reason to expect that it will

happen. Statistically, the extra data can lead to improved precision, but not necessarily improved accuracy. In science, targeted data collection has always been the most productive approach to any clearly stated experimental question.

For example, the estimated relationships among humans, chimpanzees, and gorillas did not change as a result of genome sampling rather than gene sampling (Galtier and Daubin, 2008), nor did those of malaria species (Kuo *et al.*, 2008), nor those of mammal super-orders (Hallström and Janke, 2010) or even the orders of wingless insects (Dell'Ampio *et al.*, 2014). In all four cases, the inferred relationships were just as complex after the genome sequencing as before — the resolution of controversial branches in the phylogenetic trees did not occur as a result of increased access to character data.

In this sense, a small sample of representative gene sequences should reveal just as much of the genealogical truth as will a genome-wide sample. A recent empirical example is presented by O'Neill *et al.* (2013), who found that including less informative loci added so much noise to the phylogenetic signal that the analysis eventually broke down. The issue here is that, as data volume increases, so does the potential occurrence of systematic bias owing to model misspecification.

This sort of problem can easily be visualised using phylogenetic networks. Here, genome-scale data frequently produce unresolved bushes rather than tree-like phylogenies, as shown by Beiko (2011), whose analysis involved 298 completely sequenced bacterial genomes, or Decker *et al.* (2009), who analysed 372 individuals belonging to 48 breeds of cattle. Bush-like phylogenies may represent complex evolutionary histories, but they may also represent a failure of phylogenetic analysis; and it is important to be able to distinguish between these two possibilities.

This all suggests that we will need to think carefully about how to apply phylogenetic networks to genome-scale data. Much of the lack of resolution may very well come from the nature of NGS, rather than from the actual evolutionary history.

NGS and networks

There are a number of potential problems with NGS. These may not matter so much for tree-building algorithms, but it is a different matter for networks. They each need to be thought about to assess whether they are serious problems or only of minor concern.

Increased homoplasy owing to sequencing errors

An error rate of even 0.01% is considered good in NGS (*e.g.*, Roche 454: 1%; Illumina HiSeq: 0.1%; Life SOLiD: 0.01%), but when this is extrapolated to the genome scale, it results in thousands of errors. Networks are sensitive to this magnitude of stochastic error. Indeed, one of the valuable uses of phylogenetic networks is specifically to identify data errors. For example, they have been used for detecting chimeric sequences resulting from laboratory-induced errors (Kong *et al.*, 2008), or detecting possible errors in mitochondrial DNA (miDNA) genomes sequenced to find mutations associated with particular diseases (Bandelt *et al.*, 2009).

Increased homoplasy owing to intra-gene processes

These include substitutions, deletions, duplications (especially tandem repeats), inversions and translocations. These processes can potentially reveal evolutionary history, but we have little idea about how best to process the data in a way that will reveal that history. Currently, we deal with this by lumping most of the processes together in the analysis model as 'indels'. This approach is likely to be inadequate for networks, because these very processes may be involved in horizontal evolution.

Increased homoplasy owing to inter-gene processes

The main processes known to confound attempts to identify reticulate evolution are incomplete lineage sorting and gene duplication–loss. The more genes that are sampled, then the greater will be the effect of these confounding processes. There are several methods available for addressing them in the context of estimating phylogenetic trees (*e.g.*, Knowles and Kubatko, 2010; Blair and Murphy, 2011; Bansal *et al.* 2012), but the applicability of these methods to networks is still being assessed (Kubatko, 2009).

Increased homoplasy in non-coding regions

Sanger sequencing in phylogenetics is usually targeted towards gene-coding regions or their introns, but genome-scale data can include what is currently called 'junk DNA'. The evolution-

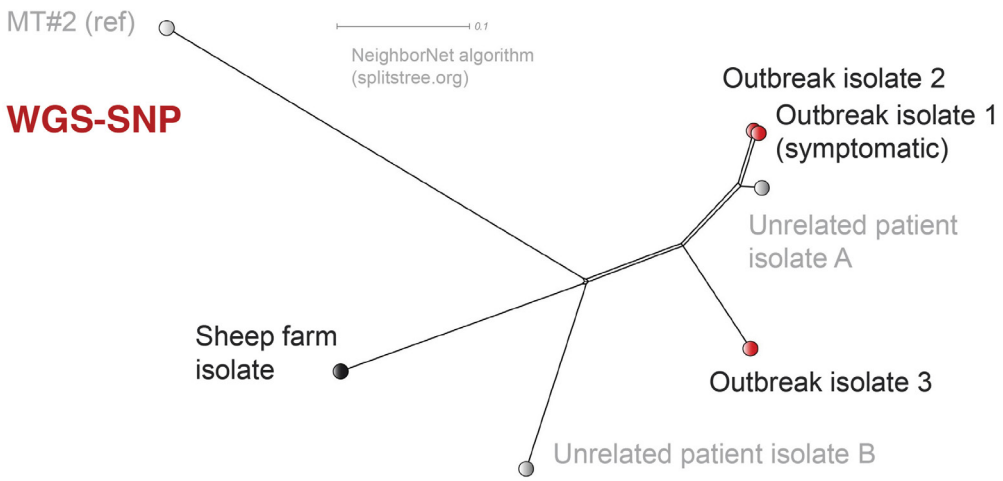


Figure 1. NeighborNet analysis of SNP data from whole-genome shotgun sequencing of seven *E. coli* strains. This network is very tree-like, so that the reticulations are unlikely to represent biologically important processes. The phylogenetic interpretation of the network is thus very straightforward.

ary processes in these regions are currently unknown, so they are difficult to model; and their applicability to phylogenetic analysis has not yet been assessed.

Inadequacies owing to data-processing methods

The analysis of NGS data is often a black art — each published paper seems to provide its own way of processing the data. This has been a cause of concern expressed in the literature (*e.g.*, Check Hayden, 2012; Editorial, 2012a, 2012b; MacArthur, 2012), especially in light of the currently poor documentation and archiving of bioinformatics programs (Cuticchia and Silk, 2004). Perhaps the most talked-about problem is ascertainment bias, especially when SNP (Single Nucleotide Polymorphisms) variants are reported only if they do not match a specified reference genotype. Non-reported variants can just as well be sequencing failures, or coverage gaps, or insufficient evidence for a non-reference variant. Networks generated from such data are likely to consist largely of artefacts.

Network analysis of NGS data

All of this might make the application of networks to phylogenomics problematic in many cases, because we already have enough challenges dealing with the data from Sanger-style sequencing, without having them be orders of magnitude worse. It will therefore be very interesting to see what emerges from the current attempts to apply phylogenetic networks to NGS data. To date, most of the analyses have been ad hoc in nature (Dagan, 2011).

There have been a few applications of EDA (Exploratory Data Analysis) programs, such as SplitsTree[1], mostly involving bacteria and viruses (*e.g.*, Beiko, 2011), and often in the context of detecting recombination. Not all of these studies have produced networks that look bushy, as shown by Figure 1, from Söderlund *et al.* (2013).

SplitsTree is mostly limited by the number of samples, not by the number of characters, so that genomic data are not a particular analysis issue for network algorithms such as NeighborNet. However, it might be necessary to calculate the inter-sample distances outside of this program, unless you want the simple p-distance (popular genome-scale distances include F_{st}).

There have also been programs developed for the study of admixture (or introgression) in human genomes, such as TreeMix[2], AdmixTools[3] and MixMapper[4], and these might repay wider exploration. Essentially, they first construct a phylogenetic tree and then add network reticulations based on various criteria. As is usual with this general approach, there is a problem constructing the initial tree in the presence of reticulation processes. Moreover, there seems to be no clear

1 www.splitstree.org
2 https://code.google.com/p/treemix/
3 genetics.med.harvard.edu/reich/Reich_Lab/Software.html
4 groups.csail.mit.edu/cb/mixmapper

criterion for when to stop adding reticulations — optimisation criteria always increase as reticulations are added, so that increasingly complex networks will always be preferred mathematically.

Conclusion

We need to make sure that we are getting the most out of NGS that we can in phylogenetics, because the times are changing and we need to move with them. However, when moving, the cart should not be leading the horse, and so the phylogenetic horse needs to think carefully about its relationship to the NGS cart. It should be exciting to see the horse and cart working together well, sometime soon.

References

Bandelt H-J, Yao Y-G, Bravi CM, Salas A, Kivisild T (2009) Median network analysis of defectively sequenced entire mitochondrial genomes from early and contemporary disease studies. *Journal of Human Genetics* **54**, 174-181. http://dx.doi.org/10.1038/jhg.2009.9

Bansal MS, Alm EJ, Kellis M (2012) Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics* **28**, i283-i291. http://dx.doi.org/10.1093/bioinformatics/bts225

Bapteste E, van Iersel L, Janke A, Kelchner S, Kelk S, McInerney JO, Morrison DA, Nakhleh L, Steel M, Stougie L, Whitfield J (2013) Networks: expanding evolutionary thinking. *Trends in Genetics* **29**, 439-441. http://dx.doi.org/10.1016/j.tig.2013.05.007

Beiko RG (2011) Telling the whole story in a 10,000-genome world. *Biology Direct* **6**, 34. http://dx.doi.org/10.1186/1745-6150-6-34

Blair C, Murphy RW (2011) Recent trends in molecular phylogenetic analysis: where to next? *Journal of Heredity* **102**, 130-138. http://dx.doi.org/10.1093/jhered/esq092

Check Hayden E (2012) RNA studies under fire. *Nature* **484**, 428. http://dx.doi.org/10.1038/484428a

Cuticchia J, Silk G (2004) Bioinformatics needs a software archive. *Nature* **429**, 241. http://dx.doi.org/10.1038/429241b

Dagan T (2011) Phylogenomic networks. *Trends in Microbiology* **19**, 483-491. http://dx.doi.org/10.1016/j.tim.2011.07.001

Decker JE, Pires JC, Conant GC, McKay SD, Heaton MP, Chen K, Cooper A, Vilkki J, Seabury CM, Caetano AR, Johnson GS, Brenneman RA, Hanotte O, Eggert LS, Wiener P, Kim J-J, Kim KS, Sonstegard TS, Van Tassell CP, Neibergs HL, McEwan JC, Brauning R, Coutinho LL, Babar ME, Wilson GA, McClure MC, Rolf MM, Kim J, Schnabel RD, Taylor JF (2009) Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. *Proceedings of the National Academy of Sciences of the USA* **106**, 18644-18649. http://dx.doi.org/10.1073/pnas.0904691106

Dell'Ampio E, Meusemann K, Szucsich NU, Peters RS, Meyer B, Borner J, Petersen M, Aberer AJ, Stamatakis A, Walzl MG, Minh BQ, von Haeseler A, Ebersberger I, Pass G, Misof B (2014) Decisive data sets in phylogenomics: lessons from studies on the phylogenetic relationships of primarily wingless insects. *Molecular Biology and Evolution* **31**, 239-249. http://dx.doi.org/10.1093/molbev/mst196

Editorial (2012a) Must try harder. *Nature* **483**, 509. http://dx.doi.org/10.1038/483509a

Editorial (2012b) Error prone. *Nature* **487**, 406. http://dx.doi.org/10.1038/487406a

Galtier N, Daubin V (2008) Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* **363**, 4023-4029. http://dx.doi.org/10.1098/rstb.2008.0144

Hallström BM, Janke A (2010) Mammalian evolution may not be strictly bifurcating. *Molecular Biology & Evolution* **27**, 2804-2816. http://dx.doi.org/10.1093/molbev/msq166

Knowles LL, Kubatko LS (eds) (2010) *Estimating Species Trees: Practical and Theoretical Aspects*. Wiley-Blackwell, Hoboken NJ. http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470526858.html

Kong Q-P, Salas A, Sun C, Fuku N, Tanaka M, Zhong L, Wang C-Y, Yao Y-G, Bandelt H- J (2008) Distilling artificial recombinants from large sets of complete mtDNA genomes. *PLOS One* **3**, e3016. http://dx.doi.org/10.1371/journal.pone.0003016

Kubatko LS (2009) Identifying hybridization events in the presence of coalescence via model selection. *Systematic Biology* **58**, 478-488. http://dx.doi.org/10.1093/sysbio/syp055

Kuo C-H, Wares JP, Kissinger JC (2008) The Apicomplexan whole-genome phylogeny: an analysis of incongruence among gene trees. *Molecular Biology & Evolution* **25**, 2689-2698. http://dx.doi.org/10.1093/molbev/msn213

Lemmon EM, Lemmon AR (2013) High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution & Systematics* **44**, 19.1–19.23. http://dx.doi.org/10.1146/annurev-ecolsys-110512-135822

MacArthur D (2012) Face up to false positives. *Nature* **487**, 427-428. http://dx.doi.org/10.1038/487427a

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution* **66**, 526-538. http://dx.doi.org/http://dx.doi.org/10.1016/j.ympev.2011.12.007

Morrison DA (2011) *Introduction to Phylogenetic Networks*. RJR Productions, Uppsala. http://www.rjr-productions.org/Networks/

O'Neill EM, Schwartz R, Bullock CT, Williams JS, Shaffer HB, Aguilar-Miguel X, Parra-Olea G, Weisrock DW (2013) Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex. *Molecular Ecology* **22**, 111-129. http://dx.doi.org/10.1111/mec.12049

Rannala B, Yang Z (2008) Phylogenetic inference using whole genomes. *Annual Review of Genomics and Human Genetics* **9**, 217-231. http://dx.doi.org/10.1146/annurev.genom.9.081307.164407

Söderlund R, Jernberg C, Källman C, Hedenström I, Eriksson E, Bongcam-Rudloff E, Aspán A (2013) Rapid whole genome sequencing investigation of a familial outbreak of *E. coli* O121:H19 with a sheep farm as the suspected source. *Bioinformatics in Action* **19** suppl.A, 89-90. http://journal.embnet.org/index.php/embnetjournal/article/view/657

Whelan N (2011) Species tree inference in the age of genomics. *Trends in Evolutionary Biology* **3**, e5. http://dx.doi.org/10.4081/eb.2011.e5

http://dx.doi.org/10.14806/ej.20.0.752

Next Generation Sequencing methods for identification of mutations and large structural variants, 11 - 12 March 2014

Vital-IT and the Bioinformatics Unravelling groups of the SIB Swiss Institute of Bioinformatics are organising a joint ALLBIO and COST SeqAhead event in Lausanne, Switzerland

Grégoire Rossier (AllBio partner) & Laurent Falquet (SeqAhead partner) are pleased to announce a joint ALLBIO and COST SeqAhead event in Lausanne, Switzerland.
It is entitled **"Next Generation Sequencing (NGS) methods for identification of mutations and large structural variants"** and will be held from Tuesday 11 March to Wednesday 12 March, 2014.

This international event covers several aspects of the identification of genomic structural variants using NGS data. The mini symposium (Day 1) will present the latest developments in the field and the workshop (Day 2) will allow participants to get used to the tools with a virtual machine prepared during a test case hackathon. Particular emphasis will be given to the comparison of the different analysis tools and how to combine their results.
The objective of the mini symposium is to provide an overview of the existing tools/pipelines available for NGS analysis, as well as to present some data using those tools. The objective of the workshop is to allow participants using the pipeline, either with our data, or with their own data.

Requirements for the workshop:

- Basic knowledge of UNIX
- A laptop with at least 4 GB RAM, 50 GB of free disk space, WIFI and VirtualBox preinstalled

The number of seats is limited to 90 for the mini symposium and to 25 for the workshop. Further information and application are available from http://edu.isb-sib.ch/course/view.php?id=104.

ALLBIO (http://www.allbioinformatics.eu) is the main sponsor of this event, and in addition 10 seats of the workshop are sponsored by the COST Action SeqAhead (http://www.seqahead.eu).

From high-throughput structural bioinformatics to integrative systems biology: NETTAB 2014


Francesca Cordero[1], **Paolo D. M. Romano**[2]✉

[1]Department of Computer Science, University of Torino, Italy
[2]IRCCS San Martino IST, Genoa, Italy

Received 12 May 2014; **Published** 13 May 2014

Cordero F and Romano PDM (2014) *EMBnet.journal* **20**, e772. http://dx.doi.org/10.14806/ej.20.0.772

The NETTAB 2014 Workshop will be held in Torino, at the Molecular Biotechnology Centre, from 16 to 17 October 2014. It will be a joint event with the "2014: Crystal (cl)Year" meeting[1], in the International Year of Crystallography 2014, and it will be followed by the annual meeting of the University of Torino's Centre for Complex Systems in Molecular Biology and Medicine.

The workshop represents a virtual bridge between these two events, showing how to manage and elaborate structural and high-throughput proteomics data so that it may be integrated with information from other life science disciplines, aiming to reach a richer description and deeper understanding of mechanisms and interactions in the human being, in its physiological and pathological states. The workshop title, *"From high-throughput structural bioinformatics to integrative systems biology"*[2], reflects this scope.

The topics of the workshop will therefore relate to methods, tools, applications and perspectives on structural bioinformatics, proteomics and integrative systems biology. These issues are very relevant for several research communities, which are invited to join forces and create synergies for an interdisciplinary effort aimed at developing new tools at the interfaces of these disciplines. Contributions to the NETTAB 2014 workshop should, then, be focused on, but not limited to, the following non-exhaustive list of topics: bioinformatics methods, tools and applications for models, standards and management of high-throughput biological data, data integration, structural bioinformatics, functional proteomics, mass spectrometry, drug discovery, systems biology.

The workshop will run from the morning of Thursday 16 to the afternoon of Friday 17 October. It will include four keynote lectures, given by Wolfgang Marwan (Otto-von-Guericke Universität), Ram Samudrala (University of Washington), Torsten Schwede (University of Basel) and Ada Yonath (Weizmann Institute of Science). It will also include oral communications from selected contributions, open discussions and posters, as well as tutorials, which will be given on the premises of the University of Torino's Department of Computer Science[3] on Wednesday 15 October.

Submissions are welcome, both for oral communications and for posters. All abstracts must be submitted through the NETTAB 2014 EasyChair submission page[4]. Abstracts for oral communications must not exceed four pages; those for posters must not exceed two pages. All abstracts

1 www.nettab.org/2014/CCY/

2 www.nettab.org/2014/
3 www.di.unito.it/
4 https://www.easychair.org/conferences/?conf=nettab2014

must be structured, and include an Introduction, as well as Methods, Results, Discussion and References sections. Two special issues will appear as supplements of BMC Bioinformatics and of BMC Systems Biology. The related Call for papers will be launched shortly after the workshop.

The workshop is held under the Patronage of the International Society for Computational Biology[5] (ISCB), which has granted the status of ISCB Affiliated Conference to the workshop; the Global Bioinformatics Network EMBnet[6]; the Italian Society of Bioinformatics[7] (BITS); the Polish Bioinformatics Society[8] (PTBI) and the Rete Ligure di Bioinformatica[9] (ReLiB).

5 www.iscb.org/
6 www.embnet.org/
7 www.bioinformatics.it/
8 https://www.linkedin.com/company/polish-bioinformatics-society
9 https://sites.google.com/site/reteliguredibioinformatica

GOBLET: achievements and goals a year on

GOBLET Consortium

GOBLET Stichting, CMBI Radboud University, Nijmegen Medical Centre, Nijmegen, The Netherlands

Received 22 January 2014; **Published** 5 March 2014

Goblet Consortium (2014) *EMBnet.journal* **20**, e751. http://dx.doi.org/10.14806/ej.20.0.751

The idea to create a Global Organisation for Bioinformatics Learning, Education and Training (GOBLET) was formulated during a satellite meeting[1] of the 24th Annual General Meeting (AGM) of EMBnet (the Global Bioinformatics Network) in Uppsala, in June 2012. Here, leaders and representatives of ten international societies, networks and institutes concluded that tangible benefits could be realised if organisations whose core business activities involve bioinformatics education and training could more readily share their experiences, expertise and resources. GOBLET[2] was subsequently established as a Dutch Foundation, and held its first meeting in Amsterdam in November 2012, hosted by The Netherlands Bioinformatics Centre (NBIC) (GOBLET Consortium, 2013). A year on, we report the outcomes of GOBLET's first AGM: we review its principal achievements, and reflect on future priorities, moving forward.

Annual General Meeting, 2013

The GOBLET AGM[3] (GOBLET Consortium, 2013), held at The Genome Analysis Centre (TGAC), Norwich, in November 2013, was attended (or represented) by all 22 eligible organisational members; one individual member also participated. Representatives from the European Bioinformatics Institute, the Wellcome Trust and the Fondazione Edmund Mach observed the meeting. The main goals were I) to report achievements since the kick-off meeting in Amsterdam; II) to announce the results of the first elections; III) to define GOBLET's immediate priorities; and IV) to discuss how the organisation should begin reaching out to the rest of the world.

Executive and Task-Force Reports

During GOBLET's first year, 14 Gold, eight Silver and two Bronze organisational members were welcomed, together with three individual members. Since the AGM, more individuals have joined, and several further organisations have pledged membership (joining has been facilitated by the recent implementation of a PayPal module for payment of fees online). To afford greater flexibility during lean funding periods, GOBLET uses a mixed financial model, including subscription fees, donations, grants, *etc.* Therefore, to supplement income from membership fees, a 'networking' grant was also applied for from the Canadian Institutes of Health Research (CIHR) – this was eventually successful, providing funds to support a future GOBLET workshop in Toronto.

This has been a very busy year, not just in terms of the recruitment of new members (membership has more than doubled since the original meeting in Uppsala) and the first successful grant application, but also in terms of the number of meetings held and attended by GOBLET members: *e.g.*, the first ELIXIR-UK/GOBLET workshop hosted at TGAC, March 2013; the *NextGenBug* meeting at the Roslin Institute, June 2013; the interim GOBLET meeting hosted by the ISCB in Berlin, July 2013; the workshop for e-infrastructure trainers at the Hartree Centre, August 2013; the pan-european bioinformatics training strategy workshop at TGAC, November 2013; and so on. Moreover, in addition to these and the forthcoming Toronto workshop, plans are also in hand to organise a GOBLET workshop alongside the Society for Experimental Biology (SEB)'s 2014 AGM, in Manchester in June.

During the year, two other tangible achievements stand out. First, members of GOBLET worked with the ISCB to create, for the first time, a new track for education posters in the annual ISMB conference – the idea was to provide a forum for exploring different models of and/or vehicles for learning and education in bioinformatics (tutorials, workshops, courses, e-learning, and so on), and how these can be used to enhance the understanding and use of bioinformatics across disparate audiences. GOBLET subsequently had its first poster accepted for ISMB 2013 (Figure 1). Second, GOBLET's training portal[4] was released shortly before the AGM. This is now gaining momentum, and the numbers of uploaded materi-

1 www.mygoblet.org/about-us/goblet-events/inaugural-b3cb-meeting
2 www.mygoblet.org/
3 www.mygoblet.org/about-us/goblet-events/tgacgoblet-meeting
4 www.mygoblet.org/training-portal

als and courses are expanding (it is planned to describe the portal more fully in an article to be published later this year).

Election Results

Following a procedure discussed and agreed during the Berlin meeting, a candidate nomination and election process was conducted online during September and October 2013. The aim was to elect members to the first formal Executive Board, and to the Chair/co-Chair positions of each of the new Committees: I) Learning, Education and Training (LET); II) Outreach and PR; III) Standardisation; IV) Fund-Raising; and V) Technical. The newly elected members (summarised in Table 1) will form GOBLET's first Operational Board, which will henceforth assume responsibility for running the daily business of the Foundation and coordinating its diverse activities.

'state of the field' manuscript from the results. To further support these activities, the Outreach and PR Committee will begin to prepare appropriate materials (brochures, a newsletter, promotional slides, and so on), and will broadcast GOBLET's work using appropriate social media. Acquiring sufficient funds (through grants, subscriptions, sponsorship, *etc.*) will be essential to support this work and to allow GOBLET to achieve its mission – inevitably, this will be the focus of the Fund-Raising Committee.

The outcomes of the break-out sessions also highlighted the need for GOBLET to clearly define its training focus. One concrete suggestion was that the LET Committee should oversee the development of a resource kit for educating the self-taught, to better help GOBLET members to train in their communities.

Table 1. Newly elected members forming GOBLET's first Operational Board.

GOBLET OPERATIONAL BOARD	
Executive Board	**Outreach & PR Committee**
Chair: Terri Attwood	Chair: Erik Bongcam-Rudloff
Vice Chair: Vicky Schneider	**Standardisation Committee**
Secretary: Michelle Brazas	Chair: Pascale Gaudet
Treasurer: Fran Lewitter	**Fund-raising Committee**
Learning, Education & Training	Chair: Patricia Palagi
Chair: Nicky Mulder	**Technical Committee**
Co-Chair: Celia van Gelder	Chair: Manuel Corpas

Defining GOBLET's priorities

During the meeting, several break-out sessions were organised to help elucidate GOBLET's next steps and to inform its outreach strategy. From the discussions, publishing GOBLET's achievements to date and broadening GOBLET's horizons emerged as the most urgent priorities: amongst other things, GOBLET needs to position itself: I) to reach out effectively to all who need bioinformatics training, II) to attract individuals, small groups, students, as well as larger organisations, and III) to promote the importance of bioinformatics training to funding bodies, to grant holders and to universities.

To inform such activities, it was agreed that GOBLET should coordinate a survey (building on the survey conducted by SEB at the beginning of the year), aiming both to give a broader picture of training needs worldwide and to generate a

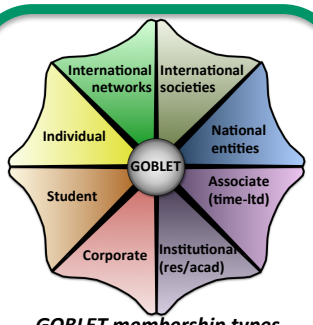
A common theme throughout the discussions was also continued development of the training portal. It was agreed that, above all, this resource needs to be non-redundant and to address real user needs. To this end, collaboration with other organisations will be essential, to identify synergies and to avoid costly duplication of effort – a particular priority for the Technical Committee is therefore to liaise with ELIXIR-UK in order to harmonise GOBLET's training portal with their plans to develop a Training e-Support Service (TeSS).

Reaching out to the rest of the world

It is important for GOBLET to reach out and attract new members for a variety of reasons: to expose new market-places for training; to open up potential new funding routes; to provide lobbying opportunities; to get recognition and buy-in from established professions; and so on. In

Bioinformatics Training and Education:
towards a sustainable global network
www.mygoblet.org

The **Global Organisation for Bioinformatics Learning, Education & Training (GOBLET)** evolved from the BTN - the Bioinformatics Training Network[1]. Arising from a recognised need to coordinate a spectrum of world-wide training activities in bioinformatics, biocuration, biocomputing and computational biology, its members aim to: **share, not duplicate, effort; share, not duplicate, cost; to work together, in a mutually respectful way, towards common solutions and a sustainable future.** Join us at www.mygoblet.org



GOBLET membership types

26 organisations have now joined

GOBLET: a global umbrella organisation for societies, networks & institutions

Steered by an Executive Board (elected by the membership) and five Committees (Education & Training, Outreach & PR, Technology, Standarisation and Fund-raising), GOBLET is about *action* - working together towards pragmatic solutions to common problems[2,3]. Building on the work of the BTN, GOBLET has launched a **Training Portal,** offering a registry of trainers, organisers, events, and materials and documents, providing a **sharing platform** for trainers and trainees alike.

