EMBnet.journal
Bioinformatics in Action

22
2017

EMBnet.journal
Bioinformatics in Action

Contents

Cover picture: Burrowing owls. Athene cunicularia, Brasilia, Brazil, 2016, © Erik Bongcam-Rudloff

Editorial

EMBnet.journal is the official publication of EMBnet, a not-for-profit foundation and legal entity based in Nijmegen, the Netherlands. EMBnet.journal is peer reviewed and has a strong focus on articles that relate to the practical use of bioinformatics in solving scientific problems in the life sciences; it also keenly promotes articles relating to bioinformatics education and training, and hence accepts papers describing the production of teaching materials, data-sets, tutorials, etc., development of competency frameworks, new pedagogical approaches, technology-enhanced-learning systems, and so on.

The Journal also reports EMBnet's activities, achievements and plans for the future. This volume publishes one interesting technical note on how to set up your own Chipster bioinformatics server in the EGI Federated Cloud, and also publishes one report from the Executive Board activities 2015-2016 and a report by EMBnet.journal´s Editorial Assistant. Two interesting articles dedicated to Education are also included in this volume, one about an innovative method on how to present bioinformatics for the layman (the Metagenomic Pizza) and one sharing the experiences in setting up an intercontinental course about "Assembly, annotation and comparison of bacterial genomes".

This closing editorial of volume 22 was the first volume with a fully revised, enhanced layout. The editorial and technical teams behind the Journal worked hard both on the layout design and on establishing new administrative routines, and now is prepared to receive the contributions from more authors and readers to the ameliorated journal.

We hope that readers will appreciate the journal improvements, and the editorial team warmly encourages the submission of new articles to volume 23 using our online publishing system.

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

EMBnet.journal
Bioinformatics in Action

Blatter et al. (2016) EMBnet.journal 22, e864
http://dx.doi.org/10.14806/ej.22.0.864

Reports

The Metagenomic Pizza: a simple recipe to introduce bioinformatics to the layman

Marie-Claude Blatter[1], Vivienne Baillie Gerritsen[1], Patricia M. Palagi[1], Lydie Bougueleret[1], Ioannis Xenarios[1,2]

[1] SIB Swiss Institute of Bioinformatics, Geneva, Switzerland; [2] Geneva University, Geneva, Switzerland.

Introduction

Bioinformatics touches many aspects of everyday life – health, nutrition, environmental care, forensics – and is a major element of modern research in the life sciences. However, it is a fairly young scientific domain and is still poorly known to the layman.

In January 2013, a widespread food contamination scandal arose regarding beef *lasagne* that contained horsemeat. This led us to imagine a workshop to explain, in a simple but engaging way, how to identify food ingredients by way of DNA and bioinformatics tools available on the Internet. 'The Metagenomic Pizza' was born.

This article describes the Metagenomic Pizza, one of several guided bioinformatics activities that are available on the 'Ateliers de Bioinformatique' website[1] (mainly in French).

The Metagenomic Pizza Workshop: step by step

1. Biodiversity in a pizza

To get people started, participants are asked to imagine different pizza recipes and ingredients (i.e., pizza crust, mozzarella, nutella, ham, etc.), and consider the various animal and vegetable species that can be found in a pizza (i.e., wheat, buffalo, palm, cacao, hazelnuts, pig). We end up with a list of 50 different organisms, including Homo Sapiens (one of the cook's hairs), horse, and other, invisible, organisms, such as yeast, bacteria and archaea – all of which can be found on a pizza, whether they should be there or not. The question of how it is possible to identify all these species in our food (or in other samples) is raised.

2. DNA: a little theory

A brief introduction describes how food comes from living organisms, and thus contains DNA. The DNA present in an uncooked pizza can be extracted and sequenced, thanks to Next Generation Sequencing (NGS) technologies. The different ways of representing DNA – from the well-known double helix explained in the biology classroom, to a DNA sequence and its digital single-strand format – is a key step for understanding the workshop. We then introduce the concept that certain regions in DNA sequences are specific to a given organism, much as a barcode is specific to a given item in a shop. Consequently, such a region can be used as a means of identification by comparison with already known data – which is where bioinformatics comes in.

3. DNA sequence analysis using the 'BLAST approach': pen-and-pencil activity

Participants are asked to fish out several 40-nucleotide DNA-reads from a box that contains hundreds (a foretaste of what is called 'big data' in metagenomics). They then manually compare these stretches of DNA with 50 annotated DNA "entries" from a "printed knowledgebase". This is presented in the form of a booklet in which are found, page after page in alphabetical order, the 50 DNA sequences. Each page also includes information on the organism it belongs to, as well as the function of the protein it encodes (Figure 1). This step involves comparing, two-by-two, stretches of DNA about 40 nucleotides long, and looking for tiny differences ("one small difference makes all the difference"); it is a painstaking task, and the participants rapidly ask for help from the computer.

The DNA read (`ttcaaaactaacaatgttaccgccaggctttttgagcgcca`) is from a tomato (*Solanum lycopersicum*), and encodes a protein involved in fruit pigmentation. A link to the accession number of the corresponding entry in UniProtKB/Swiss-Prot is provided (Q9ZNU6), as well as the name of the taxonomic group to which the organism belongs (*Dicotyledon Plant*).

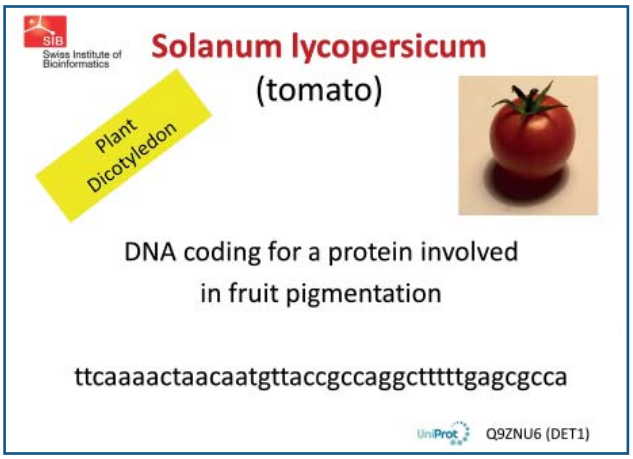


Figure 1. Example of an 'entry' out of the 50 found in the 'printed DNA knowledgebase'.

Article history
Received: 29 January 2016
Published: 24 March 2016

[1] http://education.expasy.org/bioinformatique/

4. DNA sequence analysis using BLAST: bioinformatics activity

The manually obtained results are checked on the computer. Participants perform a BLAST search (BLASTX) against the UniProtKB/Swiss-Prot knowledgebase at www.uniprot.org/blast. Little knowledge is required for someone to understand a BLAST output – the results are intuitive. Even school children easily grasp the fact that the computer is comparing their DNA sequence with all the sequences stored in an electronic knowledgebase. They are informed that the knowledgebase has 550,000 sequences, representing 12,000 species, and that the sequence most similar to theirs will be the first to appear in the list of results. To give the exercise a flavour of reality, some of our 40-nucleotide DNA-reads are not discriminative enough when using BLASTX against UniProtKB/Swiss-Prot. This provides an opportunity to discuss the fact that scientists sometimes need to use longer or other stretches of DNA for the unambiguous identification of (food) contaminants.

5. Universality of DNA and species classification

Once the taxonomic origin of the pizza DNA sequences has been identified, the participants map the DNA sequences onto a printed reference species tree. This is a nice way of illustrating the biodiversity found within a pizza, and the fact that DNA is found in every living organism (Figure 2).

Figure 2. A participant mapping the DNA sequences onto a printed reference species tree.

6. The gender of the cook

Using yet another bioinformatics tool – *i.e.*, the BLAST-like alignment tool (Blat) of the UCSC Genome Browser[2] – the participants are asked to check whether a DNA sequence discovered in the cook's hair bulb (`atgcaatcatatgcttctgctatgttaagcgtattc`) is located on chromosome Y or not. If it is, then the cook is a man.

From this point onwards, participants are left to their own devices, to extend their knowledge by using the skills they have acquired during the workshop. They can, for instance, type out a random DNA sequence of about 40 nucleotides and check whether this sequence actually exists or not, and whether it belongs to the human genome, or to the genome of another species. They can also try to see whether the previously identified tomato DNA sequence exists somewhere in the human genome.