GOBLET worked with ISCB to establish a poster track for education at ISMB conferences. It is also developing standards for disseminating life science events[4], preparing best practice guidelines for bioinformatics training[5], exploring ways to bring ***recognition and accreditation*** to training, and is liaising with ELIXIR to interface with the European research infrastructure for biological data.

1. Schneider MV *et al.* (2012) Bioinformatics Training Network: a community resource for bioinformatics trainers. *Brief.Bioinform.*, **13**, 383-9
2. Via A *et al.* (2011) 10 simple rules for developing a short bioinformatics training course. *PLoS Comput.Biol.*, **7**, e1002245
3. Schneider MV *et al.* (2010) Bioinformatics training: a review of challenges, actions & support requirements. *Brief.Bioinform.*, **11**, 544-51
4. Jimenez RC *et al.* (2013) iAnn: An Event Sharing Platform for the Life Sciences. *Bioinformatics*, 10.1093/bioinformatics/btt306
5. Via A *et al.* (2013) Best Practices in Bioinformatics Training for Life Scientists. *Brief.Bioinform.*, 10.1093/bib/bbt043

GOBLET
Global Organisation for Bioinformatics Learning, Education & Training

Figure 1. GOBLET's first poster, accepted in the new education track for ISMB2013 posters.

return, reciprocal benefits for new members include, amongst many others, the formation of new collaborations, the chance to share best practices (Via *et al.*, 2013) and, importantly, access to the trainers' registry, to a diverse array of bioinformatics training opportunities, and to a well-established bioinformatics trainer community. During the break-out discussions, numerous stake-holder groups were identified as potential targets for future outreach. Forging links with such groups will require the concerted efforts of GOBLET's new Executives and Committee Chairs; but *all* members can continue to play important roles as GOBLET ambassadors.

Staying in touch

Together, we've established the world's first global bioinformatics training organisation as a legal entity – GOBLET. The first AGM was an extremely positive and useful meeting, helping both to outline appropriate strategies to build on the foundations of the last year, and to elicit GOBLET's immediate priorities.

If you'd like to learn more about GOBLET, please visit the website[5], or follow us on Twitter via @mygobletorg. To participate directly in our activities, a range of membership options[6] is now available – we will be happy to welcome new members at any GOBLET events during the year, and especially at the next AGM, now scheduled to take place in Toronto, November 2014, hosted by bioinformatics.ca, with support from the recently awarded funds from the CIHR.

References

GOBLET Consortium. (2013) The Global Organisation for Bioinformatics Learning, Education & Training (GOBLET). *EMBnet.journal* **19**(1), 10-13. http://dx.doi.org/10.14806/ej.19.1.606

Via A, Blicher T, Bongcam-Rudloff E, Brazas MD, Brooksbank C *et al.* (2013) Best Practices in Bioinformatics Training for Life Scientists. *Brief. Bioinform.* **14**(5), 528-537. http://dx.doi.org/10.1093/bib/bbt043

Consortium members

Attwood, T.K.[1], Blackford, S.[2], Brazas, M.D.[3], Brooksbank, C.[4], Budd, A.[5], Corpas, M.[6], Davies, A.[7], Fatumo, S.[8], Fernandes, P.L.[9], Gaudet, P.[10], Hayer, J.[11], Jimenez, R.[4], Korpelainen, E.I.[12], Kumithini, J.[13], Maclean, D.[14], McGrath, A.[15], Orengo, C.[16], Palagi, P.M.[10,17], Ponting, C.[18], Sansone, S.[19], Schneider, M.V.[6], Schönbach, C.[20], Taylor, J.[7], van Gelder, C.W.G.[21,22], Vriend, G.[22]

1. Faculty of Life Sciences and School of Computational Biology, The University of Manchester, Manchester, UK; 2. The Society for Experimental Biology, Lancaster University, Lancaster, UK; 3. The Ontario Institute for Cancer Research, Ontario, Canada; 4. European Bioinformatics Institute, Hinxton, UK; 5. European Molecular Biology Laboratory, Heidelberg, Germany; 6. The Genome Analysis Centre, Norwich Research Park, Norwich, UK; 7. The Nowgen Centre, University of Manchester, UK; 8. Department of Computer and Information Sciences, Covenant University, Ota, Nigeria; 9. Instituto Gulbenkian de Ciência, Oeiras, Portugal; 10. University of Geneva, Geneva, Switzerland; 11. Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden; 12. CSC - IT Center for Science, Espoo, Finland; 13. Centre for Proteomic and Genomic Research, Cape Town, South Africa; 14. The Sainsbury Laboratory, Norwich Research Park, Norwich, UK; 15. CSIRO Computational Informatics, Canberra, Australia; 16. Faculty of Life Sciences, University College London, London, UK; 17. SIB Swiss Institute of Bioinformatics, Switzerland; 18. Computational Genomics Analysis and Training, University of Oxford, Oxford, UK; 19. BioSharing, University of Oxford, Oxford, UK; 20. Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Fukuoka, Japan; 21. Netherlands Bioinformatics Centre, Nijmegen, The Netherlands; 22. Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands.

5 www.mygoblet.org
6 www.mygoblet.org/about-us/membership

eBioKit bioinformatics workshops in Dar es Salaam, Tanzania


Etienne P. de Villiers[1]✉, **Erik Bongcam-Rudloff**[2]

[1]KEMRI, Kenya
[2]SLU-Global Bioinformatics Centre,Department animal Breedings and Genetics, SLU, Uppsala, Sweden

Received 7 February 2014; **Published** 5 March 2014

de Villers EP and Bongcam-Rudloff E (2014) *EMBnet.journal* **20**, e755. http://dx.doi.org/10.14806/ej.20.0.755

In collaboration with H3Abionet[1] and BecA-ILRI Hub[2], two eBioKit-based bioinformatics workshops were held from 10-14 December 2013 in Dar es Salaam, Tanzania. The workshops aimed to strengthen bioinformatics research capacity in Tanzania, and facilitate discussions on bioinformatics tools for diagnostics, pathogen discovery, genome evolution, and other applications relevant to improving health and increasing agricultural productivity in Tanzania.

Experiences gained by Prof. Erik Bongcam-Rudloff at the Swedish University of Agricultural Sciences[3] and Dr. Etienne de Villiers at KEMRI-Wellcome Trust Research Programme[4] in conducting bioinformatics training courses in Kenya, Uganda, Mauritius and Zimbabwe over several years, showed that it was difficult to successfully teach and demonstrate several bioinformatics resources. This was mainly owing to limited network connections and computing infrastructures in these countries. For this reason, a bioinformatics platform, the eBioKit, was engineered to ease the administrative burdens both of installing bioinformatics software and of regularly updating large databases over unreliable and slow network connections. The eBioKit is a self-contained computing platform and database system, containing more than 200 bioinformatics applications, including EMBOSS, NCBI BLAST,

Figure 1. Researchers at MARI during the bioinformatics traing using the eBioKit.

1 www.h3abionet.org
2 hub.africabiosciences.org
3 www.slu.se
4 www.kemri-wellcome.org

Figure 2. Participants to the Bioinformatics training workshop at MUHAS using the eBiokit.

Galaxy, Ensembl database systems, and several other specific crop or organism databases of relevance to African scientists. The eBioKit used in Tanzania was the latest version of the platform, version 3.

The Mikocheni Agricultural Research Institute[5], Tanzania, partnered with SLU and BecA-ILRI Hub to host this one-week intensive workshop, to inspire collaboration between bench scientists and bioinformaticians through hands-on training in sequence data analysis using the eBioKit. Ten researchers from MARI were introduced to the eBioKit's tools and databases. During the workshop, it was agreed that a second workshop with at least 20 participants would be organised during 2014.

The University of Dar es Salaam[6] acquired an eBioKit under the aegis of the H3ABionet project; in collaboration with H3ABionet, we simultaneously organised a second eBioKit bioinformatics workshop at the Muhimbili University of Health and Allied Sciences (MUHAS), Dar es Salaam[7]. This workshop focused on health researchers, and included 25 participants from MUHAS, University of Dar es Salaam, Muhimbili Wellcome Programme, PWANI University, Kenya, and several participants from nodes in H3ABionet.

During the workshop, two individuals were instructed in managing and maintaining the eBioKits that were permanently installed at MARI and PWANI University. These systems are now accessible to researchers and students at these institutions for bioinformatics teaching or research purposes.

The workshops included sessions on basic introduction to the Linux operating system, classical bioinformatics tools in the EMBOSS package, introduction to Next Generation Sequencing and Galaxy system, introduction to biological databases and Ensembl, Genome Wide Association Studies (GWAS) and UGENE, a GUI-based bioinformatics tool for desktop environments. Trainers on the course were Prof. Erik Bongcam-Rudloff (SLU), Dr. Maria Wilbe (SLU), Dr. Juliette Hayer (SLU), Dr. Etienne de Villiers (KWTRP) and Dr. Mark Wamalwa (BecA-ILRI).

Acknowledgments

The workshops were funded by the Swedish Ministry of Foreign Affairs as part of its special allocation on global food security.

5 www.ari-mikocheni.org
6 https//udsm.ac.tz
7 http://www.muchs.ac.tz

BiP-Day 2013: "Prima Giornata della Bioinformatica Pugliese" – Workshop report


Domenica D'Elia✉, **Sabino Liuni**
CNR, Institute for Biomedical Technologies, Bari, Italy

Received 11 March 2014; **Published** 18 March 2014

D'Elia D and Liuni S (2014) *EMBnet.journal* **20**, e758. http://dx.doi.org/10.14806/ej.20.0.758

On 5 December 2013, a regional workshop on *Bioinformatics in Apulia* (*BiP-Day 2013*[1]) was held in Bari (IT) under the patronage of the Italian Bioinformatics Society (BITS[2]) and EMBnet[3].

The aim of the workshop, prompted by the Italian National Node of EMBnet (CNR Institute for Biomedical Technologies[4]) and supported by the InterOmics Flagship project[5], was to stimulate tighter collaboration between life science researchers and private biotech companies in the Apulia Region around cutting-edge topics in biological and clinical research, for which bioinformatics R&D is key.

With the advent of new high-throughput technologies, in particular High-Throughput Next-Generation Sequencing (HT-NGS), the approach to biological and clinical research has completely changed. There is now a pressing need to develop new bioinformatics tools and techniques to allow researchers to more easily handle the avalanche of data produced and, more importantly, to be able to interpret the results in a holistic way. Understanding the complex molecular interactions that modulate gene expression in physiological and pathological conditions is key to discovering the genetic bases of human diseases, and for understanding the contribution of dietary habits and lifestyles on human health and disease onset. In order to improve the ability of research communities to cope with these challenging tasks, multidisciplinary approaches are necessary. Moreover, the role of biotech companies is essential for translating research achievements into the clinical, agri-food and environment domains. All these themes were featured in this first edition of the *BiP-Day 2013* event, viewing them from the perspectives of biologists, physicians, bioinformaticians, computer scientists, physicists, engineers and biotech companies operating in the Apulia Region.

The workshop was organised in collaboration with representatives of the CNR Institutes of Biomembrane and Bioenergetics (IBBE)[6] and Biosciences and BioResources (IBBR)[7], and with the Department of Biosciences, Biotechnologies and Pharmacological Sciences[8] of the University of Bari (IT).

The programme included twenty-five oral presentations from CNR and University research groups, biotech companies and representatives of i) the Regional Agency for Technology and Innovation (ARTI), Dr. Eva Milella (President); ii) the Regional Coordination Office for *Policies for economic development, employment and innovation Services, Industrial Research and Innovation*, Dr. Adriana Agrimi (Regional Executive Officer); and iii) the Apulia Industry Confederation, Dr. Michele Vinci (President of Industrial Confederation Bari-BAT).

Dr. Adriana Agrimi and Eva Milella, illustrated respectively, the Regional 2014-2020 Research & Innovation funding programmes and support actions of ARTI in line with the Horizon 2020 Research & Innovation European Programme. Dr. Michele Vinci illustrated the vision of the Industrial Confederation on the way biotech companies and Regional research groups could synergise efforts on research innovation, technology transfer and training of next-generation life science researchers.

The workshop's programme was structured into three main sessions: 1) Regional development programmes and major infrastructures for Bioinformatics in the Apulia Region; 2) Bioinformatics projects in bio-medicine, biodiversity, agri-food and bioinformatics training programmes; 3) Research & Business: the im-

1 www.ba.itb.cnr.it/bip-day
2 www.bioinformatics.it
3 www.embnet.org
4 www.itb.cnr.it
5 www.interomics.eu/home
6 www.ibbe.cnr.it
7 www.ibbr.cnr.it/ibbr
8 www.uniba.it/ricerca/dipartimenti/indice-dipartimenti-attivi/bioscienze-biotecnologie

Figure 1. Dr. Eva Milella, President of the Regional Agency for Technology and Innovation (ARTI).

portance of communication. Presentations are available from the workshop website associated to the programme[9], and from the *News* section: Presentations[10].

The workshop programme was introduced by Domenica D'Elia, who gave a snapshot of the central place of bioinformatics in biological research in the past 30 years, and its key role in modern biology. She presented EMBnet, the Global Organisation for Bioinformatics Learning, Education & Training (GOBLET[11]) and collaborating European projects (SeqAhead[12]) and AllBio[13]. Participants were invited to showcase their lines of research, their findings and expertise, and also to express their needs and wishes for collaboration and bioinformatics support. Presentations were structured in this way to give participants a global view of human and technological resources operating in the Region, to establish new contacts on the basis of their own scientific interests and needs, to detect common interests and to stimulate discussion around common goals. In order to facilitate contact after the meeting, participating groups were invited to submit their contributions as a card with the following schema: i) who (research group details and contacts); ii) what (research line description); iii) why (research aims & goals); iv) how (methods and technologies used); v) with whom (needs and requests for collaboration). Participants' cards are available from the workshop website at: www.ba.itb.cnr.it/bip-day/category/partecipanti/.

Four companies participated in the event: Eusoft s.r.l.[14], MASMEC S.p.A[15], EXPRIVIA S.p.A[16]and IBM Italia S.p.A[17]; each presented current technological advances in their own *Research & Innovation* sectors that exploit the translational nature of modern biological, biomedical and

9 www.ba.itb.cnr.it/bip-day/programma/
10 www.ba.itb.cnr.it/bip-day/category/presentazioni/page/3
11 www.mygoblet.org
12 www.seqahead.eu
13 www.allbioinformatics.eu/doku.php
14 www.eusoft.it/it
15 www.masmec.org
16 www.exprivia.it/en/home
17 www.ibm.com/it/it

Figure 2. From left to right, Gaetano Scioscia (IBM Italia S.p.A), Massimo Carella (Laboratory of Medical Genetics and Bioinformatics Unit, IRCCS Casa Sollievo della Sofferenza), Graziano Pappadà (EXPRIVIA S.p.A).

pharmaceutical research. Concluding remarks were given by Gaetano Scioscia (IBM Italia S.p.A.), who provided a fascinating overview on The role of Bioinformatics in the perspective of bio-economy[18]. Education and training was included in the programme to stimulate discussion around the importance of 'formal and non-formal' bioinformatics education & training programmes, which is emerging as one of the most pressing needs of the scientific community. Educating the next generation of scientists is one of the most important commitments that bioinformatics communities worldwide have a duty to accomplish in order to allow scientific research to advance in all domains of the life sciences.

With the support of the Flagship project *InterOmics*, a Tutorial-Day[19] on *Tools and methods for the analysis of omics data and biodiversity* was organised in association with the *BiP-Day 2013* workshop, on 6 December. The Tutorial was held at the Department of Physics *Michelangelo Merlin* of the University of Bari, with the collaboration of the BioVel project partners[20]. The Tutorial-Day was heavily over-subscribed and registration had to close early owing to space limitations.

Overall, the workshop attracted more than 110 participants, and we were delighted by the enthusiasm with which this initiative was welcomed. We would like to thank the speakers and all the attendees for their contributions to the success of the workshop. In particular, we would like to thank the trainers on the Tutorial-Day (Monica Santamaria, Bruno Fosso, Saverio Vicario, Balech Bachir, Andreas Gisel, Angelica Tulipano, Flavio Licciulli, Arianna Consiglio) and Giacinto Donvito, the University Department of Physics and the INFN[21] for providing the necessary IT infrastructure and logistics for organising the Tutorial-Day.

Figure 3. The BiP-Day's registration desk.

18 www.ba.itb.cnr.it/bip-day/2013/12/la-bioinformatica-nelle-prospettive-della-bioeconom
19 www.ba.itb.cnr.it/bip-day/tutorial
20 https://www.biovel.eu
21 https://www.ba.infn.it/

InterOmics Tutorial - Tools and methods for the analysis of omics data and biodiversity


Angelica Tulipano✉, **Andreas Gisel**
CNR, Institute for Biomedical Technologies, Bari, Italy

Received 11 March 2014; **Published** 18 March 2014

Tulipano A and Gisel A (2014) *EMBnet.journal* **20**, e759. http://dx.doi.org/10.14806/ej.20.0.759

The CNR Institute for Biomedical Technologies (ITB)[1] in Bari (IT), with support from the Italian *Flagship project InterOmics*[2], organised a *Tutorial-Day*[3] as a satellite event of the *BiP-Day 2013* workshop (see related article in the present volume). The tutorial was organised in three 3-hour events, covering metagenomics, phylogenetics and data analysis of non-coding RNA. The event took place on 6 December 2013 at the Department of Physics *Michelangelo Merlin* of the University of Bari and INFN[4], as their computing infrastructure was used to guarantee the required performance for such data analysis approaches. While the services required for the first two sessions were already hosted on the INFN computer infrastructure, the third session was run on Virtual Machines (VMs). Four VMs with the analysis pipeline pre-installed, each with 16 CPU and 200GB shared memory, were used to serve 40 participants. Giacinto Donvito, from Bari University's Department of Physics continuously monitored the infrastructure to guarantee a flawless service.

The morning session started with a tutorial on the *Classification and quantification of the mi-*

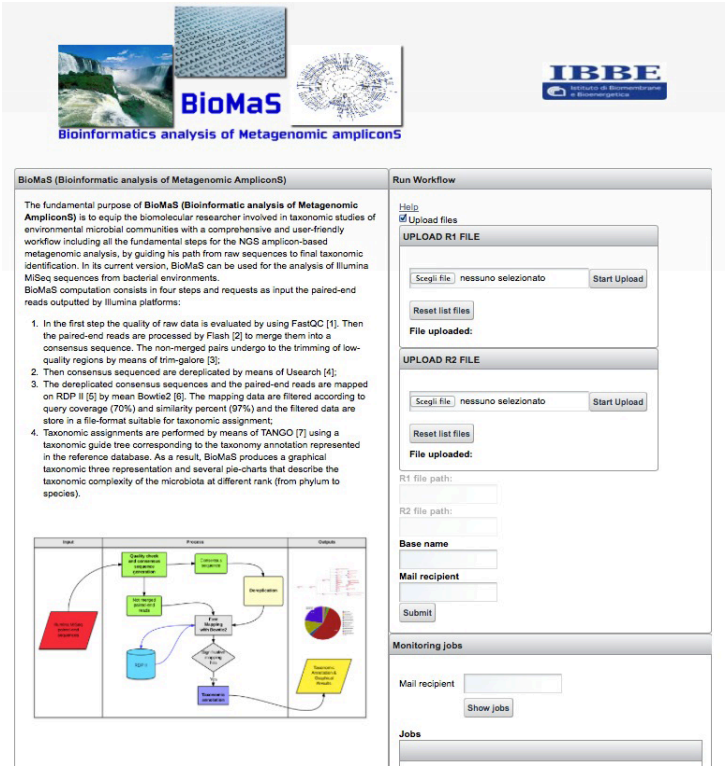


Figure 1. Screenshot from the BioMaS website.

1 www.itb.cnr.it
2 www.interomics.eu
3 www.ba.itb.cnr.it/bip-day/tutorial
4 https://www.ba.infn.it/

crobiome using metagenomic amplicons. Bruno Fosso, from the Department of Biotechnology and Biopharmaceutical Biosciences of the University of Bari (IT), and Monica Santamaria, from the CNR Institute of Biomembranes and Bioenergetics (IBBE)[5], presented a modular pipeline (BioMaS) using third-party tools and ad hoc python and bash scripts. BioMaS is a web-service on the INFN/UNIBA infrastructure (Figure 1). High-level SaaS (Software as a Service) services are applied to facilitate the use of BioMaS components that are already suitably configured and optimised to run on the dedicated infrastructure. BioMaS allows the analysis of both bacterial and fungal environments, and alternative paths can be selected to process data obtained either by Roche 454 or Illumina sequencing technology. The tutorial allowed participants to run a test data-set and understand in detail the pipeline and its functionality.

The second session covered *Instruments for the phylogenetic analysis for studies of biodiversity*. Saverio Vicario, from the CNR – ITB, and Bachir Balech, from the CNR – IBBE, presented the BioVeL[6] infrastructure. This allows users to build customised workflows (Figure 2) by selecting and applying successive 'services', or re-using existing workflows available from BioVeL's library. By giving participants the opportunity to process specially provided test data, the tutorial offered profound insights into BioVeL's significant functionality and performance.

The third session introduced participants to the world of non-coding RNA, in *Mapping and*

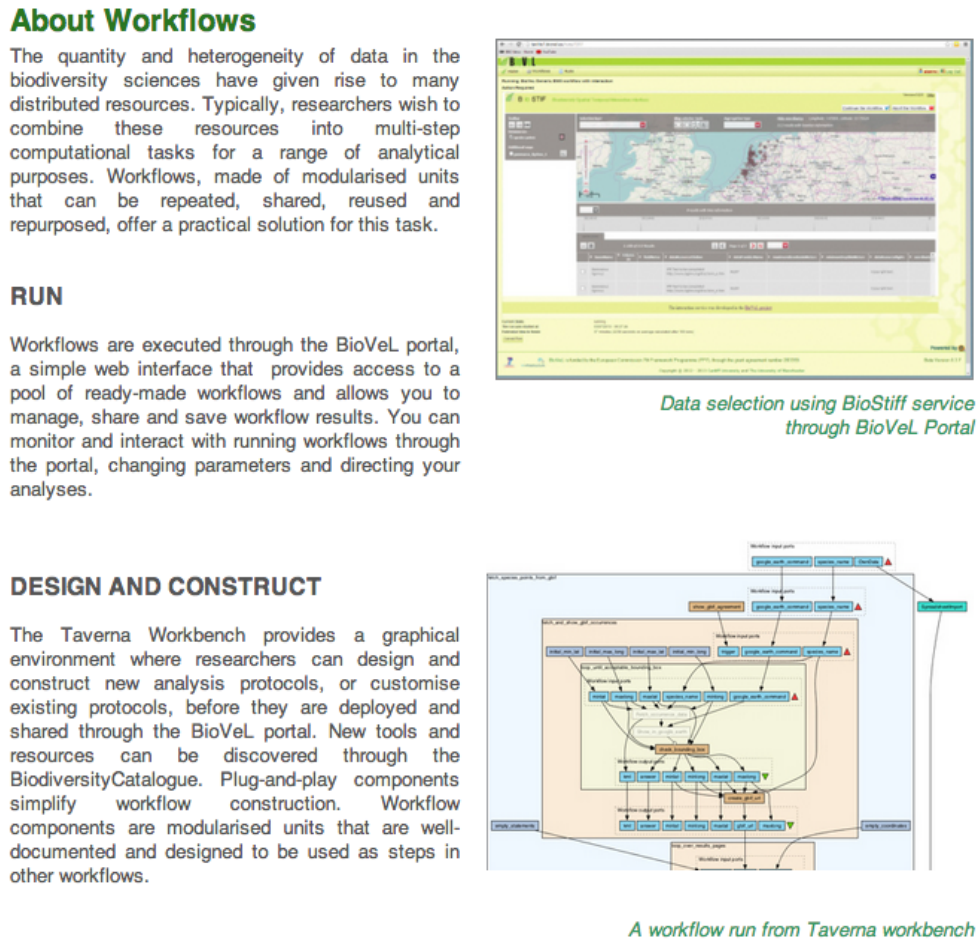


Figure 2. Screenshot from the BioVeL website illustrating its underlying workflows.

5 www.ibbe.cnr.it/

6 www.biovel.org

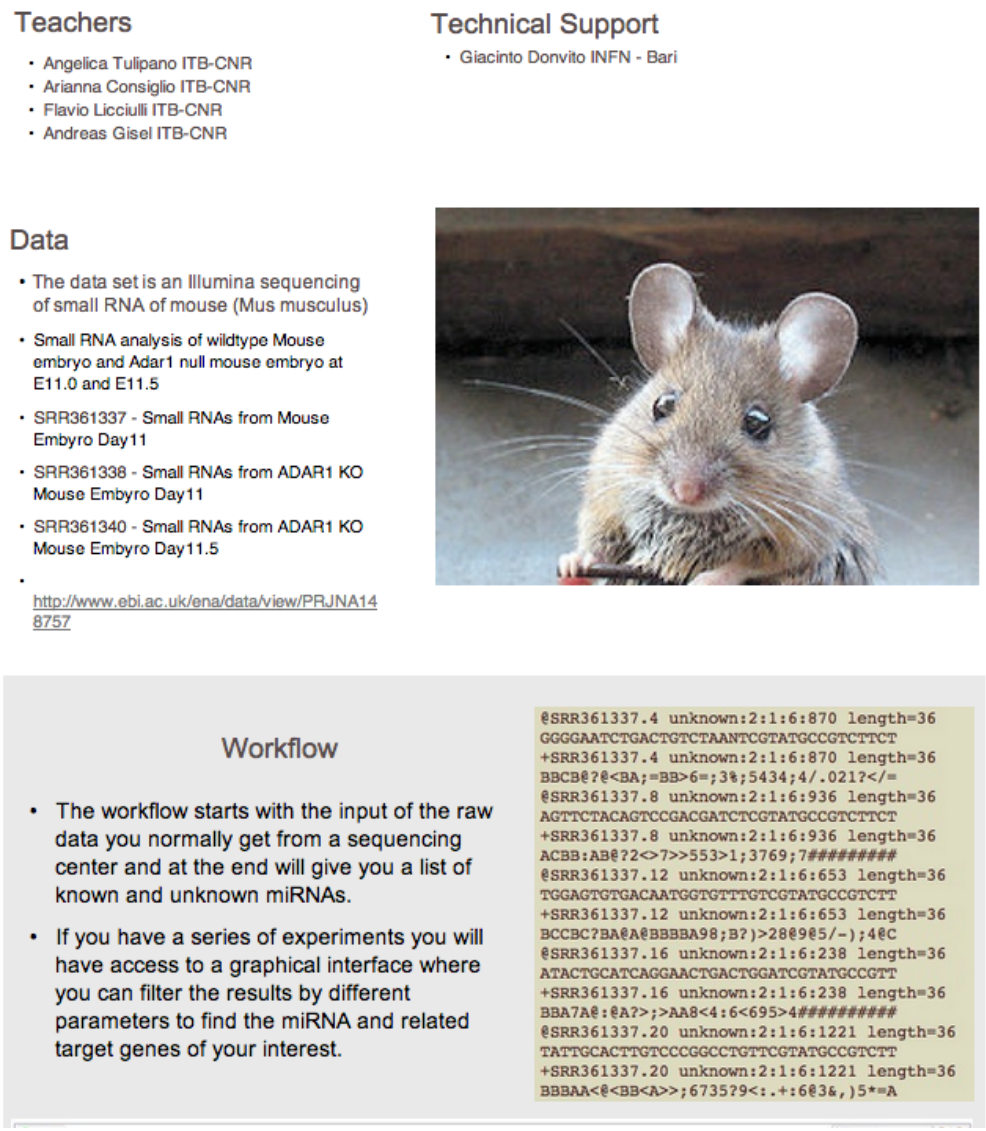


Figure 3. Screenshot from the ncRNA data-analysis website.

analysis of non-coding RNAs and small RNAs from NGS technologies. The specialist team from ITB Bari – Angelica Tulipano, Flavio Licciulli, Arianna Consiglio and Andreas Gisel – demonstrated a simple workflow to get from raw sequencing data to an expression profile of known and unknown miRNA and other non-coding RNAs. The workflow is based on publicly available software and in-house Perl scripts, assembled into a user-friendly pipeline. The results can be uploaded into a MySQL database with a simple graphical interface to visualise, sort and filter the data. A customised data-set of Illumina small RNA sequences, at three time points (Figure 3) , was provided to give the users first-hand experience of the pipeline's functionality.

More than 110 participants (on average 35 per session) attended the tutorials, demonstrating the urgent need for such events to help train life scientists to cope with the large and complex data-sets produced by NGS technologies.

Report on the ALLBIO minisymposium and workshop: "Next Generation Sequencing (NGS) methods for identification of mutations and large structural variants"



Laurent Falquet[1]✉, Grégoire Rossier[2], Tiffanie Yael Maoz[3]

[1]University of Fribourg and Swiss Institute of Bioinformatics, Biochemistry Unit, Fribourg, Switzerland
[2]Vital-IT, Training & Outreach, Swiss Institute of Bioinformatics, Lausanne, Switzerland
[3]Weizmann Institute of Science, Lab of Prof. Avi Levy, Rehovot, Israel

Received 10 April 2014; **Published** 14 April 2014

Falquet L *et al.* (2014) *EMBnet.journal* **20**, e766. http://dx.doi.org/10.14806/ej.20.0.766

This workshop was organised as one of the validation training workshops of the AllBio FP7 Coordination Action. The AllBio project[1] aims to transfer human genome-oriented bioinformatics methods to non-model organisms.

Following a first round of test-case proposals from all over Europe, a selection of 15 test- cases was presented and discussed in detail during an initial workshop in December 2012 in Milano[2], Italy, organised by Dr. Andreas Gisel, CNR Institute for Biomedical Technologies of Bari (IT). From this event, seven test-cases were selected for 'hackathon' sessions, where real data were analysed jointly by computer scientists, bioinformaticians and biologists.

Our test-case for identification of large structural variants (insertions, deletions, inversion, translocations, *etc…*) immediately attracted a lot of interest, given the current difficulty in predicting large structural variants (SVs) in all organisms, but in particular in polyploid genomes, as in plants, and chimeric genomes, as in cancer. The first hackathon session was planned during the preparatory phase, using regular video meetings to define the goals and discuss ideas on how to solve them. It was decided that a Virtual Machine (VM) would be set up to provide a standardised platform as a benchmark method to evaluate and compare all tools. A small group of participants, led by Dr. Yael Maoz (Weizmann Institute, IL), met in March 2013 in Amsterdam (NL) with the support of the local organiser, Prof. Gert Vriend, and the computational support of SURFsara[3], for an intensive hackathon session where one biologist, three bioinformaticians and four computer scientists met to solve a real case. The outcome was a preliminary version of a VM, hosting many tools and a benchmark data-set. During the following months, regular video meetings allowed the participants to combine their expertise in order to improve this first version of the VM.

In August 2013, a second hackathon session was hosted in Nijmegen (NL) and allowed testing of a merging tool that combines results of individual structural-variation prediction tools in a hierarchical manner, aiming to reduce false-positive calls. The results being promising, a publication and a validation workshop were planned for March 2014, in Lausanne, Switzerland.

The workshop was expanded to include a mini-symposium on the first day, with eight invited speakers and more than 75 participants from all over Europe (Figure 1). After the usual welcome address by the local organisers, Yael Maoz detailed the work of the hackathon team and the proposed outcomes of the AllBio project:

- creating an automated standardised pipeline for testing new tools and/or for testing existing tools on non-model organisms;
- identifying the best tool(s) for SV prediction through benchmarking;
- providing a statistically sound method of merging SV calls;
- a tool called SV-Autopilot: Structural Variation AUTOmated PIpeLine Optimisation Tool (submitted for publication).

Alexandre Reymond gave an amazing talk on the role of large SVs in human, showing the balancing effect of deletion and duplication of the same genomic location on diseases such as autism and schizophrenia, and their links with obesity (Zufferey *et al.*, 2012).

1 www.allbioinformatics.eu/doku.php
2 www.allbioinformatics.eu/doku.php?id=public:bioinf
3 https://www.surfsara.nl/

Figure 1. Participants at the mini-symposium.

Tobias Rausch presented the Delly software tool (Rausch *et al.*, 2012), and the various uses in many large-scale population genomics analyses. He also mentioned that inversions are usually not found alone: in more than 60% of cases, inversions are associated with deletions or duplications.

Bart Deplancke gave a brilliant talk on the analysis of large genomic variants in *Drosophila* lines identified using PrInSeS (Massouras *et al.*, 2010), and their effects (Massouras *et al.*, 2012).

Tobias Marshall gave a detailed talk on MATE-CLEVER, an extension of CLEVER (Marschall *et al.*, 2013) used to discover 'twilight zone' variants and their genotypes.

Valentina Boeva showed various tools to discover large SVs in cancer cell lines (Boeva *et al.*, 2013).

Yogesh Paudel introduced copy-number variation methodology used to analyse domestication events in pig lines (Paudel *et al.*, 2013).

Can Alkan provided a detailed account of the characterisation of mobile element insertions in humans and apes (Hormozdiari *et al.*, 2013).

The hands-on workshop on the second day joined 30 participants and six speakers. The attendance would have been larger, but was limited to this number for practical purposes. Participants were asked to download and install a pre-configured Ubuntu Virtual Machine using VirtualBox software on their laptops.

Yael Maoz explained the concepts of the project and of the hackathon sessions, while building and use of the VM was detailed by Wai Yi Leung (Leiden University Medical Center). Participants were able to test the VM on a subset of the original data only, as the whole genome analysis would run for too long on a laptop. The results were then detailed and discussed by Tobias Marschall and Yael Maoz. Visualisation of the results with the Integrative Genomic Viewer (IGV) was presented by Laurent Falquet and Yael Maoz (Figure 2).

In the afternoon, the participants were divided into two groups: those wishing to analyse their own data, and those wishing to discuss issues and solutions for detecting large SVs in planned projects. The workshop ended with a summary given by Yael Maoz.

In conclusion, this workshop validates the outcome of our test-case dealing with large SVs. We have shown that a small group of volunteers working on a part-time basis can develop new methods and tools. We believe that the SV-AUTOPILOT VM that we developed will be very useful both for biologists looking to get the best variant predic-

Figure 2. Hands-on workshop. In this picture, Laurent Falquet illustrates IGV.

tions, and for bioinformaticians seeking to evaluate their software performance against existing tools.

Acknowledgements

This work was supported by the AllBio FP7 work programme KBBE-2011-5-289452 «FOOD, AGRICULTURE AND FISHERIES, AND BIOTECHNOLOGY».

We thank the COST Action, SeqAhead, for supporting travel grants of participants and providing local organiser support, and the Vital-IT group of the SIB Swiss Institute of Bioinformatics, for local organisation.

References

Boeva V, Jouannet S, Daveau R, Combaret V, Pierre-Eugène C *et al.* (2013) Breakpoint features of genomic rearrangements in neuroblastoma with unbalanced translocations and chromothripsis. *PloS One* **8**, e72182. http://dx.doi.org/10.1371/journal.pone.0072182

Hormozdiari F, Konkel MK, Prado-Martinez J, Chiatante G, Herraez IH *et al.* (2013) Rates and patterns of great ape retrotransposition. *Proc Natl Acad Sci U S A* **110**, 13457–13462. http://dx.doi.org/10.1073/pnas.1310914110

Marschall T, Hajirasouliha I, Schönhuth A (2013) MATE-CLEVER: Mendelian-inheritance-aware discovery and genotyping of midsize and long indels. *Bioinformatics* **29**, 3143–3150. http://dx.doi.org/10.1093/bioinformatics/btt556

Massouras A, Hens K, Gubelmann C, Uplekar S, Decouttere F *et al.* (2010) Primer-initiated sequence synthesis to detect and assemble structural variants. *Nat Methods* **7**, 485–486. http://dx.doi.org/10.1038/nmeth.f.308

Massouras A, Waszak SM, Albarca-Aguilera M, Hens K, Holcombe W *et al.* (2012) Genomic variation and its impact on gene expression in Drosophila melanogaster. *PloS Genet* **8**, e1003055. http://dx.doi.org/10.1371/journal.pgen.1003055

Paudel Y, Madsen O, Megens H-J, Frantz LAF, Bosse M *et al.* (2013) Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. *BMC Genomics* **14**, 449. http://dx.doi.org/10.1186/1471-2164-14-449

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V *et al.* (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339. http://dx.doi.org/10.1093/bioinformatics/bts378

Zufferey F, Sherr EH, Beckmann ND, Hanson E, Maillard AM *et al.* (2012) A 600 kb deletion syndrome at 16p11.2 leads to energy imbalance and neuropsychiatric disorders. *J Med Genet* **49**, 660–668. http://dx.doi.org/10.1136/jmedgenet-2012-101203

EMBnet, the Global Bioinformatics Network: a report on the workshop and 26th AGM, Lyon, May 2014

Teresa K. Attwood

University of Manchester, Manchester, United Kingdom

Received 26 July 2014; **Published** 24 September 2014

Attwood TK (2014) *EMBnet.journal* **20**, e786. http://dx.doi.org/10.14806/ej.20.0.786

Introduction

EMBnet's 2014 Annual General Meeting (AGM) and associated events were hosted in Lyon (FR), from 26 to 30 May. This 26th formal meeting of EMBnet provided an opportunity both to partner with the *Bioinformatics for Environmental Genomics* workshop of the Pluridisciplinary Thematic Network in Environmental Genomics, and to review progress since the silver anniversary meeting in Valencia (ES) last year.

The occasion of the AGM included several events: i) a one-day EMBnet tutorial at the Pôle Rhône-Alpes de Bioinformatique (PRABI) on the Doua campus of the University of Lyon 1, entitled *From NGS data through the third dimension towards new agrochemicals and drugs* – this included sessions on RNA-seq analysis, tutored by Vincent Lacroix and Vincent Navratil; the Hope protein structure-function analysis suite, led by Gert Vriend; and the STING platform, demonstrated by Goran Neshich; ii) the two-day *Bioinformatics for Environmental Genomics* workshop, which took place in the École Supérieure de Chimie Physique Électronique de Lyon, with 200 participants, also on the Doua campus; iii) a one-day EMBnet workshop held (partly) in Lyon's Hôtel de la Cité; and finally, iv) the traditional business meeting, also held in the Hôtel de la Cité.

The EMBnet workshop got off to an interesting start. A major fault with the hotel's fire-alarms forced us to abandon the meeting room and to transfer to the lobby/bar area to begin the business of the day (in doing so, we had to turn the agenda upside-down, beginning with round-table discussions and leaving formal reports for later); after mid-morning coffee, we transferred to a meeting room in an adjacent hotel, where we continued with the Executive Board (EB) and Committee Chair reports; we then returned to the Hôtel de la Cité to pick up the pieces of the now-dishevelled agenda after lunch! Despite the disruptions, we nevertheless managed to have some very constructive discussions, with concrete outcomes for the future.

The past year has been another busy one, much of our time being devoted to supporting our allied research projects, initiatives and affiliates (SeqAhead, AllBio, GOBLET, ISCB), alongside the routine work of running the Stichting. In this report, we begin by reviewing some of the year's principal efforts to build on these initiatives and affiliations, we make a critical analysis of EMBnet's current status, and we conclude with a summary of the workshop's main conclusions.

Activities and achievements of the last year

During the last year, our efforts have been concentrated in three main areas: education and training; allied research projects; and outreach/dissemination.

Education and training

Members of EMBnet have organised, participated in and/or presented EMBnet at, a range of bioinformatics courses and summer schools. These include, but are not limited to, the EBI-Wellcome Trust Bioinformatics Summer School in Hinxton, UK (June 2013); the NGS bioinformatics course at KEMRI-Wellcome Trust in Kilifi, KE (August 2013); and, of course, the AGM tutorial (May 2014).

During the 2013 EMBnet workshop, we had agreed to devote more time to producing QuickGuides. Accordingly, two new guides were produced, one on amino acids, the other on the semi-empirical quantum mechanics software package, MOPAC, both of which have been published (see Figure 1); several others are currently in production – please visit the website for the latest information on all other published QuickGuides[1].

Overall, EMBnet's training strategy has been dominated by our leadership of the Global

1 www.embnet.org/embnet-quickguides

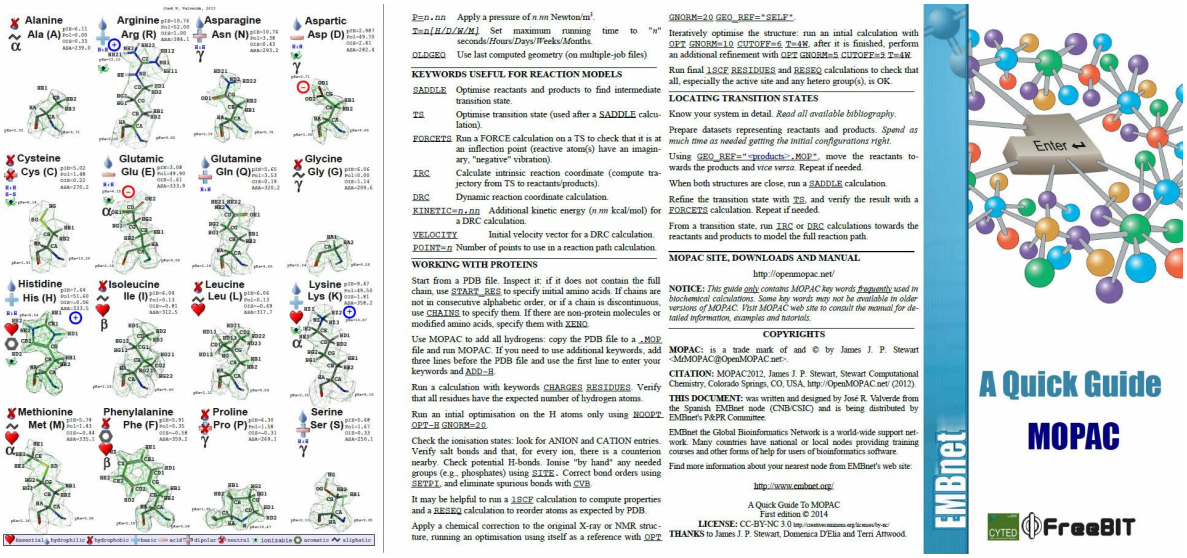


Figure 1. Illustration of panels from the Amino Acid and MOPAC QuickGuides.

Organisation for Bioinformatics Learning, Education and Training (GOBLET[2]). GOBLET, which, at the time of writing, has ~30 members, held its first formal AGM at The Genome Analysis Centre (TGAC), Norwich (UK) in November 2013, a year after its kick-off meeting in Amsterdam (NL). This event took place the day after a pan-European bioinformatics training strategy meeting (also at TGAC): this had been organised largely to be able to discuss how GOBLET and ELIXIR[3] could work together in mutually supportive ways to best serve their communities in future, to build on each other's strengths and obviate unnecessary duplication of effort.

Working through GOBLET has significantly increased our interactions and cooperation with several major international societies and networks, especially with the International Society for Computational Biology (ISCB). Notable highlights of the latter include i) establishment of the ISMB conference's first education poster track (July 2013); ii) development of a Community of Special Interest (COSI) around Computational Biology Education (CoBE), to be launched at ISMB 2014 (Boston, USA); iii) hosting a GOBLET interim meeting alongside ISMB 2013 (see Figure 2), which included stimulating presentations from Lonnie Welch (on curriculum guidelines for bioinformatics and computational biology), from Ana

Figure 2. From left to right: participants of the interim GOBLET meeting held in Berlin (July2013), and of the first formal GOBLET AGM held at TGAC, Norwich (November 2013).

2 www.mygoblet.org
3 www.elixir-europe.org

Conesa (on multidisciplinary, multi-institutional PhD programmes), from Anupama Jigisha (on the ISCB Student Council intern initiative), and from Niklas Blomberg (on how ELIXIR will safeguard life science research in Europe); and iv) preparing a GOBLET feature article for the ISCB newsletter.

Allied research projects

During the year, a significant amount of time was also devoted to working with our allied EU-funded projects. Specifically, members of EMBnet helped organise and/or attended a range of AllBio events[4]: these included a number of 'hackathons' (August 2013, Nijmegen, NL; September 2013, Amsterdam, NL; October 2013, Uppsala, SE; November 2013, Alnarp, SE), an *RNA-seq data analysis workshop* (January 2014, Espoo, FI), the *AllBio/EMBRACE metagenomics interoperability workshop* (April 2014, Amsterdam, NL), and the AllBio AGM (September 2013, London, UK).

Within SeqAhead[5], members of EMBnet were involved in events covering a range of problems in NGS data analysis: these included the *NGS data and the Variation Calling Challenge* (May 2013, Udine, IT); Hadoop and *NGS data processing hackathon III* (June 2013, Pula, IT); workshops on the *Future demands and challenges in ICT and bioinformatics tools for NGS* (June 2013, Pula, IT), *NGS methods for identification of mutations and large structural variants* (March 2014, Lausanne, CH) and *Assessment of training methods in NGS data analysis* (March 2014, Oeiras, PT); and the *NGS data after the Gold rush* workshop and Management Committee Meeting (May 2014, Norwich, UK) – see Figure 3.

Outreach/dissemination

Throughout the year, we have described these and other activities in the monthly *EMBnet.digest* and in *EMBnet.journal*. For example, we made a special report on EMBnet's silver anniversary AGM in the May 2013 digest, we provided a round-up of the year's activities in the December issue, and we provided a report on our work with GOBLET, and invited EMBnet to participate in a GOBLET survey in February's digest. The most notable change for the digest has been a new, much simpler look-and-feel, starting from the January 2014 issue, as illustrated in Figure 4.

For the Journal, the most notable change has been a move to an instant access model, in which articles are published as soon as peer-review and layout have been completed – articles will henceforth be collated into volumes only once a year, commencing with volume 20.

Figure 3. Participants of the SeqAhead NGS data after the Gold rush workshop held at TGAC, UK (May 2014).

4 www.allbioinformatics.eu/doku.php?id=public:events
5 www.seqahead.eu

Figure 4. From left to right: screen-grabs illustrating the old- and new-look EMBnet.digest, respectively.

During 2013, as illustrated in Figure 5, the bulk of the journal work involved preparation of volume 19(1) (which included reports from the EMBnet 2013 AGM, held 17-18 May in Valencia, ES) and proceedings of the *The Next NGS Challenge Conference: Data Processing and Integration* conference, held 14-16 May in Valencia, and of the NETTAB 2013 workshop on *Semantic, Social and Mobile Applications for Bioinformatics and Biomedical Laboratories*, held 16-18 October in the Venice Lido (IT).

Structural changes

Following the successful ratification of the new statutes, which became legally binding in April 2013, we have been formally able to accept individual members, a facility made possible by implementation of the online fee-payment module; accordingly, we were able to welcome a new individual member to the AGM, Axel Thieffry, who gave an entertaining talk by way of introduction. The new statutes also ushered in changes to the internal structure of the organisation. In particular, Committee Chairs now have the flexibility to convene their own working groups and task-forces, without having to proceed via cumbersome election processes.

EMBnet's 2014 workshop and AGM provided opportunities to critically assess how some of these changes have been working in practice. The events also offered a chance to discuss ways of building on some of these initiatives, of kick-

Figure 5. Covers of the principal EMBnet.journal publications from the last and current year.

starting new projects and of seeking joint funding opportunities to support them.

A critical review

Clearly, a lot has been done during the last year, but it's also evident that this is not a time to relax – with greater engagement of EMBnet members, with more coherent leadership of the Committees and more inclusive approaches by the Executive Board (EB), EMBnet could have achieved more, and could achieve a lot more in future.

A brief review of the Committees suggested that only one of these was highly active (the Publicity and Public Relations Project Committee (P&PR PC)), but that this was largely the work of one individual – it was therefore strongly recommended to delegate more of the work of this Committee to others (and EMBnet members are strongly encouraged to lend their support). At the other extreme, the Education and Training Project Committee (E&T PC) seemed largely dormant, with no Committee members being indicated on the website, and hence presumably no meetings having taken place – it was therefore strongly recommended to revitalise this Committee, to recruit members and list them on the website, and to deliver some concrete results during the coming year. The Technical Management Committee (TM PC) sat somewhere in the middle in terms of activity, but lacked coherence (*e.g.*, website maintenance was largely undertaken outside the Committee, and was hence not included as part of the annual report) – it was therefore strongly recommended to revisit the structure of this Committee and to better define its role and responsibilities. Finally, it was felt that the EB could provide more guidance to the Committee Chairs via the Operational Board Meetings, but importantly also that it should run more Virtual General Meetings (VGMs) with the full EMBnet constituency. Although these have been dogged by technical problems in the past, Adobe Connect seems to work reasonably well. It was therefore recommended to re-establish regular VGMs to reach out to, to better engage with and to better inform EMBnet members of the work of the Committees and of the EB.

Overall, then, it was agreed that the Committee Chairs and the EB should review their structure and membership; that they should begin seriously to delegate tasks and/or to recruit new members, as appropriate; that they should update the website with new information; and, most importantly of all, that they should meet on a regular basis in order to be able to deliver, and ultimately report on, real, demonstrable, tangible outcomes.

Of relevance to this critical review (albeit discussed during the AGM rather than the workshop), was an analysis of the EMBnet tutorial. A number of concrete conclusions were drawn from the experience of running this event; if properly implemented, it was felt that these could improve the professionalism and value of future tutorials. Specifically, it was agreed that the E&T PC should lead the development of a core tutorial programme, together with a set of ground rules for running and hosting these events: these would include consideration of a range of aspects, such as the capacity of the room (*e.g.*, up to a recommended maximum of ~30 participants); availability of desktops (or laptops, if participants bring their own); trouble-shooting the local infrastructure in advance; inviting guest speakers (with a budget set aside for this); placing the tutorial closer to the AGM, in order to provide better continuity between these events; and so on (it was noted that many of the suggestions made here had already been published in recent years by the Bioinformatics Training Network – *e.g.*, see Schneider *et al.*, 2010; Via *et al.*, 2011; Via *et al.*, 2013 – whose recommendations should ideally form the bedrock of future tutorial organisation). Ultimately, it was suggested that EMBnet should aim to develop an expanded tutorial, up to two days in length, and to apply a small fee so that the event covers its costs, or better, brings some small level of income back into EMBnet.

New initiatives

Perennial challenges for EMBnet, as with pretty much all professional networks and societies, are how to engage with existing members and how to attract new ones. Key to addressing these issues are the need to present both a range of active projects with which, given the opportunity, members could become involved, and a tangible set of benefits associated with their involvement. Discussion of these points, and especially how to expand EMBnet's membership, focused once again on potential target groups. One of these is the LinkedIn EMBnet group, members of which could potentially be encouraged to join

EMBnet. However, it was clear that appropriate incentives would need to be in place if we were to be successful in stimulating greater interest in and engagement with EMBnet's work. To this end, two new proposals were outlined: i) EMBnet Fellowships; ii) EMBnet Awards.

EMBnet Fellowship Programme

The proposal here was for EMBnet to inaugurate a new 'Fellowship Programme' to fund researchers in exchange for their expertise and support. Broadly speaking, the idea of the Programme would be to encourage Fellows both to build on EMBnet's work and to develop their own interests in areas such as bioinformatics resource development, policy development, capacity building and bioinformatics training.

The idea is that two prestigious 18-month Fellowships would be awarded during the next year, in which successful applicants would be allocated €2,000 to support activities that are *mutually beneficial for the Fellow and for EMBnet*. The funding is intended to be flexible in order to encompass a range of activities: to develop bioinformatics tools and resources, to conduct surveys, to run workshops, to host training events, and so on. In the first call, priority would be given to proposals that specifically identify collaborative projects with EMBnet's Executive Board, with its Committees or with its publications (*EMBnet.journal, EMBnet.digest, EMBnet QuickGuides*).

Again, for maximum flexibility, applicants would be eligible from all ages and career stages (from students, to early stage researchers and principal investigators). Although the Programme would not be open to Organisational Members (Node Managers), Individual Members would be strongly encouraged to apply. At the end of the Fellowship, successful candidates would be invited to present their work during the next AGM, to submit a report or article for publication in *EMBnet.journal*, and an executive summary both for announcement in the 'In Focus' section of *EMBnet.digest* and for publication on the website.

It was generally agreed that this would be a good initiative to pursue. A range of new Web pages would need to be established (to promote the call for proposals, to celebrate the winners and to promote their work), a new Fellowship email list would need to be set up, and a review panel would need to be instantiated in order to be able to launch the initiative. It was agreed that the EB, with support from the TMPC and P&PR PC, should progress this as soon as possible.