Discussion

Since 2013, the workshop has been successfully offered to more than 2,000 people, from the age of nine years upwards, and from different backgrounds. These events took place in classrooms, during science fairs, university open houses, bioinformatics labs at the SIB Swiss Institute of Bioinformatics or during high-school teacher training courses. It is also one of the workshops given by (R)amène ta Science[3], a concept developed by Geneva University. This involves academic experts who train students – future 'ambassadors' – how to conduct the workshop. In turn, these ambassadors conduct the workshop at their own school.

What are the advantages of such a workshop? It is highly adaptable in time (20 to 90 minutes), content and level of difficulty, and is thus convenient for all kinds of participant. Our main objective is to engage the layman in activities that are similar to authentic scientific research practice, and not to get lost in the technical know-how (Landhuis, 2015; Form and Lewitter, 2011). This way, the participants manipulate 'real' DNA sequences, either manually or with the help of bioinformatics tools used on a daily basis by scientists. It is an ideal way for them to understand the key role played by bioinformatics in the life sciences today. A few applications of current research in metagenomics are also discussed, such as the study of DNA preserved in the teeth of 1,000-year old skeletons. Participants learn how such studies are capable of identifying bacteria and food remains (Warinner *et al.*, 2014). They also discover how DNA derived from microbes extracted from agricultural soil, ocean surface water, or deep-sea whale bone (von Mering *et al.*, 2007), for example, helps to define new species or biological functions.

Acknowledgements

We thank Sandrine Pilbout for the original idea of the 'taxonomic' pizza recipe[4], and all the workshop participants, whether volunteers or not.

References

1. Form D and Lewitter F (2011) Ten simple rules for teaching Bioinformatics at the High School level *PLOS Computational Biology* **7** (10):e1002243. http://dx.doi.org/10.1371/journal.pcbi.1002243

[2] http://genome.ucsc.edu/cgi-bin/hgBlat
[3] http://ramene-ta-science.unige.ch/
[4] http://www.uniprot.org/help/2006/08/22/release

2. Landhuis E (2015) Early BLAST OFF: bringing bioinformatics to secondary schools. Biomedical Computation Review: http://biomedicalcomputationreview.org/content/early-blast

3. von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N, Bork P (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315**, 1126-1130. http://dx.doi.org/10.1126/science.1133420

4. Warinner C, Rodrigues JF, Vyas R, Trachsel C, Shved N, Grossmann J, Radini A, Hancock Y, Tito RY, Fiddyment S, Speller C, Hendy J, Charlton S, Luder HU, Salazar-Garcia DC, Eppler E, Seiler R, Hansen LH, Castruita JA, Barkow-Oesterreicher S, Teoh KY, Kelstrup CD, Olsen JV, Nanni P, Kawai T, Willerslev E, von Mering C, Lewis CM Jr., Collins MJ, Gilbert MT, Ruhli F, Cappellini E (2014) Pathogens and host immunity in the ancient human oral cavity. *Nat. Genet.* **46**, 336-344. http://dx.doi.org/10.1038/ng.2906

EMBnet.journal
Bioinformatics in Action

Falquet et al. (2016) EMBnet.journal 22, e872
http://dx.doi.org/10.14806/ej.22.0.872

Reports

Report on the Swiss-Colombian workshop: "*Assembly, annotation and comparison of bacterial genomes*"

Laurent Falquet[1], Sandra Patricia Calderon-Copete[2], Emiliano Barreto-Hernández[3], Jaime Enrique Moreno Castañeda[4]

[1]University of Fribourg and Swiss Institute of Bioinformatics, Biochemistry Unit, Fribourg, Switzerland; [2]LGTF-UniL, Génopode, Lausanne, Switzerland; [3]Centro de Bioinformática, Instituto de Biotecnología - Universidad Nacional de Colombia, Bogotá, Colombia; [4]Grupo de Microbiología-Investigación Instituto Nacional De Salud, Bogotá, Colombia.

Introduction

This workshop was organised as part of the Swiss-Colombian project, *A pilot integrative knowledgebase for the characterization and tracking of multi resistant Acinetobacter baumannii in Colombian Hospitals*, sponsored by the Leading House Cooperation and Development Centre[1] of the École Polytechnique Fédérale Lausanne (EPFL).

The aim of this project is to develop a prototype centralised knowledgebase. Initially, we selected complete genome sequences obtained from a collection of *Acinetobacter baumannii* strains collected from the Antimicrobial-Resistant Healthcare-Associated Infections Surveillance Program, during 2012–2015, by the Colombian National Health Institute (NHI) and the Biotechnology Institute of the National University of Colombia (IBUN-UNAL). In addition, complete *Acinetobacter baumannii* genome sequences were added from public databases. The prototype will consist of fully assembled and annotated genomes associated with geographical, temporal and clinical data, allowing tracking of a variety of infection outbreaks. The resulting knowledgebase will serve as a reference to help clinicians to track rapid dissemination of highly pathogenic and resistant strains.

The workshop was held at the Bioinformatics centre of the National University of Bogotá, 23-27 May 2016, and gathered 18 participants from diverse institutions in Colombia.

Programme

Each half-day was split into theoretical lectures (60-90 minutes each), followed by hands-on practicals (150 minutes each). The programme is shown in Table 1.

Table 1. Programme of the Swiss-Colombian workshop, 23-27 May 2016.

Day 1	Introduction to UNIX and computer clusters Introduction to sequencing techniques, QC and data cleaning (adapter removal, trimming, filtering, etc.)
Day 2	De novo assembly Assembly by re-mapping
Day 3	SNP and small indel calling: how to detect variants? Annotation and profiling of resistance and virulence factors
Day 4	Comparative genomics (core/pan genomes, structural variants, phylogeny distribution)
Day 5	Presentation of individual research projects of participants

Organisation of the work

Given the lack of access to a high-performance cluster, the participants were divided into nine groups of two, each being responsible for the analysis of a set of paired-end 100 bp reads from Illumina sequencing of a strain of *Acinetobacter baumannii* from the Sequence Read Archive (SRA). To distribute the workload, we divided the work across five computing nodes (16 cores, 64 Gb RAM). After a brief reminder of computing and UNIX operating system basics, participants had the opportunity to refresh their knowledge of the command line. The genome analysis comprised data quality control with FastQC[2], cleaning both the adapter content with CutAdapt (Martin, 2011) and low quality sequences with sickle[3]. The cleaned sequences were assembled with SOAPdenovo (Luo *et al.*, 2012), using various kmers, and SPAdes (Bankevich *et al.*, 2012). The draft genomes were compared using summary statistics, QUAST (Gurevich *et al.*, 2013) and MAUVE (Darling *et al.*, 2010). The best draft genome was annotated using Prokka (Seemann, 2014) and a set of HMMs built from the Virulence Factor database (Chen *et al.*, 2016) and downloaded from the ResFam database (Gibson *et al.*, 2015). The gff files of ten genomes (nine, plus reference) were compared, looking for core and pan genomes using Roary (Page *et al.*, 2015) and Phandango[4]. The reads were also re-mapped to the reference genome, and SNPs and indels called with BWA (Li and Durbin, 2010), SAMtools and BCFtools (Li, 2011). The SNP vcf files were annotated with snpEff, filtered with SnpSift (Cingolani *et al.*, 2012) and finally visualised with IGV (Thorvaldsdottir *et al.*, 2013). After conversion to multi-fasta format, a tree was constructed

Article history
Received: 22 August 2016
Published: 9 November 2016

[1] http://cooperation.epfl.ch/cms/lang/fr/pid/118892

[2] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[3] https://github.com/najoshi/sickle

[4] http://jameshadfield.github.io/phandango/

with FastTree (Price *et al.*, 2010) and visualised using the Newick viewer (Boc *et al.*, 2012).

Finally, participants had the opportunity to present their own current research work and to receive feedback from other course participants and trainers, promoting an enriching exchange of valuable research experiences in the area of genomics.

Figure 1. Participants and trainers.

Evaluation of the course

Participants from different research institutions expressed satisfaction with the high academic level of the course in general. They gave high value to the knowledge shown by trainers, and to the materials used in the lectures and practical exercises. Some respondents said that knowledge acquired during the course had allowed them to solve their own data-analysis problems.

They also made recommendations regarding the inclusion of additional practicals, and the possibility of additional access to the servers used for the hands-on sessions, in order to become more familiar with Linux. Course servers will be available to them for a few months more.

Figure 2. Workshop hand-on session.

Conclusions

According to the attendees' course evaluation and the organisers comments, this workshop was very useful both for biologists working on assembly and annotation of bacterial genomes, and researchers of the Colombian NHI, interested in tracking resistance and virulence factors in clinical isolates.