EMBnet Service Awards

In addition to encouraging new members to become involved in its work, EMBnet would also like to reward existing members for their dedication both to helping EMBnet's growth and development, and to assuring its continued position as an important global bioinformatics network. It was therefore proposed that an EMBnet Service Award could be given to an EMBnet member (Organisational or Individual) who had contributed significant effort to the success of EMBnet. The activities for receiving this award could be as varied as publishing an important work that mentions EMBnet, running an outstanding education/training event under EMBnet's patronage, developing new EMBnet-branded educational materials (*e.g.*, QuickGuides, tutorials) or tools, supporting/enhancing EMBnet's website or publications (*EMBnet.journal*, *EMBnet.digest*, *etc.*), bringing in new members or creating effective initiatives to help do so, establishing new synergies or collaborations, and so on.

The procedure would be to open an annual call to EMBnet members to nominate recipients of the award, with a short statement describing the nominee's contribution and hence why the award was deserved. In line with EMBnet's normal voting procedures, a majority of votes would be required for the award to be made – hence, the award might not be made in some years.

The award would consist of a certificate and modest sum of money or other token (details to be agreed), which would be presented by the nominator (or by the EB) during the next AGM. Here, the awardee would have the opportunity to give a short presentation on the work for which he or she received the award.

Although it seemed that this would be a good initiative to pursue, it was felt that the Fellowship Programme should take priority. Once again, if and when this were to go ahead, a range of new Web pages would need to be set up (to promote the call for nominations and to announce the winner); details of the award itself would also need to be confirmed prior to launching the initiative.

New proposals, affiliations, collaborations

In previous meetings, time has been set aside to reflect on the future of EMBnet, and the impact of globalisation on our formerly European organisation. This year was no exception. Although, originally, we were to have had a structured session at this point on the agenda, the rather unusual circumstances at the hotel obliged us to convene in the bar. In this 'relaxed' setting, we discussed the relationships between members of EMBnet and projects such as SeqAhead and AllBio, and with organisations like GOBLET and ISCB. There was a general sense that it would be advantageous for EMBnet to focus at least part of its future work around a common theme or project, which would allow us to seek funding in a more inclusive way (experience has shown that relying exclusively on European funds can be divisive in a global organisation). Given that members of EMBnet have very different research interests, it was considered unlikely that a common project could be identified; however, it was recognised that education and training (including sharing bioinformatics knowledge, bioinformatics capacity building, disseminating best practices, curriculum development, *etc.*) was a theme that cuts across all research niches.

This conclusion resonated strongly with the outcome of the survey conducted earlier in the year by the P&PR PC, which found that amongst the principal core values of EMBnet are bioinformatics education/training, networking and capacity building. This naturally led to a discussion focused around the development of bioinformatics curricula and train-the-trainer programmes, and recognition that these are currently 'hot topics' in countries across the world (*e.g.*, initiatives are currently being driven by the H3ABioNet[6] bioinformatics curriculum degree development task-force, the ISCB bioinformatics curriculum development task-force, the GOBLET Learning, Education and Training Committee, and ELIXIR-UK). To capitalise on this momentum, and to draw from the combined experiences of members of EMBnet, GOBLET, SeqAhead and AllBio in particular, it was agreed to organise a meeting to discuss development of global bioinformatics MSc curricula (ideally, in late September 2014, alongside the final AllBio AGM). Results from this meeting, if co-funded and co-organised, could be considered a joint outcome of AllBio and EMBnet, ultimately to be taken forward by EMBnet/GOBLET after the end of AllBio.

Summary of outcomes and actions

Discussions amongst the participants of the workshop and AGM (see Figure 6) were wide-ranging and remarkably productive. Several outcomes and actions were agreed, as follows:

i) each Committee Chair and the EB should review their structure and membership, to delegate tasks and/or to recruit new members as appropriate, to update the website accordingly, to meet regularly and to deliver, and subsequently report on, tangible outcomes;

ii) to take the *Fellowship Programme* forward: the EB will need to develop the details and circulate documentation to all – together with the TMPC, a page will need be set up on the website to publish the new programme, which would need to be advertised/promoted with the support of the P&PR PC;

iii) the EB should formulate the details of a Service Award scheme, but should progress this only *after* the *Fellowship Programme* has been established;

iv) to organise a meeting, late September 2014, ideally alongside the final AllBio AGM, to discuss development of global bioinformatics MSc curricula (the results could be considered a joint outcome of AllBio and EMBnet);

v) to target and invite LinkedIn EMBnet group members to become EMBnet members, using the *Fellowship Programme* as an incentive, once this has been properly launched;

vi) as part of i), ii) and v), Committee Chairs should publish on the website their current working programs and their future projects, not just to showcase their work, but also to provide potential focal points for Fellowship applications; and

vii) the E&T PC should help to develop both a core for future tutorial programmes, and a set of ground rules for running and hosting these events (including aspects such as the capacity of the room - *e.g.*, max. ~30 participants - availability of desktops, trouble-shooting the local infrastructure in advance, inviting guest speakers, placing the event closer to the AGM, *etc.*); ultimately, to develop an expanded tutorial, of up to two days,

6 h3abionet.org

Figure 6. Setting up for business, AGM 2014.

with a small fee so that the event covers its costs, or better, brings income to EMBnet.

Conclusion

It's clear that there are many opportunities ahead for EMBnet, and this meeting provided a timely opportunity to outline plans for taking some of these forward. During the 2014 meeting, we were fortunate to be able to celebrate the arrival of our latest individual member, Axel Thieffry, whom we hope will be willing to take an active role in helping us to drive some of these new plans forward, and especially to help inspire and recruit more eager and talented individual members like himself!

As always, there's still a lot more work to do. We therefore encourage you all to contribute your energies and visions to EMBnet, to ensure EMBnet's continued success as the Global Bioinformatics Network!

Acknowledgements

We are grateful to Guy Perriere for his work in organising these events, and especially for mediating with the hotel management to secure a new meeting room under rather unusual (not to mention, stressful) circumstances! This year, we would also again like to thank Domenica D'Elia for her energy and fortitude in coordinating and sustaining the many successful activities of the P&PR PC; Rafael Jimenez for his ongoing technical support of the website; and Lubos Klucar for his patient and consistent work in managing the production of *EMBnet.journal*.

T.K.Attwood
On behalf of the Executive Board

References

Schneider MV, Watson J, Attwood TK, Rother K, Budd A *et al.* (2010) Bioinformatics training: a review of challenges, actions and support requirements. Brief. Bioinform. **11**(6), 544-551. http://dx.doi.org/10.1093/bib/bbq021.

Via A, Blicher T, Bongcam-Rudloff E, Brazas MD, Brooksbank C *et al.* (2013) Best Practices in Bioinformatics Training for Life Scientists. Brief. Bioinform. **14**(5), 528-537. http://dx.doi.org/10.1093/bib/bbt043

Via A, De Las Rivas J, Attwood TK, Landsman D, Brazas MD *et al.* (2011) Ten simple rules for developing a short bioinformatics training course. PLoS Comput. Biol. **7**, e1002245. http://dx.doi.org/10.1371/journal.pcbi.1002245

2014 Annual General Meeting – Executive Board Report





Teresa K. Attwood[1]✉, Andreas Gisel[2], Etienne de Villiers[3], Erik Bongcam-Rudloff[4], Goran Neshich[5]

[1]University of Manchester, Manchester, United Kingdom
[2]International Institute of Tropical Agriculture, Ibadan, Nigeria
[3]Kenya Medical Research Institute (KEMRI), Kenya
[4]Swedish University of Agricultural Sciences, Uppsala, Sweden
[5]EMBRAPA, Brazil

Received 19 December 2014; **Published** 20 January 2015

Attwood TK *et al.* (2014) *EMBnet.journal* **20**, e798. http://dx.doi.org/10.14806/ej.20.0.798

During the past year, the Executive Board (EB) has continued to meet on a regular basis, has held regular meetings with the Operational Board (OB) and has convened additional meetings open to the full EMBnet constituency, either via Skype or using Adobe Connect for larger meetings. These meetings have allowed us to discuss a range of issues relating to the work of the Project Committees (PCs), to *EMBnet.journal*, to the website, to the Stichting accounts, to membership, *etc.*

Overall, the year has been another busy one; in this report, we provide a brief overview of our ongoing efforts to build on the foundations created in previous years, and on the initiatives and affiliations we have forged.

Following the 2013 AGM in Valencia (May 2013), members of the EB participated in, and presented EMBnet at the EBI-Wellcome Trust Summer School in Hinxton, UK (June 2013); they also organised and trained on the NGS bioinformatics course at KEMRI-Wellcome Trust in Kilifi (August 2013). The EB also supported the development of two new QuickGuides.

Overall, EMBnet's training strategy has been dominated by our leadership of GOBLET[1] (the Global Organisation for Bioinformatics Learning, Education and Training), which now has around 30 members. Working through GOBLET has significantly increased our level of interaction and cooperation with a range of major international societies and networks, and especially this year with the International Society for Computational Biology (ISCB). Notable highlights have included the establishment of the first education poster track at ISMB/ECCB 2013[2], and development of a Community of Special Interest (COSI) around Computational Biology Education (CoBE), to be launched at ISMB 2014[3]. The ISCB also hosted a GOBLET interim meeting alongside ISMB/ECCB 2013, which saw a range of presentations: *e.g.*, from Lonnie Welch (on curriculum guidelines for bioinformatics and computational biology), from Ana Conesa (on multidisciplinary, multi-institutional PhD programmes), from Anupama Jigisha (on the ISCB Student Council intern initiative), and from Niklas Blomberg (on how ELIXIR will safeguard life science research in Europe). Later, a pan-European bioinformatics training strategy meeting was held at The Genome Analysis Centre (TGAC), UK (March 2014), in part to discuss how GOBLET and ELIXIR can work together in future to best serve their communities in the coming years.

During the year, we have also devoted a lot of time to working closely with our allied EU-funded projects. Specifically, members of EMBnet helped organise and/or attended a range of AllBio events[4]: these included a number of 'hackathons' (August 2013, Nijmegen, the Netherlands; September 2013, Amsterdam, the Netherlands; October 2013, Uppsala, Sweden; November 2013, Alnarp, Sweden), an *RNA-seq data analysis workshop* (January 2014, Espoo, Finland), the *AllBio/EMBRACE metagenomics interoperability workshop* (April 2014, Amsterdam, the Netherlands), and the AllBio AGM (September 2013, London, UK).

Within SeqAhead, EMBnet Nodes were involved in events covering a range of problems in

1 www.mygoblet.org/
2 www.iscb.org/ismbeccb2013
3 www.iscb.org/ismb2014
4 www.allbioinformatics.eu/doku.php?id=public:events

NGS data analysis: these included the *NGS data and the Variation Calling Challenge* (May 2013, Udine, Italy); *Hadoop and NGS data processing hackathon III* (June 2013, Pula, Italy); workshops on the *Future demands and challenges in ICT and bioinformatics tools for NGS* (June 2013, Pula, Italy), *NGS methods for identification of mutations and large structural variants* (March 2014, Lausanne, Switzerland) and *Assessment of training methods in NGS data analysis* (March 2014, Oeiras, Portugal); and the *NGS data after the Gold rush workshop and Management Committee Meeting* (May 2014, Norwich, UK). For references, see the 'SeqAhead 2013-2014' Events Web page[5].

Throughout the year, we have described these and our other activities in the monthly *EMBnet.digest* and in *EMBnet.journal*. For example, we made a special report on EMBnet's silver anniversary AGM in the May 2013 digest[6], we provided a round-up of the year's activities in the December issue[7], and we provided a report on our work with GOBLET and invited EMBnet to participate in a GOBLET survey in February's digest[8]. The most notable change for the digest has been a new, much simpler look-and-feel, starting from the January 2014 issue.

For the Journal, the most notable change has been a move to an instant access model, in which articles are published as soon as peer-review and layout have been completed – articles will henceforth be collated into volumes only once a year, commencing with volume 20. During 2013, the bulk of the journal work involved preparation of volume 19(1)[9] (which included reports from the EMBnet 2013 AGM, held 17-18 May in Valencia) and proceedings of the *The Next NGS Challenge Conference*[10]*: Data Processing and Integration* conference, held 14-16 May in Valencia, and of the NETTAB 2013 workshop[11] on *Semantic, Social and Mobile Applications for Bioinformatics and Biomedical Laboratories*, held 16-18 October in the Venice Lido.

Following the successful ratification of the new Statutes, which became legally binding in April 2013, we have been formally able to accept individual members, a facility made possible by the implementation of the online fee-payment module. The new statues also ushered in changes to the internal structure of the organisation. In particular, Committee Chairs now have the flexibility to convene their own working groups, without having to proceed via cumbersome election processes.

The 2014 AGM will provide an opportunity to review how some of these changes have been working in practice. It will be a chance to strategically re-group, to build on some of these initiatives, and to prioritise new projects and funding opportunities. This year, one member of the EB will be up for re-election: Goran Neshich. We have not yet received any additional candidacies for this position, but welcome applicants to submit personal statements about their plans for and commitments to this role.

This year, we would like to thank, in particular, Domenica D'Elia for her energy and fortitude in coordinating and sustaining the many successful activities of the P&PR PC; Rafael Jimenez for his ongoing technical support of the website; and Lubos Klucar for his patient and consistent work in managing the production of *EMBnet.journal*. As always, there's still a lot more work to do. We therefore encourage you all to contribute your energies and visions to EMBnet, to ensure EMBnet's continued success as the Global Bioinformatics Network!

Chair: T. K. Attwood
Secretary: A. Gisel; **Treasurer:** E. de Villiers; **Members:** E. Bongcam-Rudloff, G. Neshich

5 www.seqahead.eu/
6 www.embnet.org/sites/default/files/digest/EMBnetDigest_2013-05.pdf
7 www.embnet.org/digest/embnetdigest-december-2013
8 www.embnet.org/sites/default/files/digest/EMBnetDigest2014-02_1.pdf
9 journal.embnet.org/index.php/embnetjournal/issue/view/74/showToc
10 journal.embnet.org/index.php/embnetjournal/issue/view/75/showToc
11 journal.embnet.org/index.php/embnetjournal/issue/view/76/showToc

2014 Annual General Meeting: Publicity & Public Relations Project Committee Report




Domenica D'Elia[1]✉, Vicky Schneider-Gricar[2], Rubina Kalra[2], Cesar Bonavides-Martinez[3], Rafael Jimenez[4]

[1]CNR, Institute for Biomedical Technologies, Bari, Italy
[2]The Genome Analysis Centre (TGAC), Norwich, United Kingdom
[3]Center for Genomic Sciences (CCG), Cuernavaca Morelos, Mexico
[4]Itico and iAnn project, United Kingdom

Received 15 January 2014; **Published** 2 February 2015

D'Elia D *et al.* (2014) *EMBnet.journal* **20**, e800. http://dx.doi.org/10.14806/ej.20.0.800

To properly respond to the most urgent needs raised during the EMBnet workshop held in Valencia in May 2013[1], the Publicity & Public Relations Project Committee (P&PR PC) established two task-forces: i) a website task-force, comprising Rafael Jimenez and Cesar Bonavides-Martinez; and ii) a communication strategy task-force, comprising Vicky Schneider and Rubina Kalra.

An overview of activities and achievements was given by the PC Chair, and discussed during the EMBnet 2014 workshop[2] held in Lyon, 30 May. The programme also included a "Website hands-on" by Rafael Jimenez on "How to use the EMBnet website, add and manage content" aiming to expose members to some of its basic functions and services, and to practise their use.

1 journal.embnet.org/index.php/embnetjournal/article/view/693/981
2 www.embnet.org/agm/2014/programme

This article describes the achievements of the P&PR PC since June 2013, and plans for the next year.

Committee's Composition

Chair: Domenica D'Elia
Members: Rafael Jimenez and Cesar Bonavides-Martinez *(EMBnet website)*
Vicky Schneider and Rubina Kalra *(EMBnet & EMBnet.journal branding & communication strategies)*

Committee's Activities and Achievements from June 2013 to May 2014

The P&PR PC agreed to have quarterly Virtual Meetings (VMs). Additional VMs were held according to task-specific needs. The P&PR PC regularly attended the Operational Board (OB) meetings and supported the Executive Board (EB) by working on the following tasks:

1. website development, management and content moderation;
2. *EMBnet.digest* releases;
3. EMBnet and *EMBnet.journal* branding and communication strategies;
4. public relationships with EMBnet communities and related networks/societies;
5. EMBnet Sponsorship Policy and Sponsorship management.

1. Web site development, management and content moderation

New developments and achievements include:

a. individual and organisational subscriptions for membership and fee payment:
   - the "Join Us" section of the website was revised and updated to provide clear information on different membership options;
   - the submission procedures were revised again from January-February 2014 to improve the tracing system for payments, and to make the procedure more easy-to-use.

Thanks to these modifications, membership renewals and submission of new subscriptions can be easily managed and processed through the website.

b. Structural re-organisation of the "Contact us" section through the inclusion of both a general enquiry form and task-specific forms for queries related to membership, organisation, events, training, projects and sponsorship;
c. development and implementation of the "EMBnet AGM 2014" website section.

New proposals are:

d. the creation of a website portal dedicated to Education & Training (ET), preferably linked to GOBLET[3] and including:
   - ET initiatives, schools and courses active across EMBnet Nodes;
   - programmes of research and student exchange active across EMBnet Nodes and affiliated societies and networks;
   - internship demands and offers;
e. the creation of a website section for building project proposals. This section could include:
   - announcements of project proposals demanding specific expertise;
   - a register of specialists available for projects demanding specific expertise.

2. EMBnet and EMBnet.journal branding & communication

Branding & Design are prime ways for EMBnet to differentiate itself from its competitors. EMBnet branding strategy should consist of a plan that uses a unique set of design tools created for EMBnet, and applied to every communication vehicle to convey EMBnet (brand) identity. These tools include colour palette, typefaces, format, images and language. EMBnet has some of these elements in place. Indeed, in recent years, since the development of the new website, the P&PR PC already renewed the EMBnet branding by adopting a new colour palette, by developing a new logo and new templates for EMBnet presentations, brochures and leaflets (D'Elia D, 2013). However, the plan of the P&PR PC was to redefine and implement, during 2014-2015, new core elements, including also EMBnet.journal, in order to improve the perception of EMBnet in the scientific community. Publicity material, such as EMBnet brochures, leaflets, presentation templates, etc., should be created by following the new EMBnet Branding & Design strategy.

Our initial steps involved reviewing the existing EMBnet design brief and mapping any aspects that needed immediate revision versus more radical aspects that might require a complete redesign of the brand. Vicky Schneider and Rubina Kalra drafted a document, which was distributed in January 2014 to the OB and to *EMBnet.journal*'s Executive Editorial Board (ENJ EB) for feedback. A snapshot of the main outcomes of this consultation process is illustrated in Figure 1.

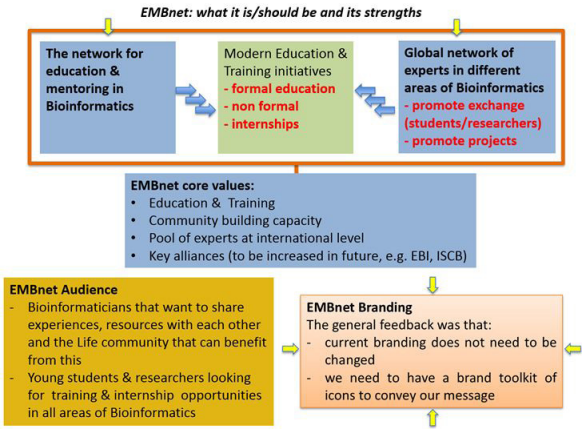


Figure 1. Summary of the most significant outcomes of the P&PR PC survey to outline a new branding & design strategy.

The major core values of EMBnet were considered to be in bioinformatics education and training, and in its networking and capacity building. To build on these strengths, it will be important to:

- develop modern education and training initiatives, and reinforce our alliances with other societies and networks proficient in the field;
- promote new initiatives, including new projects, and student and researcher exchange programmes.

The main actions are to:

a. develop and implement EMBnet education and exchange programmes;
b. establish an annual EMBnet Conference to be held jointly with the AGM, and possibly in collaboration with other affiliated societies, such as ISCB;
c. promote common projects;
d. deliver more tools and Web services;
e. publish more papers and education-related articles in EMBnet.journal;
f. produce more QuickGuides and online courses;
g. provide well-maintained services useful for teaching and well-documented training exercises;
h. promote membership subscriptions.

The points above were discussed during the EMBnet workshop and AGM, alongside proposals from the P&PR PC. A detailed report has been published in this volume by the Chair of the EB (Attwood, 2014). Realising these objectives will depend on substantial and consistent collaboration from EMBnet members, including contri-

3 mygoblet.org

butions to the website (*i.e.*, by publishing news of newly developed tools and services, new research achievements, publications and events organised across Nodes), to the *EMBnet.digest*, *EMBnet.journal* and *EMBnet QuickGuides*, and to the promotion of EMBnet and EMBnet initiatives at local levels.

As for *EMBnet.journal*, the P&PR PC has contributed in the:

a. revision of the journal Section policies;
b. rebranding of the journal Focus and Scope;
c. reshaping of the website;
d. revision of the Authors Guidelines;
e. production of a first draft proposal of the Journal's Advertisement policy and a Journal Media kit (Rubina Kalra).

A detailed report was provided by Lubos Klucar during the EMBnet workshop in Lyon.

3. Public relationships with EMBnet communities and related networks/societies

During this last year the P&PR PC has:

a. assisted members by providing support as and when requested;
b. managed and answered contacts' requests posted on the website;
c. managed the production and dissemination of EMBnet.digest and of EMBnet.journal;
d. managed sponsorships of large conferences: i.e., 2013 RECOMB-CG (Lyon, FR), NETTAB 2013 (Venice, IT), 2014 SAGS (South African Genetics Society) and SASBi (South African Society for BioinformaticsSAGS-SASBi) Conference, Kwalata Game Reserve (ZA); NETTAB 2014 (Turin, IT);
e. produced publicity material, such as a new poster and leaflet;
f. received and managed, in collaboration with the OB, the evaluation of collaboration requests from affiliated societies and networks or EMBnet Nodes;
g. contributed to the organisation of the 2014 AGM.

4. EMBnet Sponsorship Policy

A draft proposal of the 'EMBnet Sponsorship Policy' was submitted by the P&PR PC for OB evaluation in January 2014. A slightly revised version of the document was approved in March 2014. The 'EMBnet Sponsorship Policy' is available from the website in the *Contact us* section at: http://www.embnet.org/contact_form/sponsorship.

Acknowledgements

As Chair of the P&PR Committee, my personal thanks go to all Committee members, without their support, many of our activities and achievements would have been impossible. Unfortunately, owing to an overload of business at TGAC, both Vicky Schneider and Rubina Kalra were no more able to provide their support as members of the Committee since March 2014. Nevertheless, they both provided a significant contribution on key PC tasks and commitments. I would like to express all my gratitude for their valuable contribution. Special thanks go to Axel Thieffry, who joined the P&PR PC during the AGM. Axel is a young and promising individual member of EMBnet, and I'm very happy to welcome him in the P&PR PC. He is a striking example of how opening EMBnet to individual membership can greatly enrich our network. Finally, and as always, I'd like to thank the Executive Board, and in particular the Chair, Teresa K. Attwood, for her dedication and the valuable support that she has always provided.

References

D'Elia D (2013) '2013 Annual General Meeting: Publicity & Public Relations Project Committee Report'. *EMBnet.journal* **19**(1), 28-29. doi: http://dx.doi.org/10.14806/ej.19.1.705

Attwood TK (2014) EMBnet, the Global Bioinformatics Network: a report on the workshop and 26th AGM, Lyon, May 2014. *EMBnet.journal* **20**:e786. doi: http://dx.doi.org/10.14806/ej.19.1.705

Fq_delta – Efficient storage of processed versions of fastq files

Andra Veraart[1,2]✉, Henk-Jan van den Ham[2], Maarten A. Bijl[2], Arno C. Andeweg[2], Anita C. Schürch[2]

[1]School of Communication, Media and Information Technology, Rotterdam University, Rotterdam, The Netherlands
[2]Dept. Viroscience, Erasmus Medical Center, Rotterdam, The Netherlands

Received 11 July 2013; **Accepted** 2 December 2013; **Published** 5 March 2014

Veraart A *et al.*. (2014) *EMBnet.journal* **20**, e698. http://dx.doi.org/10.14806/ej.20.0.698

Competing Interests: the authors have declared that no competing interests exist.

Abstract

This technical note describes fq _delta, a python module and shell script that enables the storage of processed versions of fastq files generated by DNA and RNA sequencing technologies. By using Myer's diff algorithm to perform per-character comparisons between the original and processed fastq files, we generate delta files that describe the changes in the processed fastq file relative to the original file. While the delta files are only a fraction of the original size (0.1 – 3%), they allow lossless reconstruction of the processed fastq files. Depending on the number of processing steps, implementation of this module will lead to a significant reduction in storage required for processing sequence data.

Availability:
Fq _delta is available for download at https://github.com/averaart/fq__delta.

Introduction

Advances in sequencing technology have led to an exponential increase in the volume of sequencing data that is generated (Wetterstrand, 2013). The amount of data that is now generated poses a challenge to storage facilities, especially for primary data. Currently, the cost of sequencing is dropping faster than the cost of storage space, and will probably continue to do so in the near future (Komorowski, 2009). Additionally, data processing can require storage of intermediate analysis steps. Data compression reduces the need for storage capacity, and several compression methods have been applied to raw sequence data (Grassi *et al.*, 2012; Bhola *et al.*, 2011; Jones *et al.*, 2012; Howison, 2012; Bonfield & Mahoney, 2013; Hach *et al.*, 2012). The processing of sequence data typically consists of several steps. Runs are often split into separate samples, followed by removing sequence tags or trimming of low quality reads. Examples of popular pre-processing software include cutadapt (Martin, 2011), the FASTX-toolkit (Gordon & Hannon, 2010) or Biopython (Cock *et al.*, 2009). These create one or more versions of the original data file, thus tending to require several times the original storage capacity. To save storage capacity, processed versions are often discarded, but in many cases these files are saved to allow easy access to all intermediate steps without redoing the analysis. Moreover, a growing number of researchers advocate the publication of raw sequence data and code to improve reproducibility of results (Peng, 2011; Stodden, 2010; Barnes, 2010; Baggerly & Berry, 2011), which would be facilitated by having processed intermediate files available. Given that the differences between the original and a processed version of the data are often minor, storage and compression of only the differences between versions would be far more efficient than retaining complete versions.

For saving different versions of the same file, several general-purpose applications are available, but the specific type of manipulation that is performed in sequence data processing pre-

cludes their use. Processing often entails the removal of several bases from each read. Existing general purpose applications typically work on a line-basis or block-basis, *i.e.* a fixed number of bytes (*e.g.,* diff[1] and rdiff[2], respectively). If one base has changed, the complete line or block will be stored instead of only the changed bases. This behavior makes these applications inefficient for storing processing steps in sequence data analysis, and suggests that these data require a high-resolution method to efficiently save processed sequence data files.

Table 1. Size reduction achieved by storing the processed fastq file in a zip archive, by storing an rdiff delta file in a zip archive and by using fq _ delta. File sizes expressed in Megabytes. Percentages are based on the processed fastq file size, as indicated in the first column. The original, unprocessed fastq file was 802.5 MB.

	fastq	zipped fastq	zipped rdiff delta	fq_delta
fastq _ masker -q 10	765.29	225.58 (29%)	212.87 (28%)	7.19 (0.94%)
fastq _ masker -q 25	765.29	228.17 (30%)	227.49 (30%)	17.53 (2.3%)
fastq _ quality _ trimmer	740.10	221.32 (30%)	197.77 (27%)	2.11 (0.28%)
cutadapt	751.53	223.05 (30%)	89.68 (12%)	0.74 (<0.1%)
cutadapt trimmed only	41.59	11.47 (28%)	11.48 (28%)	0.82 (2.0%)
cutadapt untrimmed only	709.94	211.69 (30%)	81.38 (11%)	0.39 (<0.1%)
removed lines	306.58	90.67 (30%)	0.002 (<0.1%)	0.08 (<0.1%)

This paper describes fq_delta, a python module to store differences between versions of fastq files. This module compares strings on a per-character basis and stores differences between them, thereby saving all changes into a file that is a fraction of the processed fastq file. Storage of the original file and delta files therefore enables full reconstruction of processed versions of fastq files.

Design and Implementation

Fq_delta uses the google-diff-match-patch library (Fraser, 2009), which implements Myer's diff algorithm (Myers, 1986). Fq_delta applies this technique to fastq files.

Consider a given sequence line in a fastq file containing the following string:

```
ACACGTAGTATACGGCATGCTACG
```

Assume the first 11 characters comprise a sequence tag that needs to be removed before further analysis takes place, resulting in the following string:

```
ACGGCATGCTACG
```

In the delta file, this line would be listed as:

```
-11 =13
```

This describes that the first 11 characters of the original line are removed and the subsequent 13 characters remain the same. Storing this description uses far fewer bytes than storing the complete processed line. Even when storing a large number of minor changes, this method is more efficient than the standard command line diff application. The processed string can be reconstructed using the first string and the difference between the first and second strings, as documented in the delta file.

Generating a delta file

Fq_delta expects two fastq files as input. The first is assumed to be the original, the second a processed version: *i.e.*, fq_delta computes the delta of the second input file relative to the first. Four lines of each file are read, covering one sequence read. To ensure that related original and processed sequences are matched, the identifier lines (IDs) are first compared, excluding any tab-separated values. If the IDs match, the difference between each line is written to a text file, called the delta file. If the IDs do not match, the read from the original file has evidently been removed from the processed file (fq_delta thereby assumes that the original and processed fastq files have the same read order). This is written into the delta file and the next four lines are read from the original file. This process is repeated until the end of the processed file has been reached. An

1 pubs.opengroup.org/onlinepubs/9699919799/utilities/diff.html
2 linux.die.net/man/1/rdiff

md5 hash of the processed file is calculated and written to a separate checksum file so that data integrity can be verified when reconstructing the file (fq_delta raises an error when these do not match). Finally, the delta file and the checksum file are compressed into one standard zip archive.

Retrieving a processed file

Fq_delta expects two files during retrieval: the original fastq file, and the zip archive containing the delta file that represents the processed version. The delta file and checksum file are both extracted from the zip archive. If the checksum file is not found, the process is aborted immediately. The original file and the delta file are read line by line. The delta line is applied to the original line to reconstruct the processed line. The processed line is either written to a new file or printed to *standard out.* When the end of the delta file is reached, the process is stopped, effectively ignoring the last lines that were in the original but not in the processed versions.

Technical details

Fq_delta is written in Python 2.7 as a module. It provides a class that implements the same functions as typical file-like objects. The class is able to use *standard in* or *standard out* as input or output, respectively. Fastqfiles that are compressed using quip (Jones *et al.*, 2012) are decompressed on-the-fly. Fq_delta assumes an unchanged order of reads from one version to the other to generate the delta file. The Fq_delta module was tested in scenarios where data integrity was deliberately compromised. In all cases, the application detected the error and reportedit to the user.

Fq_delta can also be used as a command-line tool, using two additional shell scripts that are provided with the module. The python module, command line scripts and a test script are available at https://github.com/averaart/fq_delta.

Results & Discussion

The application was tested using fastq_masker and fastq_quality_trimmer from the FASTX-toolkit (Gordon & Hannon, 2010), and cutadapt (Martin, 2011) on the first 2,500,000 reads of a publicly available data-set (Uddenberg *et al.*, 2013) for Norway spruce (Picea abies; the test shell script is available from the codebase). To demonstrate that removed lines are handled correctly, irrespective of their location in the file, an extra test was performed where we removed 500,000 reads from the start, the middle and the end of the example set.

Using fq_delta, the processed files were successfully compressed and accurately reproduced, using the original file as a reference. Table 1 illustrates the file sizes of the processed files and the resulting delta files, showing a reduction in required storage of at least 97 percent.

The sizes of fq_delta files were compared with compressed versions of the processed files, and both uncompressed and compressed versions of diff and rdiff files. In most cases, the uncompressed diff and rdiff files were much larger than the compressed fastq files; only the compressed rdiff file was smaller in all cases (Table 1). The fq_delta files were an order of magnitude smaller in all cases, except the "removed lines" scenario.

The large difference between rdiff and fq_delta can be explained by the coarse- and fine-grained resolutions of the respective methods. The rdiff algorithm is too coarse to efficiently register small changes, whereas fq_delta works on a per-character basis. This is illustrated by the 'removed lines' test, where the fastq file was divided into five continuous sections, two of which were saved in the processed file. Only in this coarse-grained scenario did rdiff perform better than fq_delta.

In summary, fq_delta is able to store multiple versions of a fastq file at a fraction of the usual storage costs. None of the tools we are aware of show the same efficiency. There are no requirements to the fastq files, except that the order of the reads should be consistent between original and processed versions. Especially combined with compression of original files, fq_delta drastically reduces the amount of storage necessary when processing sequence data.

Acknowledgements

This study was supported by the Virgo consortium, funded by the Dutch government, project number FES0908, and by the Netherlands Genomics Initiative (NGI), project number 050-060-452.

References

Baggerly KA & Berry DA (2011) Reproducible Research | Amstat News. *AMSTAT News Blog.* http://magazine.amstat.org/blog/2011/01/01/scipolicyjan11/ (accessed 11 June 2013).

Barnes N (2010) Publish your computer code: it is good enough. *Nature* **467**, 753. http://dx.doi.org/10.1038/467753a

Bhola V, Bopardikar AS, Narayanan R, Lee K & Ahn T (2011) No-Reference Compression of Genomic Data Stored in F ASTQ Format IEEE. http://dx.doi.org/10.1109/bibm.2011.110

Bonfield JK & Mahoney M V (2013) Compression of FASTQ and SAM Format Sequencing Data. *PloS one* **8**, e59190. http://dx.doi.org/10.1371/journal.pone.0059190

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)* **25**, 1422–3. http://dx.doi.org/10.1093/bioinformatics/btp163

Fraser N (2009) google-diff-match-patch - Diff, Match and Patch libraries for Plain Text. http://code.google.com/p/google-diff-match-patch/ (accessed 16 May 2013).

Gordon A & Hannon GJ (2010) FASTX-Toolkit. *FASTQ/A short-reads pre-processing tools.* http://hannonlab.cshl.edu/fastx _ toolkit/ (accessed 31 May 2013).

Grassi E, Gregorio F Di & Molineris I (2012) KungFQ: a simple and powerful approach to compress fastq files. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* **9**, 1837–42. http://dx.doi.org/10.1109/tcbb.2012.123

Hach F, Numanagic I, Alkan C & Sahinalp SC (2012) SCALCE: boosting sequence compression algorithms using locally consistent encoding. *Bioinformatics (Oxford, England)* **28**, 3051–7. http://dx.doi.org/10.1093/bioinformatics/bts593

Howison M (2012) High-Throughput Compression of FASTQ Data with SeqDB. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* **10**(1), 213–218. http://dx.doi.org/10.1109/tcbb.2012.160

Jones DC, Ruzzo WL, Peng X & Katze MG (2012) Compression of next-generation sequencing reads aided by highly efficient de novo assembly. *Nucleic acids research* **40**, e171. http://dx.doi.org/10.1093/nar/gks754

Komorowski M (2009) A History of Storage Cost. http://www.mkomo.com/cost-per-gigabyte (accessed 31 May 2013)

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, pp. 10–12. http://dx.doi.org/10.14806/ej.17.1.200

Myers EW (1986) AnO(ND) difference algorithm and its variations. *Algorithmica* **1**, 251–266. http://dx.doi.org/10.1007/bf01840446

Peng RD (2011) Reproducible research in computational science. *Science (New York, N.Y.)* **334**, 1226–7. http://dx.doi.org/10.1126/science.1213847

Stodden V (2010) Reproducible Research. *Computing in Science & Engineering* **12**, 8–13. http://dx.doi.org/10.1109/mcse.2010.113

Uddenberg D, Reimegård J, Clapham D, Almqvist C, Von Arnold S *et al.* (2013) Early cone setting in Picea abies acrocona is associated with increased transcriptional activity of a MADS box transcription factor. *Plant physiology* **161**, 813–23. http://dx.doi.org/10.1104/pp.112.207746

Wetterstrand KA (2013) DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). http://www.genome.gov/sequencingcosts/ (accessed 14 May 2013).

Large-scale statistical analysis of genome data with Ruby and R: skipping interface libraries



Sergio R. P. Line✉, **Ana P. de Souza, Luciana S. Mofatto**

Department of Morphology, Piracicaba Dental School, Piracicaba, SP, Brazil

Received 30 January 2014; **Accepted** 29 April 2014; **Published** 26 May 2014

Line SRP *et al.* (2014) *EMBnet.journal 20*, e753. http://dx.doi.org/10.14806/ej.20.0.753

Competing Interests: none

Abstract

Ruby is a dynamic interpreted, open source, object-oriented programming language with an elegant syntax and a focus on simplicity and productivity. One factor that may hinder the dissemination of Ruby, among academic and technological communities, is that it does not contain built-in methods for statistical analysis and graph creation. Statistical analysis with numerical data generated by Ruby scripts is traditionally performed by storing data to a file, which is read into another software environment for statistical analysis, using a package such as R. In order to circumvent this limitation, libraries have been created to perform statistical analysis with Ruby. These have not gained popularity, possibly owing to its limited statistical methods and relative complex usage. In this paper, we describe a simple and dynamic procedure to connect Ruby and R scripts. We show that this approach can be used for large-scale genome-data processing and statistical analysis. Its usage is simpler than interface libraries, as it does not require the creation of methods or routines other than those already existing in R and Ruby.

Introduction

The development of high-throughput DNA sequencing techniques has led to an exponential increase in the volume of data available for analysis. This has opened new frontiers in medical and biological studies, and boosted interest in the genome research arena. However, in many cases, the analysis of large volumes of data can only be performed by computationally intensive and complex methods that include computer processing and statistical analysis of sequences. Developing statistical and programming skills is a major challenge, and frequently a discouraging factor for students and researchers in the biomedical area.

Ruby is a dynamic interpreted, open source, object-oriented programming language with a focus on productivity (Flanagan and Matsumoto, 2008); it is characterised by an elegant syntax and simplicity, it is natural to read and easy to learn (Aerts, 2009). The wide range of built-in methods for string manipulation, reflection and metaprogramming capabilities make this language especially suitable for bioinformatics, where the size and complexity of codes can hinder readability (Aerts, 2009). There are several Ruby implementations available, such as JRuby (which runs on the Java Virtual Machine), Rubinius (an alternative implementation written in Ruby and C) and the standard reference C implementation, which is now on stable version 1.9. Because of these characteristics, Ruby is an increasingly popular programming language, and has been among the most popular interpreted languages (http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html). One factor that may hinder the dissemination of Ruby, especially among academic and technological communities, is the fact that it does not contain built-in methods for statistical analysis and graph creation. In the past few years, Ruby libraries (gems) for statistical analysis have been created (*i.e.*, RSRuby, RinRuby and Statsample). Statsample[1] has a limited number of statistical methods, while RSRuby (Gutteridge, 2008) and RinRuby (Dahl

1 ruby-statsample.rubyforge.org/

Table 1. Examples of R and RSRuby syntaxes for statistical tests.

Statistical test	R	RSRuby
t-test	t.test(a,b)	r.t _ test(a,b)
Correlation test	cor(a,b, method = "spearman")	r.cor(a,b, :method => "spearman")

and Crawford, 2009) integrate Ruby with R. R is a scripting language and environment developed for statistical computing, with outstanding capacities for graphics generation (R Development Core Team, 2013). It has an extensive library of routines, with hundreds of contributors, it has been heavily used and is widely accepted by the scientific community.

The fact that RSRuby is a C extension for Ruby makes it much faster than RinRuby, which is 100% Ruby implemented. RSRuby, however, has some disadvantages, as: i) it is not available for alternative implementations of Ruby (*e.g.*, JRuby); ii) it is dependent on operating system, Ruby implementation and R version; and iii) downloading may not be trivial for people with no formal training in informatics. The main drawback of RSRuby is that it does not have the full capacity of R (*i.e.*, *p* values for correlation analysis cannot be directly obtained), and the transformations between R and Ruby are not trivial, as in many cases the methods for statistical tests have quite different syntaxes (Table 1).

RinRuby and R methods have the same syntax, and all the parameters from statistical analysis can be transformed in Ruby objects. In fact, RinRuby seems to be the less complicated bridge between R and Ruby. RinRuby, however, requires the assignment of variables to connect Ruby to R, a procedure that has to be repeated many times in complex codes. Limited by the factors listed above, the libraries for statistical analysis with Ruby have not gained much popularity, and it seems that the most common procedure to perform statistics from data generated by Ruby scripts is the storage of data in files, and later access of files through the R (or other statistical packages) console or command line (Chang, 2012). This approach, however, is time consuming and may not be feasible in high-throughput analysis of data, where hundreds or thousands of statistical tests have to be performed on a given data-set.

Ruby has the capability to execute an R script from within a Ruby script. This can be achieved using the Ruby system method (system("data")) to run an R code using the batch mode (R CMD BATCH script.R). Perhaps the easiest procedure would be to create a file (.csv or .txt) with the numeric data generated by a Ruby code, and subsequently to run an R script using batch processing (system("R CMD BATCH script.R")). In this approach, two separate scripts are created. The first is a Ruby script that generates and stores data in a file, where numeric values for each group or treatment are stored in distinct columns. Execution is then transferred to an R script that contains the commands for the statistical analysis. In our view, this is the simplest way to link Ruby and R. One possible drawback for this approach is the creation of the file to be accessed by the R script, which can delay the execution of the script, especially in files generated from very large data-sets.

In this report, we demonstrate the use of the Ruby method for statistical analysis of large-scale human coding sequences with R, and compare its time performance with RinRuby.

Methods and Scripts

Procedures

The scripts were run on Ruby version 1.9.3[2], R version 2.14.2[3] and Linux operating system Ubuntu version 12.04[4]. The performances of the Ruby system and RinRuby were compared analysing 29,064 human coding sequences from the The Consensus CDS (CCDS) project[5]. Sequences were stored in a single file (CCDSfinal.txt) in FastA format. The sequences were analysed as follows:

1. the CCDSfinal.txt file was opened, and each coding sequence was sequentially read;
2. the size, number of bases (A, C, G, T) and CpGs (cytosine followed by guanine) of the coding sequences were printed to a .csv file (Ruby system approach) or stored in arrays (RinRuby) (Supplementary file 1[6]);

2 www.ruby-lang.org/en/downloads
3 www.r-project.org
4 www.ubuntu.com/download
5 www.ncbi.nlm.nih.gov/projects/CCDS
6 journal.embnet.org/index.php/embnetjournal/article/downloadSuppFile/753/969

3. the data from the .csv files and arrays were used to plot frequency charts of bases (A, C, G and T) versus the frequency of CpGs. The Spearman rank correlation coefficients between coding sequence size and number of CpGs, and the respective p-values, were also calculated.

The time required to accomplish these procedures was obtained by the mean of five program runs, and was measured using the Ruby benchmarking method. The scripts are reported below.

Ruby system approach

Ruby script

```
File.open("results.csv","w")
def countCpG(seq)
       count = 0; cpg = 0
       while count < seq.size - 1
              cpg += 1 if (seq[count] +
seq[count +1]) == "CG"
              count += 1
       end; cpg
end

File.open("CCDSfinal.txt","r").each do
|line|
       unless line[0] == ">"
              File.open("results.
csv","a") do |f|
              f.print line.chomp.
size,",",line.count("A"),",",line.
count("C"),",",line.count("G"),",",line.
count("T"),",",countCpG(line),"\n"
              end
       end
end
system("R CMD BATCH statistics.R")
```

R script

```
file =read.csv("results.csv")
   par(mfrow = c(2,2))
   par(mar= c(4.5,5,4,4))
   bases = c("A","C","G","T")
for(i in 2:5){
   plot(file[,i]/
file[,1],(file[,6]/file[,1]),xlab =
paste("frequency",bases[i -1]), ylab =
"frequency CpG", col = "red",cex.lab = 2)
   lines(loess.smooth(file[,i]/
file[,1],file[,6]/file[,1], span= 3/5,
family="gaussian"), lwd = 2)}
corr = cor.test(file[,6],file[,1],method
= "spearman")
print(paste(corr$estimate,corr$p.value))
```

Results and Discussion

Figure 1 shows a graph of the frequency of each base (number of specific bases in a coding sequence/size of coding sequence) versus the frequency of CpG (number of CpGs in a coding sequence/size of respective coding sequence). CpG dinucleotides may be enzymatically methylated, and this chemical modification can modulate gene expression (Robertson, 2005). As can be seen, the relationship between CpG and base frequency is not linear.

The processing and statistical analysis of CCDS project coding sequences was completed in an average of 32.32 s with the Ruby system against 37.48 s with RinRuby (t-test p = 1.3e-07). The Ruby approach was also faster when we doubled the genes, or when only half of the genes were analysed (Figure 2). The scripts using the Ruby system approach had 795 characters altogether (Ruby and R scripts) against 1722 characters of RinRuby, which requires the assignment of variables to connect Ruby to R.

The Ruby system approach also works in Windows, where it requires the path to the R directory installation before the R CMD BATCH (*i.e.*, system("path R CMD BATCH script.R")) and in Mac OS X system in a way similar to that described here for the Linux operating system. This approach can probably be performed with any

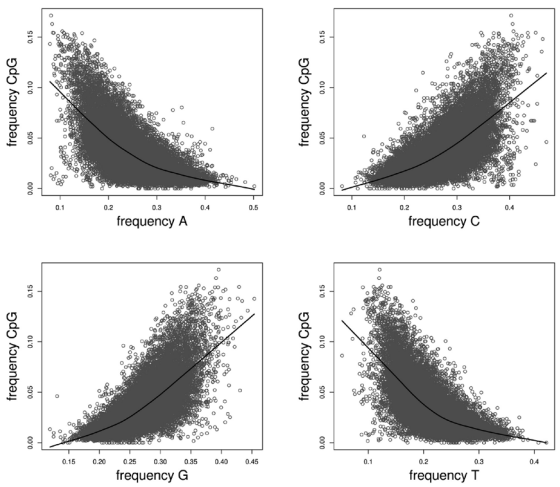


Figure 1. Plot of the frequency of each base (number of specific bases in a coding sequence/size of coding sequence) versus the frequency of CpGs (number of CpGs in a coding sequence/size of its coding sequence). The fitting line was obtained with the LOESS (locally weighted scatterplot smoothing) function of R.

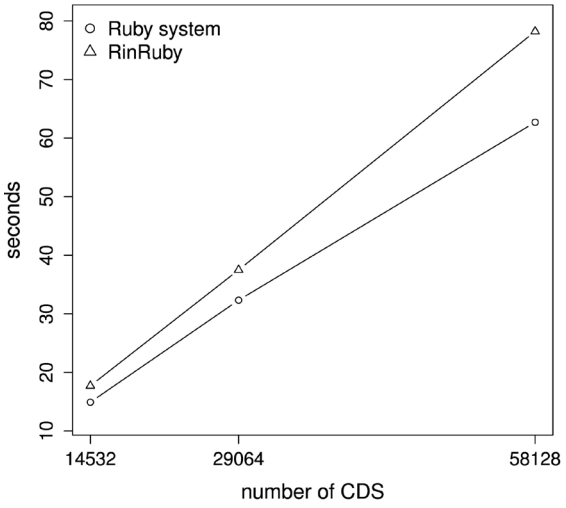


Figure 2. Time performance of the Ruby system and RinRuby approaches. Note that the Ruby system approach is faster than RinRuby, and that the time difference increases as the number of coding sequences increases. Number of CDS = number of coding sequences.

other computer language that has the capabilities to execute R scripts from within its native script.

Our analyses show that the Ruby system can be used for large-scale processing and statistical analysis of DNA sequencing data. In our view, this approach has three main advantages over other procedures: i) it avoids the installation of interface libraries; ii) it is simpler to use, as it does not require the creation of methods or routines, other than those already existent in R and Ruby; iii) the codes are shorter and more readable, as it does not require the assignment of variables to connect Ruby to R. We hope that the simplicity of the approach presented in this paper will incentive the use of Ruby in bioinformatics, as well as in other academic and technology fields, especially by professionals and students with no formal training in informatics.

Acknowledgements

LSM was supported by Fundacao de Amparo a Pesquisa do Estado de Sao Paulo-FAPESP.

References

Aerts J, Law A (2009) An introduction to scripting in Ruby for biologists. *BMC Bioinformatics* **10**, 221. http://dx.doi.org/10.1186/1471-2105-10-221

Chang SS (2012) *Exploring everyday things with R and Ruby.* O´Reilly Media. Sebastopol.

Dahl DB, Crawford S (2009) RinRuby: Accessing the R interpreter from pure Ruby. *J. Statist. Software*, **29**(4), 1-18.

Flanagan D, Matsumoto Y (2008) *The Ruby Programming Language*. O´Reilly Media, Sebastopol.

Gutteridge A (2008) RSRuby: A bridge between Ruby and the R interpreted language. Ruby package version 0.5.1. http://rubyforge.org/projects/rsruby (accessed 7 January 2014).

R Development Core Team (2013) The R project for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/ (accessed 6 January 2014).

Robertson K (2005) DNA methylation and human disease. *Nat Rev Genet* **6**, 597-610. http://dx.doi.org/10.1038/nrg1655

AACDS: A database for personal genome interpretation


Thanawadee Preeprem✉, **Greg Gibson**
Georgia Institute of Technology, Atlanta, United States

Received 11 August 2014; **Accepted** 5 September 2014; **Published** 10 October 2014

Preeprem T and Gibson G. (2014) *EMBnet.journal* **20**, e780. http://dx.doi.org/10.14806/ej.20.0.780

Competing Interests: none

Abstract
Incorporation of diverse data sources adds value to genomic studies, especially for annotation and categorisation of personal genome variants. The database for Association-Adjusted Consensus Deleterious Scheme (AACDS) and its Web application deliver a novel approach to assess genetic variations based on their putative functionality. The database is built upon integrated knowledge of variant data, with the aim of relating clinical phenotypes to predictions of variant deleteriousness. The simple but inter-related queries classify each variant into an 8-level category. The categories can be ranked, enabling straightforward interpretation of relative likelihood of functionality. The ranking thus facilitates improved efficiency in prioritising further detailed evaluation of key variants within a personal genome. The AACDS database covers more than 68 million mis-sense variants in approximately 18,000 human genes. Given a list of genetic variants, the retrieval of the AACDS category, along with known clinical data can be performed through an intuitive search platform.
Availability: The AACDS Web application is publicly available at http://cig.gatech.edu/tools.

Introduction
Non-synonymous Single Nucleotide Polymorphism (nsSNP) is one of the most common forms of genomic variability. About 60% of known disease-causing mutations are nsSNPs (Cooper *et al.*, 2010). One of the major goals for personal genomics is to identify a subset of variants that have the potential to influence an individual's health. Each individual genome is estimated to contain roughly ten thousand nsSNPs (Kim *et al.*, 2009; Ng *et al.*, 2008; Patel, *et al.*, 2013). The assessment of deleteriousness for SNPs is commonly performed on a per variant basis, by using many available computational tools that typically classify each SNP into two groups: benign and damaging. Although many prediction programmes have been proven to have acceptable accuracy, mostly in the range of 70-80% (Gonzalez-Perez and Lopez-Bigas, 2011), it is deemed an advantage to incorporate more data into the assessment (Ng and Henikoff, 2006).

In our recent study on interpretation of personal genome data (Preeprem and Gibson, 2013), we developed the "Association-Adjusted Consensus Deleterious Scheme" (AACDS) to facilitate variant prioritisation of personal genome studies. AACDS is constructed from the combination of existing databases that implicate the variant with disease or phenotype, and traditional sequence-based predictions. It classifies variants according to an 8-level category. Not only does AACDS incorporate the clinical or phenotypic annotations of the genomic variants in an individual, it also narrows down the variants to a subset that is appropriate for further follow-up experiments and validation with respect to individualised health profiles.

To promote the utility of our variant classification schema, AACDS, we have implemented the assessments into a database-driven Web application that allows users to search the AACDS categories and relevant information for user-defined variants. The AACDS website aims to provide a user-friendly platform for anyone interested in personal genome interpretation. The database schema was designed to cover the annotated list of functional variants (31,092 disease-associated amino acid variants in 3,363 genes), 4,225 pairs of gene-disease associations, 5,113 pairs

of gene-trait associations, and all possible coding genomic variants in 18,349 human genes (*i.e.*, 68,165,196 nsSNPs). Therefore, our newly developed database-driven Web application for AACDS can serve as a tool to generate the best estimate of clinical significance of each variant from the large and growing accumulation of personal genome data. In addition to identifying causal variants or variants in disease- or trait-associated genes from a list of genomic variability, the application also allows users to carry out further functional analyses of all SNPs in any gene of interest.

Although many tools and databases exist for the purpose of variant prioritisation and/or personal genome interpretation, we are not aware of any tool with similar features to ours, especially in the categorisation of genomic variants. Our AACDS tool allows SNP evaluations to be performed simultaneously on the basis of deleterious predictions, direct connections between variants to diseases, and associated traits and diseases to the genes. The tool assigns an AACDS class to each individual SNP; it also reports the overall AACDS statistics for a given genome. The classification and ranking of SNPs is particularly significant and original, as it assists effortless interpretations of whole-genome SNP searches. The results facilitate the identification of high-impact variants within a genome in an effective and efficient manner.