Acknowledgements

This work was supported by the "Leading House Cooperation and Development Centre". We thank Hermes Perez Cardona and Dra. Maria Teresa Reguero, for their help with local organisation of the workshop.

References

1. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477. http://dx.doi.org/10.1007/978-3-642-37195-0_13
2. Boc A, Diallo AB, Makarenkov V (2012) T-REX: a Web server for inferring validating and visualizing phylogenetic trees and networks. *Nucleic Acids Res.* **40**, W573–W579. http://dx.doi.org/10.1093/nar/gks485
3. Chen L, Zheng D, Liu B, Yang J and Jin Q (2016) VFDB 2016: hierarchical and refined dataset for big data analysis - 10 years on. *Nucleic Acids Res* **44**, D694–D697. http://dx.doi.org/10.1093/nar/gkv1239
4. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118 ; iso-2; iso-3. *Fly* (Austin) **6**, 80–92. http://dx.doi.org/10.4161/fly.19695
5. Darling AE, Mau B and Perna NT (2010) progressiveMauve: Multiple Genome Alignment with Gene Gain Loss and Rearrangement. *PLoS ONE* **5**, e11147. http://dx.doi.org/10.1371/journal.pone.0011147
6. Gibson MK, Forsberg KJ, Dantas G (2015) Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J* **9**, 207–216. 6. https://dx.doi.org/10.1038/ismej.2014.106
7. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075. http://dx.doi.org/10.1093/bioinformatics/btt086
8. Li H (2011) A statistical framework for SNP calling mutation discovery association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993. http://dx.doi.org/10.1093/bioinformatics/btr509
9. Li H and Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595. http://dx.doi.org/10.1093/bioinformatics/btp698
10. Luo R, Liu B, Xie Y, Li Z, Huang W *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 18. http://dx.doi.org/10.1186/2047-217x-1-18
11. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10-12. http://dx.doi.org/10.14806/ej.17.1.200
12. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S *et al.* (2015) Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693. http://dx.doi.org/10.1093/bioinformatics/btv421
13. Price MN, Dehal PS and Arkin AP (2010) FastTree 2 - Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **5**, e9490. http://dx.doi.org/10.1371/journal.pone.0009490
14. Seemann T (2014) Prokka: rapid prokaryotic genome annotation. *Bioinforma. Oxf. Engl.* **30**, 2068–2069. http://dx.doi.org/10.1093/bioinformatics/btu153
15. Thorvaldsdottir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* **14**, 178–192. http://dx.doi.org/10.1093/bib/bbs017

EMBnet.journal
Bioinformatics in Action

Theofilatos et al. (2016) EMBnet.journal 22, e871
http://dx.doi.org/10.14806/ej.22.0.871

Reports

InSyBio BioNets, an efficient tool for network-based biomarker discovery

Konstantinos Theofilatos[1,✉], Christos Dimitrakopoulos[1,✉], Christos Alexakos[1], Aigli Korfiati[1], Spiros Likothanassis[1], Seferina Mavroudi[1]

[1]InSyBio Ltd, London, UK

Abstract

Biological networks have been widely used in systems biology in order to model the complex interactions of molecular players such as proteins, genes, mRNAs, non-coding RNAs and others. However, most of the current methods for biomarker discovery do not use biological networks, but just deploy simple statistical methods to identify differentially expressed genes and gene products. In the present paper, we present InSyBio BioNets, which is a cloud-based web platform offering a unique biomarker discovery pipeline, which combines differential expression analysis and a method for comparing biological networks to identify biomarkers efficiently. As a case study, InSyBio BioNets was applied to a Parkinson dataset of gene expression measurements and outperformed a standard statistical approach by recovering a more compact and informative set of biomarkers.

Introduction

The execution of complex biological processes requires the precise interaction and regulation of thousands of molecules. These interactions can be modeled as networks, which typically consider molecular components within a cell as nodes and their direct or indirect interactions as edges. Network representation enables data integration from a wide range of studies, including protein-protein interaction (PPI) and gene expression measurements, into a single framework. The analysis of these networks can aid in understanding the disease mechanisms, but it has not been successfully linked to clinical applications until now.

Biomarker discovery is a field currently dominated by statistical analysis on the actual expression values of genes or quantitative values of proteins to identify the cellular molecules, which differ significantly in experiments between biological or clinical conditions (*i.e.* disease vs control samples). The results of this approach present certain drawbacks including the high number of discovered biomarkers that require experimental validation as well as the high number of false positives. Moreover, standard statistical approaches are prone to identify biomarkers descriptive of the disease's outcome and not of its cause. For this reason, the current trend in biomarker discovery is to detect biomarkers by comparing biological networks. Biological network metrics are more stable to changes between biological conditions compared to absolute gene expression differences and can be associated with the causes of disease mechanisms.

InSyBio BioNets is a tool providing a unique biomarker discovery pipeline that overcomes the aforementioned constraints of existing biomarker discovery methods capitalizing on biological networks' comparison. In specific, InSyBio BioNets offers a novel systems medicine approach, which provides biomarker sets with increased predictive accuracy. The proposed pipeline is offered through a flexible semi-automated web-based analysis enabling the users to easily navigate through its different steps, while also being able to use their own algorithms and methods. This can be accomplished either by selecting among a variety of algorithms offered through InSyBio BioNets web interface or by downloading intermediate results, processing them locally and uploading results to continue the analysis through the web interface.

In addition to the biomarker discovery pipeline, InSyBio BioNets provides a set of tools for the construction, preprocessing, meta-analysis and visualisation of biological networks and it supports tools for parsing and creating gene expression files to enable the construction and analysis of gene co-expression networks. Moreover, users can fast analyse large biological networks and gene expression files using the tool's user-friendly job management mechanism. Regarding the uncovered biomarkers, users have access to informative biomarker reports, which provide information from publicly available databases and from InSyBio Interact tool's PPI repository. These reports also include information about the prior knowledge linking biomarkers to diseases.

Article history
Received: 25 November 2016
Published: 1 December 2016

Theofilatos et al. (2016) EMBnet.journal 22, e871
http://dx.doi.org/10.14806/ej.22.0.871

Methodologies

Gene expression data parsing, preprocessing and analysis

InSyBio BioNets offers a set of tools for handling, preprocessing and analysing gene expression data in order to construct gene co-expression networks. It supports the universally accepted format for gene expression data named SOFT (Simple Omnibus Format in Text) supported also by Gene Expression Omnibus. InSyBio BioNets SOFT parsing includes the following preprocessing steps a) logarithmic normalisation, b) missing values estimation and c) filtering based on average expression or expression variance. The different experimental states (conditions) defined in the SOFT file are automatically recognised and a gene expression tab delimited file is constructed for each state. Gene expression files can also be used to generate weighted gene co-expression networks. A weighted edge is added to the network if the metric among the expression profiles of the two nodes adjacent to this edge exceeds a predefined threshold (which is automatically derived for each node from the dataset). InSyBio BioNets also provides a network-clustering tool that supports state of the art clustering algorithms, a network analysis tool to compute network metrics and components as well as various Cytoscape-based network visualisation tools. Most algorithms included in the InSyBio BioNets tool were implemented using Python programming language and standard python libraries such as Biopython.

Biomarker discovery pipeline based on network comparison

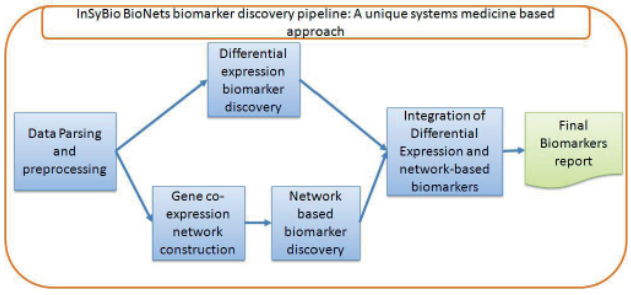


Figure 1. InSyBio BioNets biomarker discovery pipeline.

The proposed pipeline (Figure 1) is a five step procedure: I) parse and preprocess transcriptomics experiments, II) perform differential expression analysis to uncover differentially expressed biomarkers, III) construct gene-co expression networks for each one of the biological conditions examined in the transcriptomics experiments, IV) compare biological networks to uncover network-based biomarkers and V) combine differentially expressed and network-based biomarkers by a confidence score. For the networks comparison (step IV), users are able to select one of the offered network metrics (degree centrality, clustering coefficient or PageRank centrality with the PageRank centrality being the default option) and the tool detects the network's nodes for which the selected metric is significantly altered. Network-based biomarkers are merged with the differentially expressed ones (step V) by intersecting the two sets and by computing a combined confidence score. The final biomarker list is annotated with information from public available repositories including OMIM, DisGeNet, Genecards and other InSyBio Suite Tools (InSyBio Interact and InSyBio ncRNAseq). In step (II), differential expression is performed by using the Wilcoxon rank sum test and users can state a P-value threshold. Bonferroni corrections are applied to reduce the number of false positive predictions.