Compared to aggregative variant association methods such as in VAAST 2.0 (Hu *et al.*, 2013), our tool does not require that users have prior knowledge of various additional genomic attributes to perform searches and interpret the results. VAAST requires not only target and background genome data-sets, but also user-defined sets of genes and prior knowledge of genetic parameters (inheritance, penetrance, locus heterogeneity, allele frequency, *etc.*) in order to search for causal SNPs or genes. The search pipeline is neither designed for evaluation of all genomic variants, nor as a simple look-up utility.

Two recent genome analysis tools, eXtasy (Sifrim *et al.*, 2013) and Phen-Gen (Javed *et al.*, 2014), introduce a new phase of genome interpretation, in which the tools link genome variants to a specific phenotype. Although both tools have great potential for guiding diagnostics of rare disorders through the identification of phenotype-specific causal variants, the evaluations are performed on a per disease basis. Most personal genome variants are likely to be neutral and contain a minimal number of annotated disease SNPs (Preeprem and Gibson, 2013; Xue *et al.*, 2012); the individuals are healthy and unlikely to have noticeable clinical phenotypes (Patel *et al.*, 2013). These limitations represent a significant challenge for personal genome variant annotation for sub-clinical phenotypes, in whose interpretation AACDS is designed to help.

Implementation

The AACDS website serves as an interface for queries of the AACDS database, which is built to categorise nsSNPs into an 8-level class, based on their consensus predicted deleteriousness and the evidence of disease or complex trait associations with their genes. The database includes 68,165,196 nsSNPs that can be found in a human genome. The website allows users to retrieve the AACDS classification and relevant information about variants in genes of interest.

Data sources

To facilitate the variant mapping of various data types (chromosome coordinates, gene names, protein names), we chose UniProtKB (UniProt Consortium, 2012) as the core database. UniProtKB accession numbers provide unique identifiers for gene products, allowing direct look-up of the disease-association data from the selected SNP databases: MSV3d (Luu *et al.*, 2012) and SwissVar (Mottaz *et al.*, 2010). A list of 20,277 reviewed human proteins (representing the gene products of 19,700 genes) was compiled from UniProtKB (2012_06 release, accessed 1 November 2013).

Next, we used dbNSFP v2.1 (Liu *et al.*, 2011) (released 3 October 2013) to extract all possible SNP locations within each gene. The database provides translations of nucleotide variants into alternate amino acids, which we indexed with respect to the corresponding proteins. All SNP functional predictions (benign vs. damaging) were retrieved from the pre-computed scores for six sequence-based deleterious predictors available from dbNSFP. To resolve discrepancies among prediction algorithms, we assigned levels of deleteriousness using the consensus prediction. A variant is regarded as "deleterious" if $\geq$ 3/6 predictors reported the variant as "deleterious", and as "non-deleterious" if the predictions

suggest otherwise. Later, the initial set of SNPs was filtered such that only variants located in known genes were retained (68,165,196 nsSNP locations in 18,349 genes).

Gene-trait associations were retrieved from the NHGRI Genome-Wide Association Studies (GWAS) catalogue (Hindorff, *et al.*), available from dbNSFP v2.1. Additional information provided at the AACDS website includes essential information about each variant: *i.e.*, dbSNP reference SNP ID number (db138 release, downloaded from the NCBI's FTP site at ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/BED, accessed 16 January 2014), gene name and protein name from UniProtKB, and population-specific minor allele frequencies (retrieved from dbNSFP).

Database construction

AACDS was designed as a relational database on a MySQL server. The data relationships are presented in Figure 1.

In-house Perl scripts were used to extract variant information from the aforementioned data sources. Our original paper describes the AACDS as an 8-level category (variant categories 1, 2A, 2B, 3A, 3B, 4, 5, and 6) (Preeprem and Gibson, 2013). However, many SNPs cannot be exclusively defined into one class; therefore, a maximum of 12 classes are reported in this implementation to represent all distinct conditions possible when joining multiple assigned AACDS categories together (Table 1).

The list of disease associations was collected from SwissVar (accessed 1 November 2013) and MSV3d (released 29 July 2012) databases. We did not attempt to standardise the minor differences of clinical terms provided by the two data sources. Similar association records for a particular SNP or a gene from both SNP databases were dealt with by reporting only the most detailed record. Some SNPs have ambiguous clinical annotations; for example, when one of the two databases documents a SNP as a disease-associated variant, but the other suggests it is a polymorphism or has missing data, the intuition we followed was to regard the variant to have

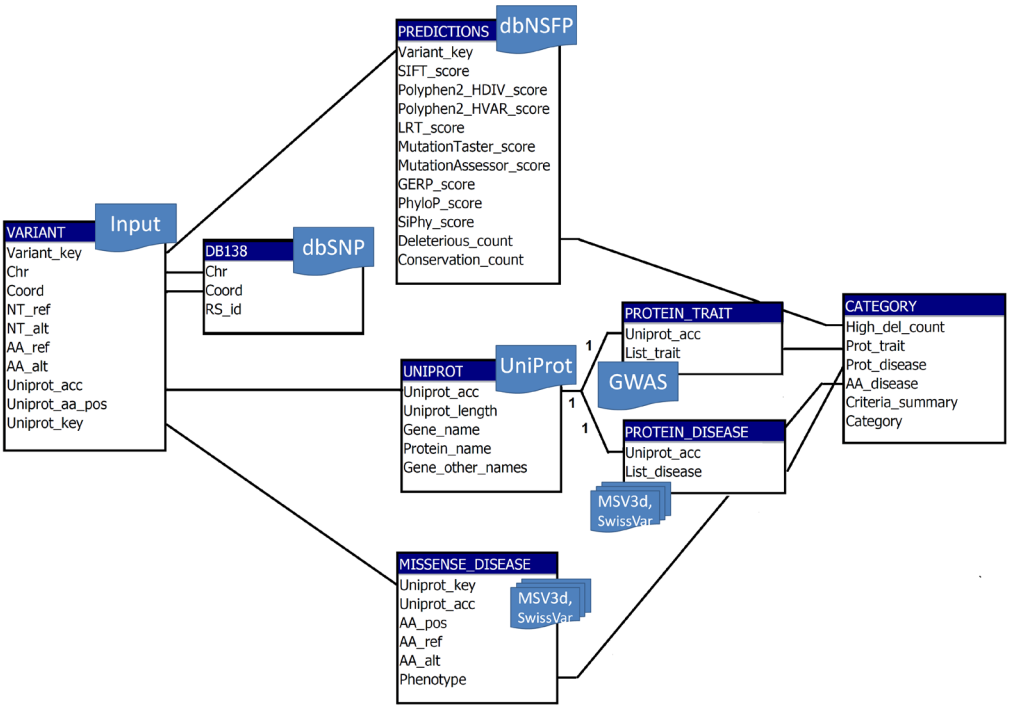


Figure 1. AACDS database schema. The database constructs its data relationships from several sources. The AACDS category for each variant is assigned according to whether the variant has a high deleterious count, whether its gene has GWAS-documented gene-trait/gene-disease association, its gene is database-documented to have disease association, or the variant is documented as a disease-causal variant.

clinical associations. In total, 31,092 instances of variant-disease associations and 4,225 pairs of gene-disease associations were included in our database. The number of genes whose gene-trait associations were identified from GWAS is 5,113.

To ensure that search results are returned quickly, we performed the computation of AACDS for all variants, and utilised the assigned categories as the pre-computed variant classification during Web searching. The online service of the AACDS database was implemented in PHP, MySQL, JavaScript and Apache. The AACDS

Table 1. Descriptions of the 12 combined AACDS classes. Column descriptions for features of nsSNPs are (i) disease-causing, if MSV3d and/or SwissVar indicate the variant is disease-causal; (ii) predicted deleterious, if ≥ 3/6 programmes predict the variant to be deleterious; (iii) in disease gene, if MSV3d and/or SwissVar indicate the gene has disease associations; (iv) in GWAS-documented trait/disease gene, if GWAS indicates the gene has trait/disease associations.

AACDS classes	Features of nsSNPs				Descriptions of nsSNPs
	(i) Disease-causing	(ii) Predicted deleterious	(iii) In disease gene	(iv) In GWAS-documented trait/disease gene	
1	X				disease-causing (but not located in gene with disease- or trait-associations nor predicted as deleterious by most programmes)
1, 2B	X	X	X		disease-causing, predicted as deleterious by most programmes, located in gene with disease-associations (but not GWAS-documented)
1, 2B, 3B	X	X	X	X	disease-causing, predicted as deleterious by most programmes, located in gene with disease and trait-associations
1, 5	X		X	X	disease-causing, located in gene with disease- and trait-associations (but most programmes predicted it to be benign)
2A			X		located in gene with database-documented disease-associations (but no other implications)
2B		X	X		predicted deleterious by most programmes, located in gene with database-documented disease-associations (but not a causal variant)
2B, 3B		X	X	X	predicted deleterious by most programmes, located in gene with disease and trait-associations (but not a causal variant)
3A				X	located in gene with GWAS-documented trait/disease associations (but no other implications)
3B		X		X	predicted deleterious by most programmes, located in gene with GWAS-documented trait/disease associations (but not a causal variant)
4		X			predicted deleterious by most programmes (but no other implications)
5			X	X	located in gene with disease and trait-associations (but not a causal variant nor predicted deleterious)
6					no clinical implications

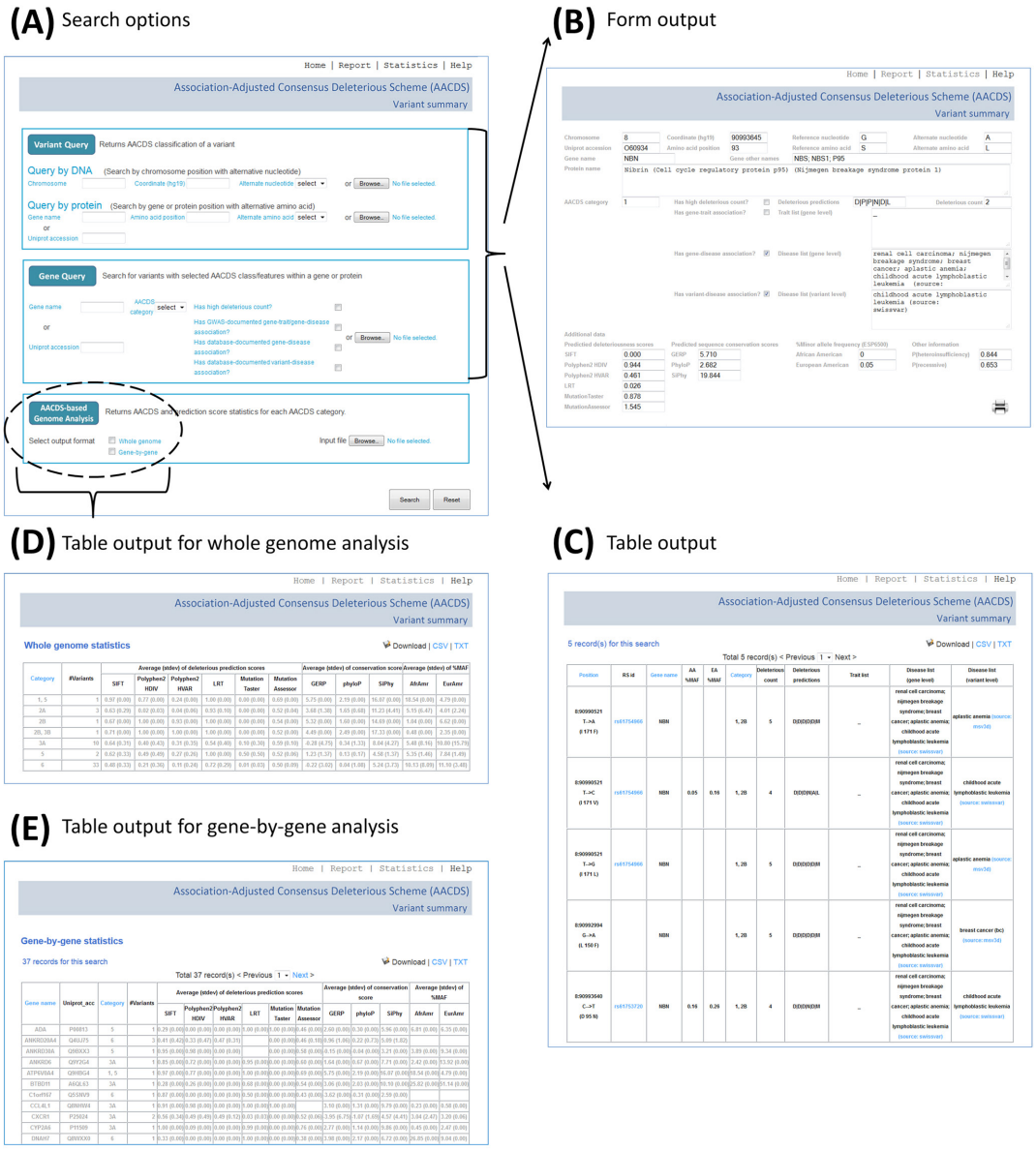


Figure 2. Overview of the AACDS Web interface. (A) The three query options: Variant query, Gene query, and AACDS-based genome analysis. (B-E) Example outputs in form and tabular formats. The form output (B) reports the AACDS category of a variant and its relevant information, along with any additional variant data. Included in the tabular output (C) are direct links to dbSNP and to the original sources of clinical data. The outputs from AACDS-based genome analysis (D-E) present numerical statistics of nsSNPs based on the assigned AACDS classes.

website can be accessed at http://cig.gatech.edu/tools. All standard browsers are supported.

Utility

Our AACDS Web application allows users to retrieve AACDS classifications and relevant information about variants or genes of interest. Figure 2A illustrates the three major components of the website: (1) Variant query, (2) Gene query, and (3) AACDS-based genome analysis. Users can search the AACDS database via single-query or batch mode. Batch mode permits practical analysis of personal genome data, as users can upload lists of variants of unlimited size and retrieve the results in plain-text formats for external use.

Table 2. File formats for batch queries. The following analyses accept a batch search if users provide a .txt file (tab delimited) with a specified format.

Queries	Query options	File formats	Column descriptions
Variant query	By DNA	Chr:10 26781257 T A Chr:10 26781257 T C Chr:10 26781257 T G	1 = chromosome number 2 = hg19 coordinate 3 = reference nucleotide 4 = alternative nucleotide
	By protein (providing gene or protein names)	Gene:AACS 8 G S Gene:GOT1 413 Q H Gene:NT5C2 515 K Q Or Uniprot:Q8IZY2 2000 N K Uniprot:Q86UK0 2000 T A Uniprot:O95477 2000 L R	1 = gene name or UniProtKB accession number 2 = amino acid position (UniProtKB numbering) 3 = reference amino acid 4 = alternative amino acid
Gene query	Providing gene or protein names	Gene:HSD3B2 1 Gene:ABCA12 1 Gene:SH3BP2 1 Or Uniprot:Q86V21 4 Uniprot:P01011 2B Uniprot:Q9NY61 4	1 = gene name, or UniProtKB accession number 2 = AACDS category (1, 2A, 2B, 3A, 3B, 4, 5, or 6)
AACDS-based genome analysis	-	Chr:10 26781257 T A Chr:10 26781257 T C Chr:10 26781257 T G	1 = chromosome number 2 = hg19 coordinate 3 = reference nucleotide 4 = alternative nucleotide

For a single-entry query, users can search for the AACDS classification of their variant of interest by providing some search parameters: for query by DNA, chromosome number, hg19 coordinate and alternative nucleotide; for query by protein, gene name or UniProtKB accession number, amino acid position, and alternative amino acid. The website outputs a variant summary page, which reports the AACDS category of the variant and its relevant information, along with any additional variant data (Figure 2B).

Users can also retrieve lists of gene variants whose characteristics match their interests. If a particular AACDS class is specified, the website returns all nsSNPs that belong to that category. If any of the four features (Table 1, Figure 2A) are specified, a list of variants whose characteristics are compatible with the search features is returned. When more than one variant meets the search criteria, a form (Figure 2B) and summary table (Figure 2C) are returned. The table provides a short description (11 attributes) of the variants; users can also download the complete table (37 attributes) through the "download" button.

We also provide the overall statistics for a set of nsSNPs found in an individual's genome via the AACDS-based genome analysis option. Users can perform the analysis on two levels: whole genome statistics and gene-by-gene statistics – Figures 2D and 2E show example outputs from the two analysis types, respectively. In either case, the schema classifies nsSNPs into several groups, based on the assigned AACDS classes. The results can be ranked by gene names or by AACDS groups. In addition to the number of variants within each AACDS class, the tabular output also presents the average (and the standard deviation) for all six deleterious scores, three conservation scores, and two population-specific minor allele frequencies.

For each of the above analyses, a batch search is possible. The required information for input file formats is described in Table2.

Discussion

The integration of both sequence-based deleterious prediction and clinical-association data in our AACDS algorithm provides a novel approach to integrative variant classification for personal genomes. Manual inspection of a variant for both predicted deleteriousness and phenotypic association is possible, but certainly not practical

for analysing large genome data. For this reason, the implementation of a database-driven Web application is considered an important tool for promoting the utility of the AACDS. We believe that with the scope of our database coverage, both in terms of genomic variations and phenotypic data, this application will help to bring a comprehensive framework of personal genome interpretation to a more practical level.

The current implementation does not have an automatic online update feature, but we will regularly check for new releases of our selected external databases so that it offers AACDS classes for the most complete set of SNPs in a human genome. Further improvements may include subsequent addition of variants in the remaining genes once their curated protein sequences are available, the inclusion of clinical and trait associations from other data sources, and the implementation of an automatic online update with the selected data sources.

Key Points

- Association-Adjusted Consensus Deleterious Scheme (AACDS) is an integrative approach for interpreting genomic variations, using variant deleteriousness predictors and publicly available genomics data.
- AACDS is specifically designed for personal genome analysis (variants likely to be neutral).
- AACDS database covers all missense variants (induce amino acid changes) of over 18,000 human genes.
- AACDS Web application enables the evaluations of variants on a per variant, per gene, and per genome basis.
- AACDS facilities the identification and prioritisation of significant variants.

References

Cooper DN, Chen JM, Ball EV, Howells K, Mort M, *et al.* (2010) Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. *Human Mutation* **31**, 631-655. http://dx.doi.org/10.1002/humu.21260

Gonzalez-Perez A and Lopez-Bigas N (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. American *Journal of Human Genetics* **88**, 440-449. http://dx.doi.org/10.1016/j.ajhg.2011.03.004

Hindorff LA, MacArthur J, Morales J, Junkins HA, Hall PN, *et al.* A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies.

Hu H, Huff CD, Moore B, Flygare S, Reese MG, *et al.* (2013) VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genetic Epidemiology* **37**, 622-634. http://dx.doi.org/10.1002/gepi.21743

Javed A, Agrawal S and Ng PC (2014) Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nature Methods* **11**, 935–937. http://dx.doi.org/10.1038/nmeth.3046

Kim JI, Ju YS, Park H, Kim S, Lee S, *et al.* (2009) A highly annotated whole-genome sequence of a Korean individual, *Nature* **460**, 1011-1015. http://dx.doi.org/10.1038/nature08211

Liu X, Jian X and Boerwinkle E (2011) dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions, *Human Mutation* **32**, 894-899. http://dx.doi.org/10.1002/humu.21517

Luu TD, Rusu AM, Walter V, Ripp R, Moulinier L, *et al.* (2012) MSV3d: database of human MisSense Variants mapped to 3D protein structure, *Database* **2012**, bas018. http://dx.doi.org/10.1093/database/bas018

Mottaz A, David FP, Veuthey AL and Yip YL (2010) Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar, *Bioinformatics* **26**, 851-852. http://dx.doi.org/10.1093/bioinformatics/btq028

Ng PC and Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function, *Annual Review of Genomics and Human Genetics* **7**, 61-80. http://dx.doi.org/10.1146/annurev.genom.7.080505.115630

Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, *et al.* (2008) Genetic variation in an individual human exome, *PLoS Genetics* **4**, e1000160. http://dx.doi.org/10.1371/journal.pgen.1000160

Patel CJ, Sivadas A, Tabassum R, Preeprem T, Zhao J, *et al.* (2013) Whole genome sequencing in support of wellness and health maintenance, *Genome Medicine* **5**, 58. http://dx.doi.org/10.1186/gm462

Preeprem T and Gibson G (2013) An association-adjusted consensus deleterious scheme to classify homozygous Mis-sense mutations for personal genome interpretation, *BioData Mining*, **6**, 24. http://dx.doi.org/10.1186/1756-0381-6-24

Sifrim A, Popovic D, Tranchevent LC, Ardeshirdavani A, Sakai R, *et al.* (2013) eXtasy: variant prioritization by genomic data fusion, *Nature Methods*, **10**, 1083-1084. http://dx.doi.org/10.1038/nmeth.2656

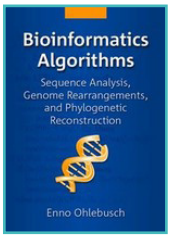
UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt), *Nucleic Acids Research*, **40**, D71-75. http://dx.doi.org/10.1093/nar/gkr981

Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, *et al.* (2012) Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing, *American Journal of Human Genetics* **91**, 1022-1032. http://dx.doi.org/10.1016/j.ajhg.2012.10.015

Acknowledgements

This work was supported by start-up funds from the Georgia Tech Research Foundation to GG, and TP was supported by the School of Biology at Georgia Tech.

Bioinformatics Algorithms - Sequence Analysis, Genome Rearrangements, and Phylogenetic Reconstruction



Enno Ohlebusch

Oldenbusch Verlag, 2013

ISBN: 978-3000413162

604 pp.

Josè R. Valverde

CSIC, Centro Nacional de Biotecnología, Madrid, Spain

Received 11 June 2014; **Published** 16 July 2014

Valverde JR (2014) *EMBnet.journal* **20**, e781. http://dx.doi.org/10.14806/ej.20.0.781

Competing Interest: none

There are many books on bioinformatics in circulation, many of them dealing with issues concerning how analytical methods are implemented, and the algorithms that underpin them. I was therefore curious to know whether a new book on Bioinformatics Algorithms was actually needed, and whether this book really could provide something that others didn't. Consequently, in this review, I will try to give an idea of what the book offers, what you may expect from it, and who is most likely to benefit from it.

The roots of bioinformatics were, and still are, firmly embedded in the management and analysis of biological sequence information. This book therefore focuses on the core technologies that underlie modern sequence analysis in the context of genomics and phylogeny.

The book starts with a general introduction (chapter 1), followed by traditional string-comparison methods (chapter 2), and quickly moves to the core of modern genomic methods: suffix arrays (chapters 3 and 4). Suffix arrays have gained increasing popularity, as they provide an efficient way to perform linearly scaling queries of huge textual data-sets; common practical applications of these algorithms are thus presented in chapter 5. Chapters 6 and 7 then address how to make these algorithms and data structures more efficient, introducing methods that work with compressed data, such as the Burrows-Wheeler Transform. These methods address exact string matching efficiently (in linear time), and have direct applications in problems such as genome assembly and short-read mapping. The traditional approaches to sequence comparison are then introduced in chapter 8, where the Needleman-Wunsch algorithm is described, followed by methods used to build multiple sequence alignments and whole genome alignments, touching on topics such as genome rearrangements. Chapter 9 deals with sorting by reversals to introduce methods that can be used to address these issues. Finally, chapter 10 ties everything together in a clear exposition of a practical application: phylogenetic analysis.

The overall layout is organised as a book on Computer Science (CS). This means that algorithms are minutely described, generally with an accompanying step-by-step walk-through using a small example data-set, followed by detailed algorithm validation and complexity analysis of its time and space requirements in 'big O' notation. The descriptions are clear, concise and illustrated with many opportune Figures, easy to follow and understand. The description of algorithm goals, and validation and complexity analysis, use a formal language that will appeal to pure computer scientists.

The orientation towards CS is also shown in the topic layout: purely algorithmic issues are often presented before their practical application in bioinformatics, often with forward references to later chapters. Many surrogate techniques related to these algorithms (e.g., SVMs, which use the kernel methods presented), alternative basic algorithms (e.g., Smith-Waterman or BLAST) or methodologies (evolutionary algorithms, machine learning, clustering or modern statistical methods, *etc.*) that would deviate too much from the central line of discourse are omitted. Similarly, many biological applications are described only to the extent needed to demonstrate the application of the algorithms (e.g., the chapter on phylogeny describes traditional techniques but does not delve into methods like Bayesian inference); this should not pose problems for computer scientists and practical bioinformaticians who are already familiar with these techniques.

Summarising, this is a very good, readable CS book on the core techniques of sequence analysis, as seen from the point of view of a modern family of algorithms (derived from suf-

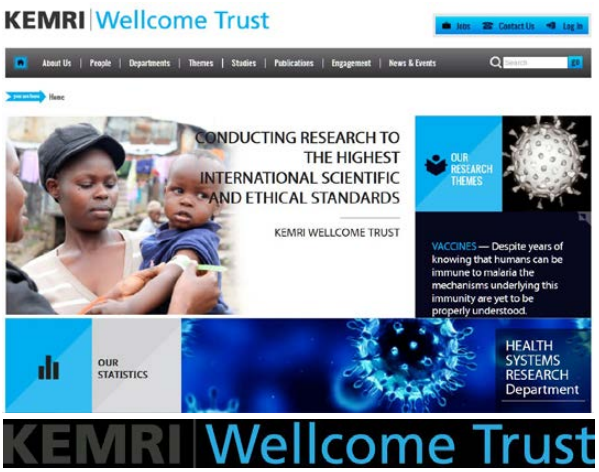
fix arrays) that have acquired major relevance in bioinformatics and other text-analysis fields, and are slowly overtaking most traditional techniques. The book does a good job of presenting them in the context of their application to genome analysis.

Advanced computer scientists will enjoy the detailed formal analysis of most algorithms in the book. The average pragmatic bioinformatician will probably be more interested in the (very good) description of the algorithms and their practical consequences (time and space complexity). Hence, I believe the book is most likely to appeal to seasoned computer scientists and CS students, but will also appeal to practical bioinformaticians who want to get up to date in modern genomics research, and to practitioners of other text-analysis fields.

After reading it in full, I enjoyed this book and found it informative, inspiring and entertaining. It is well written and readable, and although it is not a general bioinformatics text, it succeeds in explaining a complex family of methods that underlie novel applications.

KEMRI-Wellcome Trust Research Programme (KWTRP)

by Etienne de Villiers



The KEMRI-Wellcome Trust Research Programme, KWTRP (http://www.kemri-wellcome.org/), was formally established in 1989 as a partnership between the KEnyan Medical Research Institute (KEMRI), Oxford University and the Wellcome Trust.

It conducts basic, epidemiological and clinical research in parallel, with results feeding directly into local and international health policy, and aims to expand the country's capacity to conduct multidisciplinary research that is strong, sustainable and internationally competitive. Strong community links are at the heart of the Programme, with a particular emphasis on capacity building and training to build scientific leadership.

The advent of Next Generation Sequencing (NGS) platforms has opened exciting new research avenues to life scientists. To support application of NGS within the Programme, the KWTRP-Bioinformatics Core (KBiC) was established with the aim to provide a single point of contact for computational biology, providing a venue for applying genomic approaches to basic biology, epidemiology and bioinformatics to develop novel approaches to improve human health. By providing KWTRP researchers with access to a bioinformatics infrastructure and expertise, the KBiC opens new areas of research, enhances the quality and consistency of high-throughput data analysis, and improves the Progamme's ability to support research in this area.

As a further important commitment, the core periodically organises courses and workshops in order to train biologists in the implementation of the main bioinformatics tools in their research.

Contact:
http://www.kemri-wellcome.org/index.php/en/contact_page

Protein structure databases and resources at the CMBI

by Gert Vriend

Many of 'us' are involved in some form of sequence analysis, but the EMBnet community also includes protein structure bioinformaticians. When crystallographers or NMR spectroscopists solve macromolecular structures, the coordinates must be made available before publication. The international wwPDB collects these structures, annotates them and releases them to the world in formatted, keyword-based PDB files – PDB's data collection started back when data storage was done with punched cards, so the PDB format is still fixed at 80 characters per line.

PDB files are routinely used in biomedical research. Unfortunately, they all contain errors owing to poor data, human error, the fact that supercomputers once had less CPU power than today's mobile phones, and that we didn't know so much about protein structures back then. We therefore set out to use today's software and computers to redo all PDB files [1]. The results are available in the PDB_REDO database. We can't correct all errors, so we release error reports for all PDB files in the PDBREPORT database [2]; we also maintain DSSP [3], holding records of secondary structures for each protein-containing PDB file, and HSSP [4], containing sequence alignments for all proteins. For those wanting to search the PDB, we maintain PDBFINDER [5], which holds the essential metadata of PDB entries in an easy-to-parse format. And our user-friendly MRS [6] search engine can, in <1 second, query all these databases. For nucleic acid structures, where DSSP, HSSP, *etc*. files don't exist, the WHY_NOT server [7] explains why a file is missing.

These databases (some from before the Internet) are regularly maintained. For users who want other things, we made a series of ~60 Web servers [8,9], and will make new servers upon request; and if you want to perform structure calculations on large numbers of files, you can get programmatic access to our software through Web services [10] – we also write Web services on request.

All of these and other resources are available via http://swift.cmbi.ru.nl/gv/facilities/. It's no longer the same hardware, but swift was one of the first 2,000 computers attached to the Internet, and certainly the first computer ever on the Internet to provide bioinformatics services. I'm not sure if that makes me sad or happy.

EMBnet Workshop & AGM 2014, Lyon, FR

by Teresa K. Attwood

EMBnet's 2014 AGM and associated events were hosted in the Hôtel de la Cité and on the Doua campus of the University of Lyon, from 26 to 30 May. Included were a 1-day tutorial entitled "*From NGS data through the third dimension towards new agrochemicals and drugs*"; the 2-day *Bioinformatics for Environmental Genomics* workshop of the Pluridisciplinary Thematic Network in Environmental Genomics; a 1-day EMBnet workshop; and the traditional business meeting.

These were stimulating events, serving to highlight the achievements of the last year, and how much remains to be done (full details will be reported in *EMBnet.journal*). Discussions were wide-ranging and productive: it was agreed to: i) review the structure and membership of the Committees, publish their current and future projects on the website, and deliver tangible outcomes; ii) create a Fellowship Programme; iii) formulate details of a Service Award scheme; iv) organise a meeting alongside the final AllBio AGM, to discuss development of global bioinformatics MSc curricula; v) invite members of target groups to join EMBnet, following the launch of the Fellowship Programme; and vi) develop a core for future tutorials, and ground rules for running and hosting them, aiming to expand the programme and recover costs.

The work of the last year has been led by a group of dedicated individuals, who are responsible for running *EMBnet.journal*, coordinating AllBio and SeqAhead, leading GOBLET, and championing EMBnet's PR activities. This year, we celebrated the arrival of a new individual member, Axel Thieffry, whom we hope will help to drive some of EMBnet's new initiatives forward, and especially to inspire and recruit more members!

As always, there's more to be done. We therefore warmly encourage you to contribute your energies and visions to EMBnet, to ensure its continued success as the Global Bioinformatics Network!

EMBnet.digest

EMBnet.Spotlight is a quarterly release of InFocus sections published in EMBnet.digest (www.embnet.org/embnet-digest), EMBnet's monthly publication that provides a round-up of news from the community. The InFocus section features member activities, projects, initiatives, *etc.*, especially from new members, that may be of interest both to the network and to EMBnet's associated communities, societies and projects.

EMBnet membership goes individual

by Axel Thieffry

EMBnet AGM 2014: Axel & Domenica sharing the passion of the P&PR PC!

From its creation in the spring of 1988, EMBnet never stopped growing and gathering forces through a membership process based on the ratification of National Nodes, Specialist and Industrial Nodes.

Since 2013, opportunities for individuals to become involved and participate in EMBnet's activities have been made available, providing a wide range of training, collaboration and professional networking opportunities.

As such, I am one of the very first to join EMBnet as an individual member. I am a bioinformatics engineer, from Belgium, having specialised in next generation sequencing data analysis during my Master degree at the Swedish University of Agricultural Sciences, in Uppsala, Sweden. Aside from working for a private French biotech company, I have now been warmly welcomed by EMBnet's Publicity & Public Relations Project Committee (P&PR PC) to help with drafting and releasing the monthly *EMBnet.digest*, bringing new ideas and generally strengthening the team. Numerous projects are currently under development and require the input and collaboration of people with skills from different horizons; the recently published Quick Guide on the Velvet & Oases *de novo* assemblers is a simple, yet concrete, example.

The reasons for joining EMBnet as an individual member are legion. As for me, I'm convinced it's vital to stay current in the methods, techniques and new software tools that bioinformatics is generating at an outstanding pace. The numerous international events, workshops and conferences in which EMBnet is involved are great ways to fulfil this need, while developing a robust and friendly international network. As an individual member, I gladly experienced that I can express ideas and opinions, which are then carefully considered and discussed in a team-oriented manner.

Finally, EMBnet allows me to communicate my passion for Science and Bioinformatics, to share knowledge through the production of materials (*e.g.*, like Quick Guides) and *EMBnet.journal* publications, and to contribute ideas for the website, opportunities that I believe are appealing for junior and established scientists alike.

Getting COSI with a GOBLET at ISMB

by Terri Attwood

ISMB 2014 has just taken place in Boston, USA. While the buzz of the conference still rings in my ears, I thought it would be a good time to offer an update on some of its key events. As usual, it was a packed conference, and it was exhausting just trying to navigate the myriad parallel scientific sessions, special sessions, keynotes, satellite meetings, Special Interest Groups (SIGs), workshops, tutorials, Birds-of-a-Feather (BoF) and poster sessions!

Ironically, at an event this large, it's easy to feel isolated. Therefore, to try to combat the sense of 'disconnection' that many feel during this meeting, the ISCB launched a new initiative to create Communities of Special Interest (COSIs), where smaller groups of like-minded scientists can meet more readily within the main conference. One of these COSIs was on Computational Biology Education (CoBE). CoBE aims to foster a collaborative community in which bioscientists can share education and training resources and experiences, and facilitate the development of education programmes, courses, curricula, teaching tools and methods: it especially seeks, at least initially, to bring the ISCB and GOBLET communities closer together.

Other significant events within the conference were ISMB's second education poster track, where GOBLET presented some preliminary results from the 'training needs' survey it disseminated earlier this year, in part via EMBnet's email lists (see February InFocus). There was also an interesting, well-attended workshop on education in bioinformatics (WEB 2014), featuring the online world of bioinformatics education, with talks on MOOCs and gamification, followed by a panel discussion on the merits and pitfalls of online; finally, there was a GOBLET lightning presentation at the BoF session on curriculum development.

I'm happy to report that, throughout the event's formal meetings, GOBLET received a lot of support from ISCB's outgoing president, Burkhard Rost. Hopefully, with the ISCB and EMBnet communities behind it, GOBLET will go from strength to strength – I encourage you to get involved.

GOBLET website: www.mygoblet.org

A Hive of Activity at CPGR

Part I

by Judit Kumuthini

KNOWLEDGE TRANSFER PROGRAMME for Bioinformatics

The Knowledge Transfer Program (KTP) is an initiative to bring together scientific experts and trainees for the purposes of transferring knowledge in a natural and cost effective way in Africa. KTP targets postgraduates, postdoctoral fellows and researchers.Expert knowledge providers are selected through peer review and matched to specific projects based on requests from a PI and the level of training required.

Bioinformatics experts are invited from around the world to establish an **accessible online database of expertise** for the **KTP**. Experts can be selected from the bioinformatics experts database and assigned to relevant training projects using a match-making method.

To become part of KTP's expert panel requires prospective experts to register on the KTP website to upload their personal details and Curriculum Vitae. Applicants will be screened by a **Scientific Advisory** and **Review Committee** to assess an applicant's suitability as an expert using a predetermined scoring system. Applicants will be grouped as either junior or senior experts based on the Review Committee's assessment.

It is free to register and become involved in the KTP as an internationally recognised Bioinformatics expert. Register today on the **KTP's website** and see the benefits of becoming a KTP expert!

INGENUITY PATHWAY ANALYSIS

CPGR, in collaboration with QIAGEN, offer annual online trainings to develop reliable knowledge capacity on how to model, analyse and understand multifaceted biological and chemical systems at the core of life science research **http://www.cpgr.org.za/training/online-training/ingenuity-pathway-analysis/**.
Course registration will open with CPGR in the near future.

EMBnet The Global Bioinformatics Network

EMBnet.digest

EMBnet.Spotlight is a quarterly release of InFocus sections published in EMBnet.digest (www.embnet.org/embnet-digest), EMBnet's monthly publication that provides a round-up of news from the community. The InFocus section features member activities, projects, initiatives, *etc.*, especially from new members, that may be of interest both to the network and to EMBnet's associated communities, societies and projects.

A Hive of Activity at CPGR - Part II - *by Judit Kumuthini*

South Africa joins COST Action SeqAHead

South Africa joined COST action BM1006 in 2013. Current members are Prof. Fourie Joubert (University of Pretoria), A.Prof. Nicola Mulder (University of Cape Town (UCT)) and Dr. Judit Kumuthini. To help us assess the demand for future NGS workshops, please fill in the COST action questionnaire.

First Genomics Lab in Africa To Be ISO Certified

The CPGR selected ISO 9001:2008 as the standard for its Quality Management System. We believe that compliance with this standard will enhance our ability to support projects across the entire life science innovation chain, situated in the basic, translational and clinical science arenas. The ISO 9001:2008 standard is premised on a culture of continuous improvement, from product development to client engagement. This requires a commitment to flexibility and adaptation – crucial organisational features in the highly dynamic area of science and business such as ours.

The New Website and NGS Lab Launch

CPGR's website recently had an overhaul. In May 2014, CPGR launched NGS services, with acquisition of Illumina MiSeq, Ion Proton and two Ion PGM bench-top sequencers. CPGR is implementing complete genomics workflows in RNA-Seq, microbial and fungal genome sequencing, metagenomic sequencing and whole exome sequencing.

CPGR Workshop Trainings Attended

H3ABioNet is geared towards developing technical bioinformatics capacity and support. To this end, the consortium has hosted multiple training workshops: H3ABioNet Data Management, H3ABioNet NABDA Node Training Course on Visual Analytics of Human Genome Variation Datasets, H3ABioNet Introduction to Bioinformatics using the eBioKit platform, H3ABioNet Train-the-Trainer Bioinformatics Course, H3ABioNet Technical Training Course, and H3ABioNet Grants Management Course. CPGR node members have been fortunate to participate in all these workshops.

CPGR Workshop Trainings Hosted

CPGR has also hosted several workshops: 1) a 2-day NGS workshop funded by H3ABioNet, organised through the KTP at the CHPC. This involved lectures and computer exercises, with hands-on data analysis of NGS applications for variant detection and *de novo* assembly – 31 participants attended from across South Africa (see newsletter); 2) a workshop based on Chipster, held at the CHPC, funded by the DST through the SANBI, and jointly organised by the CPGR, SANBI and the Computational Biology Group (Cbio) at UCT – 30 participants attended; 3) a 3-day workshop, the first African Affymetrix University Data Analysis workshop, held at SANBI at the University of the Western Cape (UWC) in Cape Town, funded by the DST and jointly organised by the CPGR, Cbio, SANBI and Affymetrix – it had 30 participants; 4) with help from Dr. Panu Somervuo from the University of Helsinki (Metapopulation Research Group), a 3-day NGS road-show was delivered at UCT, UWC, Stellenbosch University and the African Institute for Mathematics and Statistics. Other workshops included an introduction to computational medical population genetics and genome-wide association meta-analysis for complex disease, CPGR data analysis and CPGR GeneTitan Axiom Training. CPGR also attended Monte Carlo Inference for High-Dimensional Statistical Models and NGS data after the Gold rush workshops this year.

2015 Training Activities at SIB

by Grégoire Rossier,
Vital-IT and Training & Outreach groups, SIB

The SIB Swiss Institute of Bioinformatics has extensive training experience, built over the last years. Its ***Training and Outreach Group***, which provides and coordinates a growing number of courses and workshops each year, has recently published the list of training-related events scheduled for 2015 (www.isb-sib.ch/training.html).

At the time of writing, this list detailed 26 events – spanning around 65 days of training – covering key bioinformatics topics, such as large-scale data analysis, systems modelling, statistics, data mining, programming, High Performance Computing and SIB resources. To address an increasing demand for basic training in specific topics, the SIB inaugurated, last September, a new "**First Steps with**" series, with a one-day course about UNIX for biologists. The course was a success, and will be repeated three times per year in different locations in Switzerland. Other "First Steps with" are also planned in topics such as statistics, R, sequence analysis and comparison, *etc*.

Additional SIB courses are currently being devised, and aim to reach at least 35 events next year. Currently, the courses are held at six SIB locations in the main Swiss cities, and are open to international participants, who currently only represent 10% of registrations.

Centro de BIOINFORMÁTICA
Instituto de Biotecnología
Universidad Nacional de Colombia

CBIB 2015 Activities

by Emiliano Barreto Hernandez, Colombia EMBnet Node Manager

The Bioinformatics Centre of the National University of Colombia Biotechnology Institute (CBIB) has provided bioinformatics tools, databases, training and support to the Colombian research community for 14 years. As part of EMBnet since 2003, the CBIB has been both witness to and actor in the development of bioinformatics in Colombia.

As part of the bioinformatics community in our country, the CBIB is working side-by-side with Universidad de Antioquia, Universidad EAFIT (Escuela de Administraciòn, Finanzas y Tecnologìa) and other academic partners in the organisation and planning of the 3rd Computational Biology and Bioinformatics Colombian Congress, which will take place in Medellín, September 2015 (http://ccbcol.co, under construction). It promises to be the meeting point for more than 200 Colombian researchers.

The CBIB has helped and inspired the National University of Colombia to create the 1st Bioinformatics Masters program in Colombia, taking its first students in February 2014, and continuing to receive students every semester. The next application process will commence in March 2015.

Likewise, in June 2015, the university has proposed a Research School, where their bioinformatics research groups, including the CBIB, will be responsible for planning and conducting a "*Clinical Genomics and Personalised Medicine*" course with the collaboration of international experts. The main aim of this course is to familiarise the research community with the theoretical, practical and clinical applications of genomics and personalised medicine (more info in February 2015 at: www.unal.edu.co/diracad/einternacional).

EMBnet.digest

EMBnet.Spotlight is a quarterly release of InFocus sections published in EMBnet.digest (www.embnet.org/embnet-digest), EMBnet's monthly publication that provides a round-up of news from the community. The InFocus section features member activities, projects, initiatives, *etc.*, especially from new members, that may be of interest both to the network and to EMBnet's associated communities, societies and projects.

International Society for Biocuration

In this issue:

DATABASE, official journal of ISB

DATABASE, The Journal of Biological Databases and Curation, is now the official journal of the International Society for Biocuration.

The International Society for Biocuration was created to promote biocuration, the product of multidisciplinary teams of database curators, software developers and bioinformaticians. Biocurators, whose work facilitates research and education across the life sciences, create and maintain a wide variety of online tools and databases essential to the biological community in their daily work. Such important efforts now have a rightful home at DATABASE.

DATABASE, The Journal of Biological Databases and Curation, supports the growing need of the research community to discuss a range of issues related to the creation, development and maintenance of biological databases, and to strengthen communication between database developers, curators and users. As this resonates strongly with the mission of the ISB, we are delighted to announce that DATABASE has now become the Society's official journal. DATABASE has published more than 250 papers, 50 of which have appeared in the Biocuration Virtual Issue, a special collection of articles describing work presented at the annual International Biocuration Conference.

The ISB Executive Committee.

Dr. Jennifer Harrow

Dr. Melissa Haendel

ISB Executive Committee Changes

ISB has a new chairperson and two new members of the executive committee.

ISB has elected a new Chairperson.

Congratulations are in order for **Dr. Alex Bateman**! Alex was elected as Chair of the International Society for Biocuration in an election held by the Executive Committee.

After many years of service, we are very thankful to **Dr. Pascale Gaudet** for her contributions and hard work for the improvement of the International Society for Biocuration since its inception. Pascale remains a member of the ISB EC.

The New ISB Executive Committee.

Please join us in giving a warm **Welcome!** to **Melissa Haendel** and **Jennifer Harrow** to the Executive Committee. We bid farewell and sincerely thank Renate Kania and Chisato Yamasaki, who served on the committee since 2011.

To read more about the work of Dr. Melissa Haendel, please visit http://goo.gl/R7ztlX. Learn more about Dr. Jennifer Harrow's career at http://goo.gl/735tGm

The Executive Committee is also composed, Monica-Munoz-Torres (Secretary), Marc Robinson-Rechavi (Treasurer), Teresa Attwood, Alex Bateman, J. Michael Cherry, Pascale Gaudet, and Claire O'Donovan.

Micro-Grants for Regional ISB Meetings!

ISB Micro-Grants are meant to sponsor local/regional short meetings of ISB members to synergize their work efforts, to generate a space to share your work, and to further help to advance the goals of the society.

Micro-Grants are awarded in the amount of (US) $250 per group, and applications may be submitted at any time (i.e. no deadlines). If awarded, the group will write a 300-word report informing the ISB membership about the outcomes of the meeting. This report must be sent to intsocbio@gmail.com no later than 30 days after the end of the meeting. The report will be also posted on the ISB Newsletter.

To apply, members fill out and sign a form and submit a description of the purpose of the meeting, its target audience and possible affiliations, and how the meeting will benefit the members of the ISB community. To request and submit the filled out form, please contact us at intsocbio@gmail.com with the subject line 'ISB Micro-Grants Application'.

Apply Today!

News & Views from around the ISB Community

Register now for Biocuration 2014!

The Organising Committee for the 7th International Biocuration Conference is very glad to announce that the website and registration are now open. Please visit http://biocuration2014.events.oicr.on.ca/

The conference will be held at Hart House, in Toronto from 6-9th April 2014 and the Biocuration 2014 organising committee is very much looking forward to seeing you there!

Our four confirmed keynote speakers will be **Dr. Tim Hubbard** (Wellcome Trust Sanger Institute), **Dr. Suzanna Lewis** (Lawrence Berkeley National Laboratory), **Dr. Patricia Babbitt** (California Institute for Quantitative Biosciences (QB3)), and **Dr. Lincoln Stein** (Ontario Institute for Cancer Research).

Early bird registration rates apply until 7th March 2014. The organisers have secured discount rates at three hotels in Toronto, and you can find more information about booking on the conference website.
The deadline for the abstract submission to present at the conference is 10th February 2014.

See you in Toronto next April!

CAFA 2: The Second Critical Assessment of protein Function Annotations

We are pleased to announce the Second Critical Assessment of protein Function Annotation (CAFA) challenge. CAFA 1 was highly successful experiment, involving 23 groups from around the world. In CAFA 2 we will evaluate the performance of protein function prediction tools and methods, and also expand the challenge to include prediction of human phenotypes associated with genes and gene products. As the last time, CAFA will be a part of the Automated Function Prediction Special Interest Group (AFP-SIG) meeting that will be held alongside the ISMB conference. AFP-SIG will be held as a two-day meeting in July 2014 in Boston, MA (USA).

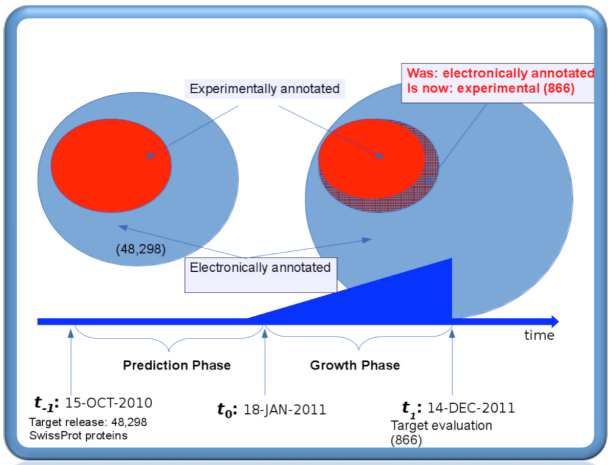
The targets and all information about the CAFA challenge are now available at http://biofunctionprediction.org. The submission deadline for predictions is January 15, 2014. The initial evaluation will be done during the AFP-SIG meeting in Boston.

Anyone in the world is welcome to participate!

How to participate in CAFA 2:
1. Download target proteins, available August 27, 2013
2. Submit predictions on or before January 15, 2014
3. Join us at the AFP-SIG, July 11-12, 2014 in Boston for the Eighth Automated protein Function Prediction meeting, to hear the CAFA 2 results, to present your work, and to learn about the latest research in computational protein function prediction

More information is available at http://biofunctionprediction.org and via email at cafa.2014@gmail.com

Submitted by Iddo Friedberg, CAFA/AFP co-chair



Online Crowd-Verification of Biological Networks – Be a Part of the Crowd!

The Network Verification Challenge (NVC) – part of the sbv IMPROVER program - is a crowd- approach for the verification and expansion of pre-defined biological networks (*Bioinform Biol Insights*, 2013).

What is it?

sbv IMPROVER is a challenge-based program with a specific focus on the verification of industrial research processes related to systems biology (read more on www.sbvimprover.com).

Biological networks, with their structured syntax, are a powerful way of representing biological information; however, they can become unwieldy to manage as their size and complexity increase.

In NVC, web-based graphical interfaces are used to visualize biological relationships. Crowdsourcing principles enable participants to annotate these relationships based on literature evidences. Gamification aspects are incorporated, to encourage biological domain experts to gather robust peer-reviewed information from which relationships can be identified and verified.

Why participating?

Best performers in the crowd-verification phase will be invited to a 3-day "jamboree" to resolve controversies with subject matter experts, finalizing and publishing the network models.

The resulting network models will represent the current status of biological knowledge within the defined boundaries. For some period following conclusion of the challenge, the published models will remain available for continuous use and expansion by the scientific community.

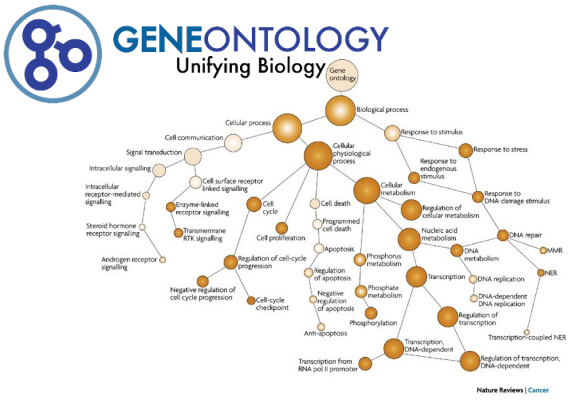
How to participate?

Learn more on the NVC website www.sbvimprover.com/nvc and become part of the scientific community improving the understanding of biological processes.

The project team includes scientists from Philip Morris International's (PMI) Research and Development department and IBM's Thomas J. Watson Research Center. The project is funded by PMI.

Submitted by
Immanuel Luhn, Marianne Charaf, and Dr. Julia Hoeng

Ten Quick Tips for Using The Gene Ontology



The Gene Ontology (GO) provides core biological knowledge representation for modern biologists, whether computationally or experimentally based. GO resources include biomedical ontologies that cover molecular domains of all life forms as well as extensive compilations of gene product annotations to these ontologies that provide largely species-neutral, comprehensive statements about what gene products do. Although extensively used in data analysis workflows, and widely incorporated into numerous data analysis platforms and applications, the general user of GO resources often misses fundamental distinctions about GO structures, GO annotations, and what can and can not be extrapolated from GO resources. The article referenced here offers ten quick tips for using the Gene Ontology. To read more, visit http://goo.gl/8EQKpu

Submitted by Judith Blake.

Upcoming Conferences

6th International Workshop on Semantic Web Applications and Tools for the Life Sciences (SWAT4LS 2013)
http://www.swat4ls.org/workshops/edinburgh2013/
When: Dec. 9-12, 2013
Where: Edinburgh, UK
What: SWAT4LS is a workshop that provides a venue to present and discuss benefits and limits of the adoption of Web based information systems and Semantic technologies in biomedical informatics and computational biology.
Notes: Proceedings of SWAT4LS 2013 will be published in CEUR Workshop proceedings (http://ceur-ws.org/)

International Workshop on Knowledge Support for Deep Phenotyping, co-located with IEEE International Conference on Bioinformatics and Biomedicine
http://skeletome.org/deepphenotyping2013/
When: Dec 18 - 21, 2013
Where: Shanghai, China
What: In this workshop we propose to trigger a comprehensive and coherent approach to studying (and ultimately facilitating) the process of knowledge acquisition and support for Deep Phenotyping by bringing together researchers and practitioners that include but are not limited to the fields of computational biology, genomics, clinical genetics, pharmacogenomics, healthcare, text/data mining and knowledge discovery, and knowledge representation and ontology engineering.
Notes: Extended versions of all accepted full papers will be included in a Special Issue of the Journal of Biomedical Semantics.