Results and conclusions

As a case study, InSyBio BioNets was used to uncover Parkinson Disease (PD) biomarkers from gene expression measurements in blood samples. PD remains a disease whose diagnosis is based on clinically detectable symptoms. When these symptoms arise, it is quite late for the effective patients' treatment. PD has been attributed to genetic and environmental causes in the relevant scientific literature. However, until now, there exist only a few early stage diagnostic tests of limited predictive accuracy. We analysed a Gene Expression Omnibus dataset (GDS2519) which has been constructed by microarray experiments on blood samples of 50 early stage Parkinson Disease patients, 22 control patients and 33 other neurodegenerative disease control patients. We used InSyBio BioNets to detect biomarkers for PD and we compared our results with the Wilcoxon rank sum test (used by Scherzer *et al.*, 2016) which detects biomarkers based on differential expression and ranks them based on their P-values. InSyBio BioNets uncovered a more specific set of biomarkers with increased predictive accuracy for PD-related genes (Table 1). These biomarkers were further reduced to five (HNRNPA3, ZFC3H1, SSR1, ATRX, SNCA) without the loss of classification accuracy when an ensemble genetic algorithm/SVM method was applied for feature selection.

Table 1. InSyBio BioNets vs. a standard differential expression analysis for identifying PD biomarkers from transcriptomics experiments.

Method	#PD Biomarkers	Precision using genes associated with PD from DisGeneNet DB	Classifiers Accuracy (SVM classifiers used) with 10-fold cross validation
Standard Approach (Wilcoxon Rank Sum method)	834	7.47%	96.4%
InSyBio BioNets	24	12.5%	100%

Availability

InSyBio BioNets is one of the tools included in the integrated bioinformatics web platform of InSyBio named InSyBio Suite. A Demo version of InSyBio BioNets is freely available at http://demo.insybio.com. A free evaluation version includes a one-month free license and it can be purchased by sending an email at info@insybio.com. To purchase the commercial version of InSyBio BioNets users can contact sales@insybio.com for the detailed quota and information. InSyBio is registered with the Information Commissioner's Office under registration reference number ZA182885 to provide data security.

Acknowledgements

InSyBio participates in the NBG Business Seeds program by the National Bank of Greece.

References

1. Theofilatos, KA, Likothanassis S, Mavroudi S (2015) Quo vadis computational analysis of PPI data or why the future isn't here yet. *Frontiers in genetics* **6**, 289. http://dx.doi.org/10.3389/fgene.2015.00289
2. Tong H, Faloutsos C, Pan JY (2006) Fast random walk with restart and its applications. *Sixth International Conference on Data Mining (ICDM'06)*, Hong Kong, pp. 613-622 https://doi.org/10.1109/ICDM.2006.70
3. Adler CH, Beach TG, Hentz JG, Shill HA, Caviness JN, Driver-Dunckley E, Dugger BN *et al.* (2014) Low clinical diagnostic accuracy of early vs advanced Parkinson disease Clinicopathologic study. *Neurology* **83**(5), 406-412. https://dx.doi.org/10.1212/WNL.0000000000000641
4. Scherzer CR, Eklund AC, Morse LJ, Liao Z *et al.* (2007) Molecular markers of early Parkinson's disease based on gene expression in blood. *Proc Natl Acad Sci USA*, **104**(3), 955-960. https://dx.doi.org/10.1073/pnas.0610204104

EMBnet.journal
Bioinformatics in Action

Santovito (2016) EMBnet.journal 22, e881
http://dx.doi.org/10.14806/ej.22.0.881

EMBnet AGM 2016: *EMBnet.journal* Editorial Assistant Activity Report

Antonio Santovito

Introduction

"Bioinformatics without borders". Those three words guided all my work during the first year of collaboration with EMBnet and EMBnet.journal (ENJ)[1]. How to help the journal reach a larger audience? How to improve EMBnet's communication and Public Relations (PR) activities? These have been the questions I've tried to answer, managing the tasks assigned to me and proposing new initiatives and useful tools. The first part of my work as Editorial Assistant (EA) was based on three key concepts: restyling, restoring and relaunching. During the first months, I created some multimedia support for incoming events, and tackled content recovery of ENJ, owing to a severe hack to the journal server.

Figure 2. The new ENJ printed layout.

Restyling

With the support of the ENJ Executive Editorial Board (EEB), Domenica D'Elia and I created a new version of the EMBnet leaflet and a short EMBnet presentation, using the restyled design agreed with the EEB, to be used for all EMBnet dissemination materials. Then we adapted this general presentation to be shown during the upcoming "NETTAB 2015 Workshop and the Integrative Bioinformatics Symposium", held as a joint event in Bari (IT) in October 2015. Afterwards, with the approval of EMBnet's Executive Board (EB), we started to create and collect comments for a new ENJ logo (shown in Figure 1). It was great teamwork, and everybody contributed with his/her suggestions for improvements. The result of our brainstorming was shared with all EMBnet members, receiving good feedback and approval.

EMBnet.journal
Bioinformatics in Action

Figure 1. The new ENJ logo.

Once the new logo was approved, the next step was restyling the ENJ website and of the layouts of the different types of journal article (Figure 2). Once again, the teamwork made the difference, and the new year started with a fresh and more appealing design.

Restoring

While we were working on the restyling, we still had a huge issue to solve: most ENJ articles were gone, following the hack, and it became necessary to consider how to make everything accessible and safe again. Lubos Klucar had a very important role in this phase. He taught me how to manage the Open Journal Systems[2], the Content Management System (CMS) of the ENJ website, and worked with me to recover all the missing contents and to restore them online. It was hard work, which we solved with the help of members of the EEB, who shared with us their personal article backups. When everything was back online, we applied the new website design and started working on new contents. As Lubos suggested, the Layout Editor role was assigned to me, owing to my work on the new layout restyle.

Relaunching

At the beginning of 2016, ENJ was back online, and everything was ready for the relaunch. We worked on a general call for papers, which was shared with the EEB for comments and suggestions, mainly relating to the type of target audience to whom to address the call.

Meanwhile, I started working with Axel Thieffry, Chair of the Publicity & Public Relations Special Interest Group (P&PR SIG), to manage EMBnet's LinkedIn group and website contents. A new social media dissemination strategy was also proposed to the EB, to improve the

Article history
Received: 27 November 2016
Published: 27 February 2017

[1] http://journal.embnet.org/
[2] https://pkp.sfu.ca/ojs/

Santovito (2016) EMBnet.journal 22, e881
http://dx.doi.org/10.14806/ej.22.0.881

visibility of ENJ articles on EMBnet's official social media accounts, sharing articles, relevant news about events and job opportunities every day. After six months, the results showed, among others, an increase of 70% Facebook fans and 150% Twitter followers, only using organic and non-sponsored posts.

Starting from May 2016, I also dealt with management of EMBnet's website, by editing and creating contents, as required by the Operational Board. In November, I volunteered to restyle EMBnet.org. The main goal of the project was to change the CMS, migrating everything from Drupal to Wordpress, a safer and more powerful CMS, with better customisation possibilities and an easier back-end system for creating, updating and managing content. The new website will be released in February 2017, once we have access to the new server. It will have a more modern design, in line with the new ENJ look, and will conform to the latest Web-design standards. It will also be responsive and accessible from mobile devices. It is a very complex project, because of the need to migrate the current contents to the new CMS, and for the complexity of the different custom-content taxonomies, as well as the e-commerce module for EMBnet membership. This is my personal gift to EMBnet, as a sign of gratitude for allowing me to have this magnificent job experience, which I hope will continue in the future.

Acknowledgements

This is where my report ends, hoping I've listed the most important points, being brief and, above all, not boring. This year has been very important for me, for the type of human and professional experiences I had. I met wonderful and willing people that helped me with their opinions, suggestions and tips. It has been an unforgettable experience and I thank everyone in EMBnet for the opportunity to work with such an important global scientific network. I hope my little contribution helps you to go forward to a successful 2017. It has been an honour to work with you.

EMBnet.journal
Bioinformatics in Action

D'Elia et al. (2017) EMBnet.journal 22, e887
http://dx.doi.org/10.14806/ej.22.0.887

Reports

2016 EMBnet Annual General Meeting – Executive Board Report

Domenica D'Elia[1], Emiliano Barreto Hernandez[2], Lubos Klucar[3], Erik Bongcam-Rudloff[4]

[1]CNR, Institute for Biomedical Technologies, Bari, Italy; [2]Centro de Bioinformática, Instituto de Biotecnología - Universidad Nacional de Colombia, Bogotá; [3]Institute of Molecular Biology, Slovak Academy of Sciences, Bratislava, Slovakia; [4]Swedish University of Agricultural Sciences, Uppsala, Sweden.

Abstract

During the past year, the Executive Board (EB) held regular monthly meetings either via Skype or using Adobe Connect. These meetings were carried out with the Interim Board (IB), comprising members of the EB and Teresa Attwood and Etienne de Villiers. The IB was established during the 2015 Annual General Meeting (AGM) both to support the new EB in its first steps forward (as three of its members were new), and to help oversee implementation and delivery of the investment strategy. The EB also regularly invited Special Interest Group (SIG) Chairs to participate at EB-IB meetings. Additional monthly meetings open to the full EMBnet constituency were also convened. In this report, we provide a brief overview of activities and achievements from June 2015 to October 2016.

Introduction

The first EB-IB meeting was held on 1 July. Erik Bongcam-Rudloff was elected Chair, Domenica D'Elia Secretary, Emiliano Barreto Treasurer. The fourth member, Lubos Klucar, retained his pivotal role in the maintenance of *EMBnet.journal* (ENJ).

The first actions of the EB were related to the investment strategy approved during the 2015 AGM in Oeiras (PT). In particular, priority was given to the following:

1. hiring an ENJ Editorial Assistant (EA) for one year to help the Executive Editorial Board (EEdB) in the daily management of editorial procedures, including contacts with authors and article copy editing;
2. assigning three fellowships for:
   - re-designing the ENJ website;
   - upgrading and updating of ENJ Open Journal System (OJS);
   - improving usability, style and content of EMBnet's website;
3. making a ring-fenced donation to GOBLET to hire an Educational Assistant for two years to develop jointly-branded materials. This proposal aimed to benefit EMBnet, by making its commitment to E&T and to GOBLET clear; and to benefit GOBLET, by boosting the Foundation's income and helping to create tangible, branded products.

In addition, the EB has been working on a range of other tasks relating to the work of the SIGs, ENJ, the website, the Stitching bank account, membership, sponsorships, *etc*.

EMBnet.journal's Editorial Assistant: commitments and achievements

The EB convened a joint meeting with the ENJ EEdB to agree procedures and deadlines for hiring an assistant. A call[1] was prepared and launched at the beginning of August, with a 1 September deadline for submission of candidacies.

From seven candidates, three were selected, on the basis of their skills and expertise, for interview on 29 September. The ENJ EEdB unanimously converged on Antonio Santovito, who has proved to be a reliable, motivated and enthusiastic professional.

Antonio was officially hired at the beginning of October 2016 and immediately started work on the most urgent tasks. First was to repair hacker damage to ENJ, which resulted in the loss of almost all *EMBnet.journal* and *EMBnet.news* PDF and HTML galley files (*i.e.*, all files not directly included in the SQL database). This problem affected the productivity of the journal and delayed Antonio's work to develop a strategy to increase the number of published articles.

Nevertheless, his work was crucial for the recovery of ENJ's archives and much more.

After several weeks (and thanks to the personal archives of Laurent Falquet, Matej Stano and Lubos Klucar), Lubos, Antonio and George Magklaras (Chair of the Technical Management SIG) were able to restore all public content and to make the ENJ archives available again.

Erik's team set up a new backup strategy, eliminating the possibility of this kind of catastrophe in future. However, this did not prevent the occurrence of fake

[1] http://www.embnet.org/news/editorial-assistant-position-embnetjournal

Article history
Received: 2 March 2017
Published: 14 April 2017

ENJ article comments (advertisements). The worst attack happened in November 2015, when several hundred fake comments appeared in article Web pages within a few days. The comments were removed, corresponding users accounts were disabled and captcha was implemented on both 'user registration' and 'add comment' pages. This almost completely halted the problem – only a few such comments have since appeared on the site; however, it cannot be completely eliminated if we want to keep the commenting feature available. Hence, for now, the only effective prevention is manual removal of fake comments by Lubos and Antonio.

Fortunately, the ENJ OJS installation hasn't shown any major technical problems, although we are still running on an outdated version of the OJS (2.4.3.0, January 2014). To ensure our installation is robust and reliable, it is essential to update to the newest OJS release as soon as possible, and to make regular updates. At the same time, the server running the OJS must be protected by security measures.

Thanks to the professional support of Antonio, the EB was able to release, in the same period, a re-styled EMBnet leaflet and a short presentation, which were immediately used for EMBnet-sponsored events. The ENJ logo was re-designed, as well as the layout of the different ENJ article types. Lubos and Matej also contributed to this work.

The dissemination strategy elaborated by Antonio, and its implementation, brought increased visibility to EMBnet and to the ENJ, as shown in the statistics in Antonio's activity report (see Santovito, 2016). From May 2016, he was also involved in the curation of the content (*e.g.*, announcement of events and news) of EMBnet's website, and in editorial work relating to the production of *EMBnet.digest*.

Prior to the AGM, Antonio expressed his willingness to restyle EMBnet's website. This was a personal gift from him in recognition of the opportunity to work with EMBnet. The website will have a more modern design, in line with ENJ's new look-and-feel, will conform to the latest Web-design standards, and will be accessible from mobile devices. The new website will be released in spring 2017, hopefully on a new server.

Overall, the investment in an EA has been extremely positive and fruitful. Thanks to Antonio's work, we obtained professional dissemination products, implemented an effective dissemination strategy, and improved ENJ's visibility on the Web. Moreover, thanks to his availability to work on tasks that were not included in his duties, such as re-designing the ENJ and EMBnet websites, we did not need to assign the other fellowships forecast in the investment strategy to accomplish these tasks.

The opportunity to renew the EA contract will be evaluated during the 2016 AGM. Technical issues relating to the management and security of the EMBnet and ENJ websites, and updating the OJS, have yet to be solved, and will also be discussed.

EMBnet/GOBLET ring-fenced donation and Educational Assistant

In May 2016, EMBnet's bank account became locked owing to increased international security measures. To gain access to our account, and allow EMBnet to continue its business, the EB was involved in a long, complex negotiation with the bank. The bank mandated that at least two Dutch residents must be registered at the Camera of Commerce as Board representatives. Gert Vriend and Barbara van Kampen kindly agreed to be registered, to allow the process of cancelling the old EB members and registration of new ones to go ahead. This procedure is still ongoing; meanwhile, we re-gained access to the bank account at the end of August 2016.

This issue both prevented the EB from finalising the agreement with GOBLET, and from renewing our ISCB affiliation on time.

Considering the time that had elapsed since the initial proposal, the EB proposed to update the Board regarding the status of the EMBnet/GOBLET agreement during the 2016 AGM.

EMBnet digest, sponsorships and organisation of AGM 2016

With the support of Teresa Attwood, Axel Thieffry (Chair of the P&PR SIG) and Antonio, the EB produced the monthly EMBnet.digest, including many of the InFocus articles, and disseminated information about their release via the mail list and LinkedIn, Twitter and Facebook social networks. Axel has reported positive access statistics.

As for sponsorship, during this last year, EMBnet sponsored two big bioinformatics events, the "*Joint NETTAB 2015 International Workshop and Integrative Bioinformatics International Symposium*[2]", held 14-16 October, 2015 in Bari (IT), and the "*Fourth International Society for Computational Biology Latin America Bioinformatics Conference (ISCB-LA)*[3]", held 21-23 November 2016 in Buenos Aires (AR), which was co-organised with the Asociación Argentina de Bioinformática y Biología Computacional (A2B2C).

The NETTAB & IB 2015 event was organised by Domenica, co-Chair of the event, alongside Paolo Romano (IRCCS San Martino IST, Geno (IT), Ralf Hofestädt (University of Bielefeld (DE) and Matthias Lange (IPK, Gatersleben (DE). The event included six keynote lectures, original research scientific talks and poster sessions, and attracted more than 80 participants from diverse European countries.