Job Opportunities

Bioinformatics Software Engineer at dictyBase, Chicago, IL, USA. Posted November 13, 2013. Details available at http://dictybase.org/dictybase_jobs.html

Content Operations Manager, Qiagen, Redwood City, US. Posted November 13, 2013. More information available at http://goo.gl/6FLafh

Ontology Engineer, Qiagen, Redwood City, US. Posted November 13, 2013. Find more about this position at http://goo.gl/n3Z8py

Bioinformaticians, including in training and outreach, University of Edinburgh, UK. Posted October 31, 2013. Details available at http://goo.gl/XpDZpS

Biocurator Position at Institut Curie in Computational Systems Biology of Cancer Department, Construction of Atlas of Cancer Signaling Networks (ACSN), Paris, France. Posted October 31, 2013. More information available at http://goo.gl/u6GP6I

Knowledge Engineer, Novartis, Cambridge, MA. Posted October 31, 2013. Find more about this position at http://goo.gl/n7zgj0

Scientific Data Curator, Novartis, Cambridge, MA. Posted October 31, 2013. Details available at http://goo.gl/63JwEi

Bioinformatician, PlantLink, Swedish University of Agricultural Sciences. Posted October 31, 2013. More information available at http://goo.gl/SyjuIM

Project Director, VIVO. Posted September 22, 2013. More information available at http://goo.gl/OA9EPP

Postdoctoral scholar position at Plant Pathway database, Oregon State University, OR, USA. Posted September 22, 2013. Find more about this position at http://goo.gl/Fv0dBM

Bioinformatics Analyst, Mouse Genome Database (MGD), Bar Harbor, MA, USA. Posted September 8, 2013. Details available at http://goo.gl/uxdpjH

Scientific Curator, Mouse Genome Database (MGD), Bar Harbor, MA, USA. Posted September 8, 2013. Find more about this position at http://goo.gl/3iRlLq

Community Manager, National Center for Biomedical Ontology (NCBO), Stanford University, CA, USA. Posted August 13, 2013. Details available at http://biocurator.org/jobs/NCBO_Community_Manager.pdf

Job Opportunities (Ctd)

Data wrangler, Oregon Health & Science University, Eugene, OR, USA. Posted August 4, 2013. More information available at http://biocurator.org/jobs/Data_wrangler_OHSU.pdf

Senior Software engineer, Genestack, Cambridge, UK. Posted May 24, 2013. More information can be found at http://www.genestack.com/careers

Software Engineer, Zebrafish Model Organism Database, Eugene, OR, USA. Posted March 15, 2013. Find more about this position at http://jobs.uoregon.edu/unclassified.php?id=4196

Executive Committee Meetings

The ISB Executive Committee meets monthly. Minutes from the meetings are posted on the ISB website http://www.biocurator.org/executive_committee_minutes.html

ISB Newsletter Archive

Previous issues of this Newsletter can be found on the ISB website at http://biocurator.org/newsletter.shtml
M. Munoz-Torres edits the ISB Newsletter.

Share your ideas with members of ISB!

Prepare a news article written to biocurators as the audience. In 260 words or less spread the word about new tools in your site, or tools for distribution to the wider community. Broadcast announcements and advances from your team to all the members of ISB by sending an email to intsocbio@gmail.com. Submission deadline for the November Newsletter is Friday, December 6th at midnight (UTC).

Kind regards,
The ISB Executive Committee.

ISB Spotlight provides a snapshot of some of the work and activities of members of the International Society for Biocuration (ISB). The Spotlight features brief descriptions of a range of databases and biocuration tools, re-published, with permission, from the 'News & Views' section of ISB's monthly newsletter. The newsletter, with the complete 'News & Views from the ISB Community', is freely available from http://biocurator.org/newsletter.shtml.

a pain soothed

Vivienne Baillie Gerritsen

Pain is part of an animal's life. It is there to tell us that something is wrong, and needs to be attended to. There is moral pain. And physical pain, the more definable of the two, which serves two purposes. The first, to warn us of tissue damage and, more often than not, its localisation. The second, to understand where danger lies, so as to avoid it in the future. Unless, of course, it has been lethal. Ever since Life emerged, Nature has been using pain as a means of communication. Though perhaps violent, it is usually very conclusive, which is why many animals have developed toxins they inject into potential predators to ward them off. Among these toxins are the well-known venom cocktails snakes, scorpions and spiders are able to conjure up. In answer to this, a few animals have developed mechanisms to ease the pain – or even suppress it altogether. This is the case of one species of mouse – the Southern grasshopper mouse from the Texan desert – who feels next to no pain when stung by the bark scorpion. As a consequence, the mouse is able to ignore the sting and eat the scorpion. Recent studies have demonstrated that this extraordinary ability is due to changes in the structure of a given type of pore: sodium channel protein type 10 subunit alpha, or Nav1.8.

Feel no pain, by Kim Roberti

Courtesy of the artist

Reducing the effects of pain sounds very attractive. Which it is, if the origin of the pain is known and you wish to alleviate it. However, if an animal is insensitive to pain, or has a pain threshold which is high, it may not be aware of damage made to a part of its body. In this respect, there is the very rare infliction found in humans, known as congenital indifference to pain (CIP)[1]. A well-known case is that of a young Pakistani street performer who was able to run knives through his arms or walk on red hot coal, without ever feeling pain. His body, however, suffered since he frequently ended up visiting the hospital for repair. And he died at the very early age of fourteen[1].

So, if the Southern grasshopper mouse – *Onychomys torridus* – has developed resistance to bark scorpion venom, it must be for a very good reason. And, the reason is: food. This specific mouse lives in the deserts of North America and Mexico, a part of the world where the usual diet – fruit and grains – of a rodent is difficult to come by. So, instead of finding another place to live, the grasshopper mouse developed a system so that its means of sustenance became the animal it shares the desert with, i.e. the bark scorpion, or *Centruroides sculpturatus*. This process must have taken a very long time from an evolutionary point of view but the outcome is surprising. The grasshopper mouse takes hardly any notice of the scorpion's multiple stings, and even begins to feast on it by beginning with the stinger and the bulb which contains the venom.

Why does this particular mouse feel nothing in response to the scorpion's venom? In animals, pain is transmitted from its origin to the spinal cord and the brain, via sensory neurons known

as nocireceptors. In nocireceptors, the pain signal is relayed via transmembrane channels that are scattered along their length. In mammals, acute pain is transmitted to the central nervous system by way of two specific voltage-gated sodium channels: Nav1.7 and Nav1.8. The former initiates the pain signal, while the latter makes sure it is propagated. In the case of the grasshopper mouse, pain is actually triggered off quite normally, while Nav1.8 – instead of propagating the pain signal – stops it from going any further. As a result, the mouse does feel a little sting, of little consequence, however, because Nav1.8 checks it and the scorpion's venom ends up acting, in effect, as an analgesic.

On the molecular level, what is happening? Nav1.8 is a transmembrane protein made up of four domains, each of which has six transmembrane segments. It is the second domain which interacts with the scorpion's venom, itself composed of multiple small peptides. When comparing the sodium channel's amino-acid sequence of the grasshopper mouse and the common mouse, *Mus musculus*, the scientists discovered that there were a number of mutations in the second domain – the peptide-binding domain. More specifically, there is one important amino-acid change where glutamine is replaced by glutamic acid. This amino-acid swap hinders the transmission of the pain signal, though it is not yet known exactly how. Could it be that a venom peptide simply blocks the channel pore so that transmission is arrested? Or do venom peptides bind to the inner side of the pore, producing the same effect?

What is remarkable here is that time has thought up a strategy, not to modify the target channel of venom peptides –Nav1.7 – but to make changes to a secondary channel, Nav1.8. Moreover, scorpion peptides typically activate channels and hence the current whereas, in the case of the grasshopper mouse, the current is checked, and the peptides act as painkillers. *C. sculpturatus* and *O. torridus* provide a first-class model for understanding how pain is transmitted, and for designing novel drugs in the treatment of pain. What is more, Nav1.8 has more sequence diversity in mammals, making it a far better target for drug design than Nav1.7, which is highly conserved with other sodium channels in the brain and the body. There is no doubt that pain will continue to divert many a scientist, as it continues to be the silent guide of animal instinct.

Cross-references to UniProt

Sodium channel protein type 10 subunit alpha, *Onychomys torridus* (Southern grasshopper mouse) : P0DMA5

References

1. Rowe A.H., Xiao Y., Rowe M.P., Cummins T.R., Zakon H.H.
Voltage-gated sodium channel in grasshopper mice defends against bark scorpion toxin
Science 342:441-446(2013)
PMID: 24159039

2. Sutherland S.
Evolutionary adaptation turns painful toxin into analgesic
Pain Research Forum (1 November 2013)
http://painresearchforum.org/news/33366-scorpion-toxin-blocks-nav18-channel-pain-grasshopper-mouse

proteinspotlight

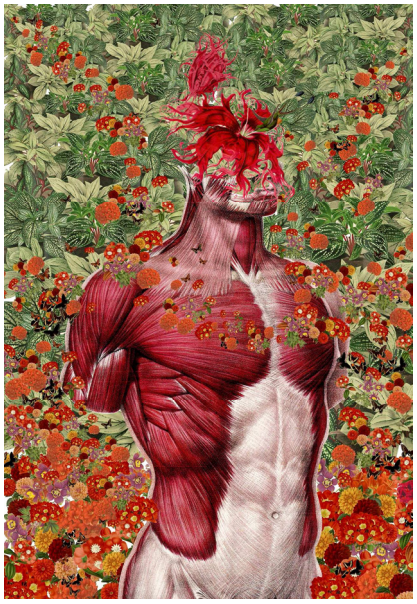
Swiss Institute of Bioinformatics

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/

the senses confused

Vivienne Baillie Gerritsen

Life is made of smells. Because smells play an important part in an organism's ordinary day to day life – it's a question of survival. And who says survival, says reproduction and food. Flowers exude perfumes to attract pollinators. There is evidence that spermatozoa sniff their way to eggs. Animals avoid eating what smells bad, but will be seduced by what smells good. While others let off putrid scents to ward off predators or, on the contrary, discharge encouraging ones to lure their prey. On the whole, the process is simple. If a fragrance is pleasant, an organism will be attracted by it. If it is not, it will turn away. This relatively direct means of communication between organisms is carried out by a more or less elaborate olfactory system. Recently, scientists managed to modify an odorant receptor – known as Orco – of *Aedes aegypti*, the mosquito responsible for transmitting yellow fever and dengue fever to humans. In so doing, the mosquito seemed to lose its taste for human skin – a valuable fact which could be used to develop powerful insect repellents.



Wellness, by Travis Bedel

Courtesy of the artist

Mosquitoes use their sense of smell to track down specific mammals, guided by exhaled CO_2, body temperature and the specific scent of chemicals produced by mammalian skin. It is now a well-known fact that only female mosquitoes bite. This is because they need vertebrate blood for their eggs to complete the reproductive cycle. Why Nature happened upon such a process remains a mystery – especially since there seems to be little evidence of mammals benefiting in any way from this kind of attention. Two mosquitoes show a marked preference for humans, namely *Aedes aegypti* and *Anopheles gambiae*. And, in so doing, they transmit diseases they carry, such as dengue fever, yellow fever and malaria. *A.gambia*, for instance, infects hundreds of millions of people with malaria every year, giving rise to large-scale health and economic issues.

An animal's olfactory system depends on odorant receptors – ORs – that are expressed on olfactory neurons. Unexpectedly, the number of ORs in animals varies hugely. As an example, fruit flies carry about 60 different ORs, while mosquitoes only carry about 27 – none of which have corresponding genes in the fruit fly! Though it may be expected that all animals depend on an olfactory system that works on the same basis, it is not the case. Indeed, it was initially thought that, upon ligand binding, as in the mammalian system, all ORs underwent conformational changes which ultimately relayed the signal down the olfactory neuron. However, scientists discovered that, in *Drosophila* for instance, the system was very different and signal transduction depended on odorant-gated cation channels. This also seems to be the case for the mosquito's olfactory system.

The system works thanks to the combined efforts of two entities: a specific OR and Odorant Receptor Coreceptor, also known as Orco. Orco always accompanies an OR, and is expressed in all olfactory neurons. Hence, a typical insect olfactory system depends on a heteromeric odour-gated ion channel, made up of an OR which binds a specific odour ligand and the coreceptor Orco. Is Orco really all that important? Well, when scientists mutated it in *A.aegypti*, they found that the engineered mosquitoes, though still eager to feed on mammals, ignored humans. This suggested that Orco is essential for recognising a particular scent given off by human skin. Such a finding is of great interest since it makes the coreceptor an ideal target for developing insect repellents.

From a molecular point of view, what is happening? Orco is a classical transmembrane G-protein coupled receptor which is activated the moment an odour molecule binds to the OR-Orco heteromer. Orco was engineered by unleashing zinc-finger nucleases onto it, which proceeded to make targeted mutations. The mutated Orco showed a clear reduced preference for humans – yet were still attracted to other vertebrates but seemed to be incapable of discriminating between them. The obvious conclusion is that, for mosquitoes, Orco is necessary for identifying human skin. However, how Orco interacts with ORs and relays the olfactory signal is unknown. And this is what needs to be understood to develop the ideal mosquito repellent to fight against the infectious diseases they transmit.

So Orco seems to be the part of a mosquito's olfactory system that gives it the ability to identify human skin. Or, a provocative thought, is it Orco that prevents mosquitoes from being attracted to any other vertebrate…? The question remains open, but doesn't change the end result: in mosquitoes, the OR-Orco pathway provides information on the specific identity of a host and, in *A.aegypti* and *A.gambiae*, the heteromer mediates a clear preference for humans, and the concomitant unfortunate transmission of infectious diseases.

Pest insects, such as mosquitoes but also other blood-sucking insects like ticks, can be disastrous for agriculture and human health, besides having an effect on a nation's economy. If the olfactory systems of these insects are understood on a very fine level, repellents can be designed to counter the damage they cause when feeding on their preferred animal – or indeed plant – host. It is obviously not possible to engineer the Orco of each mosquito… but studying Orco in its mutated form will help to understand the molecular mechanisms underlying a mosquito's preference for human skin – following which, an Orco-specific insect repellent could be designed and, perhaps, the lives of millions of people could be saved.

Cross-references to UniProt

Odorant receptor coreceptor, *Aedes aegypti* (Yellowfever mosquito) : Q178U6
Odorant receptor coreceptor, *Anopheles gambiae* (African malaria mosquito) : Q7QCC7

References

1. DeGennaro M., McBride C.S., Seeholzer L., Nakagawa T., Dennis E.J., Goldman C., Jasinskiene N., James A.A., Vosshall L.B.
Orco mutant mosquitoes lose strong preference for humans and are not repelled by volatile DEET
Nature 498:487-491(2013)
PMID: 23719379

2. Ha T.S., Smith D.P.
Insect odorant receptors: Channelling scent
Cell 133:761-762(2008)
PMID: 18510917

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/

SIB Swiss Institute of Bioinformatics

two's company

Vivienne Baillie Gerritsen

Pairing up is sometimes paramount to life. On the molecular scale, dimerization in our bodies is at the heart of many fundamental biological processes, such as the transduction of signals from the outside of a cell to the inside for instance. Split two molecules apart and, just like taking the propeller away from a ship, things are sure to change drastically. Signal transduction, on which life depends, is hugely due to protein-protein interaction. A ligand recognises its receptor, binds to it, thereby triggering off biological processes downstream. In the case of Kit ligand, and its receptor Kit, their binding is subject to the dimerization of both the ligand and its receptor, following which signals are transduced further downstream triggering off other biological processes. Kit ligand and Kit are a case of substantial conformational change on the molecular level – dimerization but also angles which bring about flexibility – that are necessary for Kit to get on with its job.

The Beautiful Spotted Negro Boy, ca. 1810-1811

Artist unknown

Kit ligand is a cytokine. When it binds to its membrane-anchored receptor – Kit – it activates a tyrosine kinase domain which ultimately initiates a horde of cellular responses downstream. This only happens once Kit has dimerized – a structural change which is driven by Kit ligand whose sole function seems to be to bring together the two Kit monomers. When this happens, Kit undergoes multiple autophosphorylations which stimulate downstream signalling pathways involved in hematopoiesis as well as the survival, proliferation and differentiation of mast cells, melanocytes and germ cells.

Kit ligand is both soluble and membrane-bound – though, under physiological conditions and extracellular, it seems to exist, mainly, as a monomer. Each monomer is a short chain of four helix bundles, characteristic of a helical cytokine topology – though Kit ligand structure is quite unique, and it is thought that its dimerization probably plays a regulatory role in its binding affinity to Kit receptor, as well as in its own activation.

The two Kit ligand monomers dimerize head to head, thus forming a longer chain, which is bent in its middle - possibly giving Kit ligand the power it needs to bind to Kit and then bring about the receptor's dimerization. Kit ligand then binds to Kit – where one end slips into one Kit monomer, and the other slips into the other Kit monomer – a little like two hands grabbing each Kit monomer to bring them closer to one another. In its middle, the Kit ligand dimer is connected by a flexible joint. Kit ligand is thus perpendicular to the Kit monomers. The overall structure sports a large cavity in its centre. And the whole complex is stabilized by a host of additional interactions, also mediated by the conformational changes of Kit ligand and its receptor.

Amongst other pathways, the Kit ligand-Kit complex mediates a pathway involved in

pigmentation, i.e. melanocyte differentiation and proliferation. Mutations in Kit receptor, for example, are known to be responsible for the congenital pigmentation disorder, piebaldism, characterised by depigmented patches on the body. In the early 1900s, a black African boy became a famous case of piebaldism when he was taken from his parents in the Caribbean, at the early age of 15 months, and sold in England to John Richardson, a travelling showman. The small boy was known as the "Beautiful Spotted Negro Boy" and became the star attraction. He died three years later – no doubt due, in part, to the number of hours he was exhibited on a daily basis.

Kit ligand also seems to be the cause of another kind of pigmentation, of a very different nature: blond hair. There is a variant which is found almost exclusively – and in high frequency – in populations of European ancestry and seems to be responsible for fair hair. Characteristically, the further humans live from the equator, the lighter their eye, hair and skin pigmentation is. Pigmentation depends on the type of melanin synthesized, and the size, shape and quantity of the cells they belong to – the melanosomes. The lighter skin of Europeans is probably to facilitate the synthesis of vitamin D3 at a latitude where levels of UVR are low. However, the physiological benefits of lighter eyes and hair remain a mystery.

Understanding, in its minute detail, what goes on between Kit ligand and its receptor will help to clarify the mechanisms of receptor activation as a whole. And the more the structure of such a complex is known, the more it can be used to design novel therapeutic interventions for the treatment of cancers and other diseases driven by activated receptors. Kit ligand, for instance, is known to promote hematopoietic recovery; in combination with other cytokines, it is used to reduce the haematological damage of chemotherapy. The design of Kit ligand analogues could be more potent, for example, and also be used to treat anaemia. As for the pigmentation properties of Kit ligand, a greater insight into its workings would not only help scientist track down past human migration from Africa to the rest of the world, but could also be a promising candidate for forensic geneticists, as well as in the study of eye and skin diseases that are known to be related with such traits.

Cross-references to UniProt

Stem cell growth factor Kit ligand, *Homo sapiens* (Human): P21583
Stem cell growth factor receptor Kit, *Homo sapiens* (Human): P10721

References

1. Yuzawa S., Opatowsky Y., Zhang Z., Mandiyan V., Lax I., Schlessinger J.
Structural basis for activation of the receptor tyrosine kinase KIT by stem cell factor
Cell 130:323-334(2007)
PMID: 17662946

2. Sulem P., Gudbjartsson D.F., Stacey S.N., Helgason A., Rafnar T., Magnusson K.P., Manolescu A., Karason A., Palsson A., Thorleifsson G., Jakobsdottir M., Steinberg S., Pálsson S., Jonasson F., Sigurgeirsson B., Thorisdottir K., Ragnarsson R., Benediktsdottir K.R., Aben K.K. Kiemeney L.A., Olafsson J.H., Gulcher J., Kong A., Thorsteinsdottir U., Stefansson K.
Genetic determinants of hair, eye and skin pigmentation in Europeans
Nature Genetics 39:1443-1452(2007)
PMID: 17952075

3. Zhang Z., Zhang R., Joachimiak A., Schlessinger J., Kong X.-P.
Crystal structure of human stem cell factor: Implication for stem cell factor receptor dimerization and activation
PNAS 97:7732-7737(2000)
PMID: 10884405

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/

SIB Swiss Institute of Bioinformatics

the hidden things

Vivienne Baillie Gerritsen

Nature has its secret ways. During the course of the 19th century, the Augustinian friar Gregor Mendel worked out the basics of genetic inheritance as he crossbred pea plants. About a century later, it has become obvious that the inheritance of a given trait is in fact not so straightforward. What is more, there seems to be growing evidence that a given trait can actually be handed down generations – even skipping generations – without it being frankly dictated by a gene; a notion which, in the realm of biological dogmas, is like a crack at the base of a sturdy building. The concept is not really new but scientists may have strengthened it following studies on folate metabolism – one of whose major protagonists is methionine synthase reductase, or Mtrr – by suggesting a mechanism of inheritance that is driven by entities which are not an actual part of a gene, otherwise known as epigenetic inheritance.

The Family, by Samuel Back

wikipedia

What is folate? Folate, or folic acid, is also known as pteroly-L-glutamate and more commonly as vitamin B9. 'Folate' or 'folic' is derived from the Latin 'folium' meaning 'leaf', since folates occur naturally and especially in plants that have dark green leaves; the compound received its name in 1941 when it was isolated from spinach. Over the years, it has become apparent that folic acid plays an essential role in nucleic acid synthesis and methionine regeneration – which both play an important part in genetic inheritance. And diets that are poor in folic acid – one of the most common deficiencies in humans – can be at the heart of serious ailments amongst which macrocytic anaemia, intrauterine growth restriction, placental abnormalities, neural tube defects, and perhaps even psychiatric and cognitive disorders.

The first person to discover the link between health and folic acid was the English haematologist Lucy Wills (1888-1964) who travelled to Bombay in 1928 to investigate anaemia in pregnancy, which was prevalent in female textile workers. It became apparent that the poorer populations had diets that were deficient in protein, fruit and vegetables. Wills solved the problem by adding yeast to their diet. It was only years later – in 1941, precisely – that scientists managed to single out the chemical component in the yeast that was lacking in their diet: folic acid. During the 1950s and 1960s, folate metabolism was finally elucidated, and one of its reactions was later shown to be particularly important in DNA synthesis. And who says DNA synthesis says genes, and hence development and inheritance.

One of the major enzymes to be part of the folic acid cycle is methionine synthase reductase, or Mtrr, which is the actual link between folate metabolism and methionine cycles, in other words: development. Indeed, folic acid carries methyl groups that are relayed downstream and ultimately used for DNA methylation. Mtrr uses the methyl groups from the folic acid cycle to form methionine which, in turn, is a precursor of S-adenosylmethionine, or SAM, the methyl donor for many cellular substrates including

proteins, RNA and DNA. Mtrr is therefore essential for the normal progression of the folate and methionine cycles, without which the expression of genes would be seriously hindered.

DNA methylation has an important role in the expression of genes since it operates as a sort of on/off switch. When such a system is impeded, it is not difficult to understand that a wide spectrum of phenotypes can suffer. A diet that is deficient in folic acid can cause this but the same sort of effect can occur if Mtrr is mutated. Indeed when scientists knocked out Mtrr in mice, it caused extensive DNA demethylation and abnormalities such as developmental drawbacks, neural tube defects, placental defects and placental lethality. This demonstrated that DNA methylation per se is an important epigenetic determinant in gene expression and DNA stability.

The intriguing part was that Mtrr deficiency could lead to the appearance of the same congenital malformations in wild-type progeny up to five generations down the line. How? What was it that was maintaining information that was passed on from parents to children in a non-Mendelian manner? One plausible explanation is that DNA methylation is inherited; in other words, the on/off switch is passed down to offspring. Up to now, it had been thought that these switches were wiped away "after use" so to speak, and hence never transmitted to future generations. But it now seems that, much like when you forget to wipe away all the chalk marks on a blackboard, some of the switches remain and are relayed to progeny through the germline. Furthermore, it seems that once an epigenetic defect is generated, it may never completely revert back to the way it was in its ancestors – which has important evolutionary repercussions.

So, what this demonstrates is that a given trait which is transmitted from generation to generation may not only be genetic, but also epigenetic which adds a spicy ingredient to the world of inheritance. And that, in the case of folic acid, the ingredient can actually be caused by a diet deficiency, i.e. something purely environmental, or even societal. It is believed that disorders such as depression or obesity can be passed on in the same way. There is still a long way to go to understand how such a mode of inheritance works and the weight it may have in passing on certain unfortunate traits. However, studying Mtrr to gain a better understanding of epigenetic transmission should prove to be informative on the inheritance of developmental disorders between generations. And, in the long run, perhaps in the treatment of congenital anomalies in humans.

Cross-references to UniProt

Methionine synthase reductase, *Homo sapiens* (Human) : Q9UBK8
Methionine synthase reductase, *Mus musculus* (Mouse) : Q8C1A3

References

1. Padmanabhan N., Jia D. Geary-Joo C., Wu X., Ferguson-Smith A.C., Fung E., Bieda M.C., Snyder F.F., Gravel R.A., Cross J.C., Watson E.D.
Mutation in folate metabolism causes epigenetic instability and transgenerational effects on development
Cell 155:81-93(2013)
PMID: 24074862

2. Greer E.L., Shi Y.
What's the Mtrr with your grandparents?
Cell Metabolism 18:457-459(2013)
PMID: 24093670

3. Nazki F.H., Sameer A.S., Ganaie B.A.
Folate: Metabolism, genes, polymorphisms and the associated diseases
Gene 533:11-20(2014)
PMID: 24091066

proteinspotlight

Swiss Institute of Bioinformatics

Protein Spotlight (ISSN 1424-4721) is a monthly 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/

National Nodes

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

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

Finland
CSC, Espoo

France
ReNaBi, French bioinformatics platforms network

Greece
Biomedical Research Foundation of the Academy of Athens, Athens

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

Norway
The Norwegian EMBnet Node, The Biotechnology Centre of Oslo

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

Russia
Biocomputing Group, Belozersky Institute, Moscow

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

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

Sweden
Uppsala Biomedical Centre, Computing Department, Uppsala

Switzerland
Swiss Institute of Bioinformatics, Lausanne

United Kingdom
The Genome Analysis Centre (TGAC), Norwich

Specialist- and Assoc. Nodes

ILRI/BECA
International Livestock Research Institute, Nairobi, Kenya

KEMRI
Wellcome Trust Research Programme, Kilifi, Kenya

UMBER
Faculty of Life Sciences, The University of Manchester, UK

CPGR
Centre for Proteomic and Genomic Research, Cape Town, South Africa

for more information visit our Web site
www.EMBnet.org

EMBnet.journal
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 "Instructions for authors" at http://journal.EMBnet.org/index.php/EMBnetnews/about and send your manuscript and supplementary files using our on-line submission system at http://journal.EMBnet.org/index.php/EMBnetnews/about/submissions#onlineSubmissions.

Past issues are available as PDF files from the Web site:
http://journal.EMBnet.org/index.php/EMBnetnews/issue/archive

Publisher:
EMBnet Stichting p/a
CMBI Radboud University
Nijmegen Medical Centre
6581 GB Nijmegen
The Netherlands

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