In association with the workshop, Domenica and Paolo also organised a "Two-Day Hands-on Tutorial[4]" on cutting-edge methods and approaches to key issues

[2] http://www.igst.it/nettab/2015/
[3] https://www.iscb.org/iscb-latinamerica2016
[4] http://www.igst.it/nettab/2015/programme/tutorials/

D'Elia et al. (2017) EMBnet.journal 22, e887
http://dx.doi.org/10.14806/ej.22.0.887

Figure 1. NETTAB & IB joint workshop 2015. A group photo taken during the tour in Bari's downtown.

in bioinformatics analysis of omics data. The tutorial took place on 12-13 October, and was hosted by the Department of Computer Science of the University of Bari and the INFN, which kindly provided the computational services and IT resources of ReCaS, a project financed by the Italian Ministry for Education, University and Research.

For this event, EMBnet granted 10 free tutorial registrations for young researchers and students attending the meeting. The tutorial consisted of full- and half-day parallel tracks on "*Genome re-Sequencing for the Detection of Genomic Variations in Human Diseases*" (full day), "*Structural Bioinformatics in Drug Discovery*" (half-day), "*Pathway Commons and BioPAX*" (half-day) and "*Analysis of small RNA-Seq Data and Identification of microRNA controlled Pathways*" (full-day). Experienced trainers were invited: Fabio Iannelli, from the FIRC Institute of Molecular Oncology, Milan (IT); Anna De Grassi, Professor at the University of Bari (IT); Pasqualina D'Ursi, from the CNR Institute of Biomedical Technologies of Milan (IT); Emek Demir, from the Sloan Kettering Institute, New York (USA); and Ioannis Vlachos, from the DIANA-Lab, Hellenic Pasteur Institute/Department of Computer & Communication Engineering, University of Thessaly (GR).

ISCB-LA aims to inspire and foster collaborations between regional scientists and students to advance research in bioinformatics and computational biology in Latin America. The request for sponsorship came from Ignacio E. Sánchez, representive of the Argentinian EMBnet centre (Universidad de Buenos Aires) and member of the organising and scientific committee. EMBnet will have an exposition space, where our banner and publicity materials will be shown and distributed to conference participants. Erik will attend the event as EMBnet's representative, and will demonstrate the eBioKit (more details will be reported during the AGM). A report on the conference will be published in ENJ.

2016 EMBnet Annual General Meeting

A proposal to organise the 2016 EMBnet AGM in April 2016, in conjunction with the SolBio International Conference and Workshop 2016 "Bioinformatics and Computational Biology for Innovative Genomics"[5], was sent to the EB by Cesar Bonavides-Martinez and presented during the virtual general meeting held on 16 September 2015. Although the proposal was interesting, the EB declined the invitation owing to the costs, which were high relative to the need to honour the investments agreed during the 2015 AGM. Therefore, Erik and Hadrien Gourle helped to organise a two-day practical course on next-generation sequencing, on 22-23 April 2016, in collaboration with Ana Conesa, coordinator of the DEANN project[6].

In October 2015, the EB launched a survey among EMBnet members seeking preferences for the location of the 2016 AGM. 15 Organisational Members answered the survey. Most preferred a European country, where costs would be cheaper. In addition to the proposal from Mexico, Erik, Domenica and Pedro Fernandes submitted offers. The cheapest proposal was from Domenica, to hold the AGM in Rome in conjunction with the 2016 NETTAB workshop, in collaboration with ELIXIR-IT. The proposal included, as satellite events, an ELIXIR Hackathon (organised by Paolo and Rafael Jimenez) and an ELIXIR/GOBLET tutorial (organised by Allegra Via and Terri), to be held in parallel before the workshop; a special session on Education & Training in Bioinformatics

[5] http://icmexico2016.soibio.org/
[6] http://www.deann.eu

was included in the main programme (also organised by Allegra and Terri). The EB-IB and Chairs of SIGs agreed on Domenica's proposal.

In March 2016, the kick-off meeting of the CHARME COST Action took place in Brussels (BE); Erik and Domenica attended. Erik was elected vice-Chair and Grant Holder and Domenica Chair of WG4. The Management Committee (MC) established the 2016-2017 action plan. Domenica and Erik proposed the NETTAB-EMBnet workshop as an event co-organised with CHARME, owing to the match in scope and aims. The main focus of the workshop is "*Reproducibility, standards and SOP in bioinformatics*", and the event will represent an exceptional opportunity to present CHARME to a large and relevant audience of bioinformaticians from different organisations, institutions and European countries. The proposal was warmly supported by the CHARME MC, and the dates fixed on 25-26 October for the workshop, and 24 October for the satellite events (tutorial and hackathon). The EMBnet AGM will take place immediately after, 27-28 October 2016.

Acknowledgements

Despite the problems of the last year, relating to the bank account and the hack of ENJ, we are proud of the year's achievements. Erik's unshakable optimism gave us the energy to keep on going in these situations, without hesitation. We also thank T.K. Attwood and E. de Villiers for their support, and P. Fernandes, A. Thieffry and A. Santovito for their valuable contributions.

References

1. Attwood TK (2016) An Active Investment Strategy for EMBnet - AGM workshop report, Oeiras, June 2015. EMBnet.journal **21**:e867. http://dx.doi. org/10.14806/ej.21.0.867
2. Santovito A (2016) EMBnet AGM 2016: EMBnet.journal Editorial Assistant Activity Report. EMBnet.journal **22**:e881. http://dx.doi.org/10.14806/ej.22.0.881

EMBnet.journal
Bioinformatics in Action

Mattila et al. (2017) EMBnet.journal 22, e879
http://dx.doi.org/10.14806/ej.22.0.879

Technical Notes

Set up your own bioinformatics server: Chipster in EGI Federated Cloud

Kimmo Mattila[1], Diego Scardaci[2], Marica Antonacci[3], Catalin Condurache[4]

[1]CSC - IT Center For Science, Espoo, Finland; [2]EGI Foundation; [3]INFN, Bari, Italy; [4]Rutherford Appleton Laboratory, Oxfordshire, United Kingdom.
Competing interests: KM none; DS none; MA none; CC none.

Abstract

Chipster is an easy to use data analysis platform for bioinformatics. It provides an uniform graphical interface for over 360 commonly used bioinformatics tools including several R/Bioconductor-based tools and standalone programs (*e.g.* BWA, TopHat). Chipster is based on a client-server system where the user runs locally a Chipster-client that submits analysis tasks to a Chipster server. Even though Chipster is an open source tool, there is no public Chipster server that would be open for everybody. Due to that, a researcher needs to have an access to some of the existing Chipster servers to be able to use this platform. Alternatively, a researcher can set up his own Chipster server. In this paper, we describe how a Chipster server can be launched EGI Federated Could environment, that provides resources for all European researchers. With the instructions provided here, any European researcher can launch and manage his own Chipster server, suited for needs of a small research group or a bioinformatics course. The setup described here is based on a collaboration of several European instances. Chipster[1] is developed by CSC – IT Center for Science Ltd. in Finland. European Grid Infrastructure (EGI) has fitted Chipster to cloud environment[2] and provides the cloud computing resources. Finally, Rutherford Appleton Laboratory hosts the CVMFS server[3] that provides the scientific tools and data sets for the Chipster servers running in EGI federated cloud.

1. Preparatory steps

The EGI Federated cloud environment can be used from Linux or Mac OSX machines. In order to launch a Chipster server in EGI Federated Cloud, the machine that is used to manage the Chipster server must have the following tools and files installed: a valid personal X.509 certificate; rOCCI command line client for managing cloud computing environment; `voms-proxy-init` command to create proxy certificates; settings files to connect the VOMS server hosting chipster.csc.fi VO.

In addition, the manager of the Chipster server must join the chipster.csc.fi Virtual Organisation. The manager needs to do these preparatory steps only once. After that, Chispster servers can be managed with the FedCloud_Chipster_manager tool. Note that the end-users who wish just to use the Chipster server running in EGI Federated Cloud, do not need to do any of these preparatory steps.

1.1 Grid certificates and VO membership

EGI Federated Cloud uses X.509 certificates for user authentication. Researchers from the member countries of GÉANT network[4] can use the DigiCert certificate service[5] to obtain a personal grid certificate. Users from other countries should use their local certification authorities. Once you have a grid certificate installed in your browser, you can join the chipster.csc.fi Virtual Organisation(VO) in the VO home page[6].

1.2 Installing rOCCI and VOMS client

The management of Federated Cloud resources is done using rOCCI, a ruby based implementation of OCCI standard. The authentication in EGI federated cloud is done using proxy certificates generated with command voms-proxy-init. The instructions to install these tools to your local machine can be found from EGI wiki site[7]. Once you have installed the rOCCI and `voms-proxy-init` commands, you must still define the connection to Chipster.csc.fi VO management server (VOMS). To do this, first create directory "/etc/grid-security/vomsdir/chipster.csc.fi" and go to this directory:

```
mkdir /etc/grid-security/vomsdir/
chipster.csc.fi
cd /etc/grid-security/vomsdir/chipster.
csc.fi
```

[1] http://chipster.csc.fi
[2] https://www.egi.eu/federation/egi-federated-cloud
[3] https://www.gridpp.ac.uk/wiki/RAL_Tier1_CVMFS
[4] http://www.geant.org
[5] https://www.digicert.com/sso
[6] https://voms.fgi.csc.fi:8443/voms/chipster.csc.fi
[7] https://wiki.egi.eu/wiki/Fedcloud-tf:CLI_Environment

Article history
Accepted: 5 January 2017
Received: 18 January 2017
Published: 06 February 2017

Mattila et al. (2017) EMBnet.journal 22, e879
http://dx.doi.org/10.14806/ej.22.0.879

Then create a file "voms.fgi.csc.fi.lsc" that contains the following 2 lines:

```
/O=Grid/O=NorduGrid/CN=host/voms.fgi.csc.fi
/O=Grid/O=NorduGrid/CN=NorduGrid Certification Authority
```

If you already have file "/etc/vomses", move the file "/etc/vomses" to "/etc/vomses/old_vomses" (vomses will be a directory now). Create a file "chipster.csc.fi-voms.fgi.csc.fi" in "/etc/vomses" and write inside the following line:

```
"chipster.csc.fi" "voms.fgi.csc.fi" "15010" "/O=Grid/O=NorduGrid/CN=host/voms.fgi.csc.fi" "chipster.csc.fi"
```

1.3 Obtaining keys and FedCloud_chipster_manager

FedCloud_chipster_manager is a help tool that can be used to manage Chipster instances in EGI Federated Cloud. It can be downloaded from the Chipster git-hub[8]. Some of the FedCloud_chipster_manager operations require that user provides encryption key pair that is used to access the virtual machine. The key pair can be created for example with command:

```
ssh-keygen -t rsa -b 2048 -f FedCloudKey
```

2. Managing Chipster server

2.1 Setting up VOMS proxy

Before launching or managing virtual Chipster servers, you have to create a temporary proxy certificate that is used to authenticate to EGI Federated Cloud environment. If you have the voms-proxy-init command installed and a valid X.509 certificate in your ".globus" directory, you can create a temporary proxy certificate with command:

```
voms-proxy-init --voms chipster.csc.fi --rfc --dont_verify_ac
```

The command above asks the password of your certificate and creates a proxy certificate that is valid for 12 hours. Note that voms-proxy-init requires that you are using OpenJDK-based Java environment. Other Java environments cause error messages like: "Credentials couldn't be loaded".

2.2 Launching a Chipster server

Once you have done all the preparations, you can launch a new Chipster Virtual Server with command (assuming you have the FedCloud_chipster_manager tool in your current working directory):

```
./FedCloud_chipster_manager -key keyfile -launch
```

This launching command uses default values, for resources and user accounts linked to the Chipster. Option `-volume_size` can be added to modify the size of the data volume (in gigabytes) that is used to store the data during the computing. The default size of the volume is only 20 GB, which is enough for testing, but for real usage a bigger data volume may be needed. By default, only one Chipster account (user: chipster, password: chipster) is created to a new Chipster server. A list of user accounts for a new Chipster server can be defined with option -users. The argument for this option should be a file containing a list of accounts in format:

```
user_name:password:expiration_date
```

The expiration date is defined with format: yyyy-mm-dd. For example file "accounts.txt" could look like following:

```
trng1:4eoU8hmx:2017-05-15
trng2:4eoU8hmx:2017-05-15
```

Note that these accounts are just Chipster server accounts, not Linux accounts that could be used to open terminal connections to the virtual machine. Launching a Chipster server with these accounts and 100 GB storage size could be done with command:

```
./FedCloud_chipster_manager -launch -key FedCloudKey -volume_size 100 -users accounts.txt
```

The launching process can take tens of minutes. In the end the launching process prints out information about how the server can be accessed. For example:

```
----------------------------------------
You can now connect your virtual machine with command:

ssh -i FedCloudkey ubuntu@90.147.102.41

The Chipster server can be connected with URL: http://90.147.102.41:8081
```

The users can now use the URL to use the Chipster server while the ssh connection is intended for managing the Chipster server. Note that each Chipster server will get a unique IP address. The address is assigned by the cloud environment and it can't be set or modified by the user.

2.3 Other management tasks

In addition to launching Chipster servers, FedCloud_chipster_manager tool can be used to manage an existing server. You can use FeCloud_chipster_manager with option `-list`, to list your virtual Chipster servers running in the EGI Federated Cloud. Option: `-status` makes FedCloud_chipster_manager to look for Chipster VMs launched by the user, and to check the status of the Chipster server running in the VMs found. In this

[8] https://raw.githubusercontent.com/chipster/chipster/master/src/main/admin/fedcloud/FedCloud_chipster_manager

Mattila et al. (2017) EMBnet.journal 22, e879
http://dx.doi.org/10.14806/ej.22.0.879

case, you must also use the `-key` option to define the key file, that was used to launch the server. The password for the key file is asked for each server to be connected. The option `-restart` makes FedCloud_chipster_manager restart the Chipster server running in the given Federated Cloud VM instance. This option can be used for example to fix the server if the Chipster server is using internal IP address instead of public IP address. For example, restarting the Chipster server running in instance `/compute/86b97ed5-e256-4bce-83b5-aa3a41920975` can be done with command:

```
./FedCloud_chipster_manager -key
FedCloudKey -restart
/compute/86b97ed5-e256-4bce-83b5-
aa3a41920975
```

To completely delete the virtual machine running in EGI Federated Cloud you can use option `-delete`:

```
./FedCloud_chipster_manager -delete
instance-ID
```

For more detailed management, you can open a terminal connection to the virtual machine and apply the instructions in the Chipster technical manual[9].

[9] https://github.com/chipster/chipster/wiki/TechnicalManual

3. Using your Chipster server in EGI Federated Cloud

The Chipster Virtual organisation can provide only limited resources for the Chipster user community. By default, the FedCloud_chipster_manager starts a Chipster server on a virtual machine that has 4 computing cores with a total of 8 gigabytes of memory. This is not much, but it should be enough to serve the needs of a small research group (only a few simultaneous users). If you wish to use a larger virtual machine, please contact the Chipster VO manager. Once launched, the server can be kept up and running as long as the data processing continues. This can be weeks or months, but finally the Chipster server should be shut down by the owner of the server.

If your server has been running longer than 4 months, the VO manager can ask the owner of Chipster server to send a report about the usage of the server. When using the Chipster in EGI Federated Cloud, you should remember that the intermediate data at the servers is not back-upped. If you need to rebuild your Chipster server, the data in the previous version will be lost when the old version is removed. Further, you should remember that current setup for running Chipster in EGI Federated Cloud is still under testing and development. We do not guarantee uninterrupted access to the resources at all times.

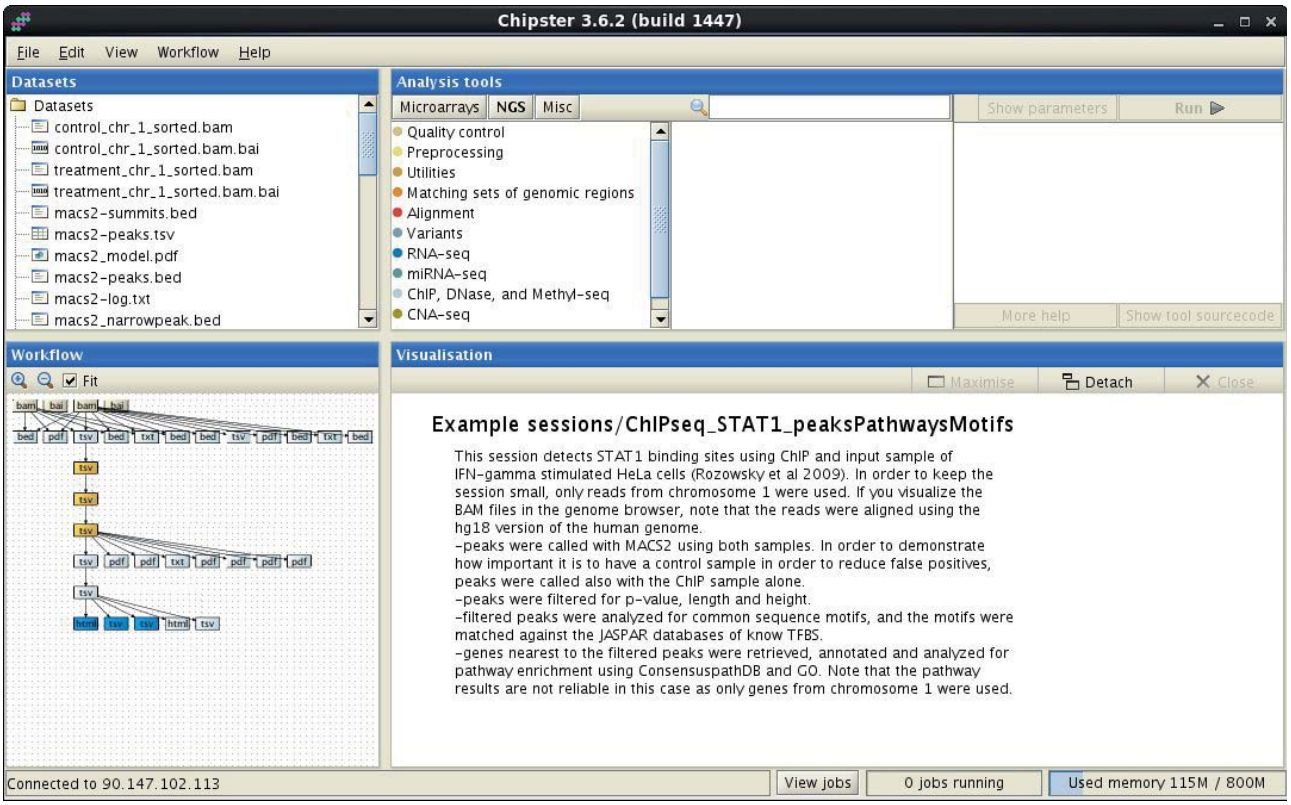


Figure 1. Chipster client started from a Chipster server running in EGI Federated Cloud.

With CHARME towards standardisation in life sciences

by Domenica D'Elia[1], Babette Regierer[2], Susanne Hollmann[3]

[1] CNR-Institute for Biomedical Technologies, Italy, [2] SB ScienceManagement UG, Berlin, Germany, [3] University of Potsdam, Potsdam, Germany

On March 21st, representatives from 26 countries met in Brussels to execute the kick-off meeting of the new COST Action (Cooperation in Science and Technology) **CA15110: "Harmonising standardisation strategies to increase efficiency and competitiveness of European lifescience research (CHARME)"**. The participants exchanged information about the need for understanding formats and standards for biological data and computer models in systems biology research, and elected Chair, Vice Chair and working group leaders.

An essential prerequisite of modern life-science R&D is high quality research data. By enabling the reuse of research assets, research becomes considerably more efficient and economical. This can only be achieved reliably and efficiently if these are generated according to standards and Standard-Operating-Procedures (SOPs). Thus, standards represent important drivers in the life sciences and technology transfer because they guarantee that data become accessible, shareable and comparable along the value chain.

Several initiatives have launched the development and implementation of standards. Unfortunately, these efforts remain fragmented and largely disconnected. CHARME aims to merge different approaches in the field, with a particular emphasis on systems biology, and thus to avoid too many different solutions being generated in parallel universes that 'in the worst case' are neither compatible nor suitable for largescale approaches.

CHARME will increase awareness of the need for standards, enabling the reuse of research data and its interoperability within the scientific community. CHARME provides a common ground for researchers from academia, research institutes, SMEs and multinational organisations. Representatives of each participating country can be found at: http:// www.cost.eu/COST Actions/ca/CA15110? management.

For further information, please feel free to send an e-mail to info@cost-charme.eu. From May, further information will be available on the Action's website: http:// www.cost-charme.eu

InSyBio joins EMBnet

by Konstantinos Theofilatos

Marketing & Sales Manager, InSyBio

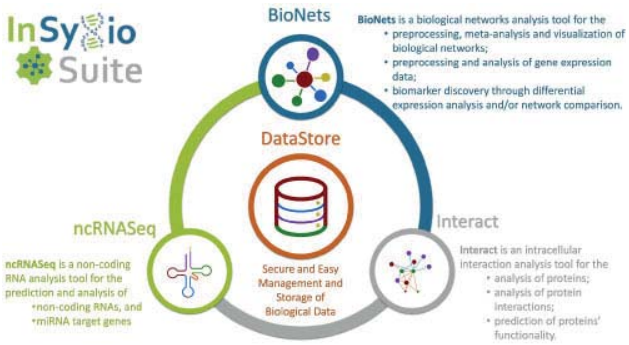
InSyBio (**In**telligent **Sy**stems **Bio**logy) is a bioinformatics company that focuses on developing computational frameworks and tools for the analysis of complex biological data. The key objective of our analysis lies in the discovery of predictive integrated biomarkers with increased prognostic and diagnostic aspects for the personalised Healthcare Industry.

InSyBio has developed an on-line clouddriven suite of software tools – InSyBio Suite – that helps researchers to identify early-stage biomarkers. By contrast with existing solutions, the InSyBio Suite consists of tools that integrate data from various sources and provide comprehensive results and meaningful knowledge using advanced big data-oriented artificial intelligence methods. Some of its methodologies have already been published in international scientific journals and conferences, while others are submitted or under submission for patent approval.

InSyBio's services can save time and money from pharmaceutical, biotechnology and nutrition-focused companies by providing:

- biomarker-discovery tools from transcriptomics experiments using its unique systems-medicine pipeline and biological network-analysis tools;
- access to its supreme human proteinprotein interactions database (and protein complexes) and interactomics analysis tools;
- fast, accurate and easy-to-use bioinformatics analysis of non-coding RNAs.

InSyBio recently launched its first commercial version, and its evaluation version is currently being tested by hundreds of users. The first feedback from InSyBio Suite's users has confirmed that it improves and significantly speeds up their biomarker discovery process. It is noteworthy that several scientific projects, including bioinformatics tasks that have been conducted with the InSyBio Suite, have begun to be published in prestigious journals. The 3- and 12-month licences for the InSyBio Suite are being sold via Internet- and direct sales.



A demo version of InSyBio Suite is freely available at demo.insybio.com.

To request a free one-month licence or to purchase the InSyBio Suite, please email us at info@insybio.com. For more information, visit our Web page at www.insybio.com.

Bioinformatics community attracts young professionals

by Pedro Fernandes

EMBnet Portugal

In February 2016, the University of Minho (PT) held its annual bioinformatics event, Bioinformatics Open Days. This student-led meeting uses a lightweight format that brings together bioinformatics students, teachers and professionals.

For students, it provides a window of opportunity to see how their newly acquired knowledge can project outside the school environment, opening their minds to contacts in an unprecedented way, not only to academic perspectives, to continue their studies, but also to the industrial world, where they may seek employment.

For teachers, it is a unique occasion to review their work and strategies for keeping their teaching content up-to-date and in consonance with external demand.

For bioinformatics professionals, it allows networking with like-minded people, promoting the discipline and best practices in its application.

At this event, I again had the chance to show that training is needed in this area, and how it can complement formal education. In this context, I explained what EMBnet provides in terms of links to worldwide bioinformatics communities, and used the opportunity to offer an '*entry package*' to students; this resulted in 17 new members.

As a welcome token, these new members will not pay the 2016 membership fee. As part of the package, all were invited to supply links to their LinkedIn profiles, which EMBnet displays in its new student-member page: http://www.embnet.org/embnet-studentmembers.
In this way, students may showcase their profiles to members of our community and beyond, open new possibilities for cooperation, further studies, etc. For EMBnet, the initiative is an innovative way to showcase recently qualified Bioinformatics MSc. students with a keen interest in community interactions and cooperation, and brings with it new blood.

We welcome this new cohort to our activities, and look forward to seeing how it may progress with mutual benefits.

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.journal
Bioinformatics in Action

The art of biocuration

a special article to celebrate Swiss-Prot's 30th Anniversary

Vivienne Baillie Gerritsen, Marie-Claude Blatter

Museums have their curators. Art galleries too. Their job is to look after collections they are knowledgeable about and present them to an audience in a way that makes sense and is informative. Biocurators do the same. Ever since the advent of computers and advanced technology in the life sciences, the quantity of biological data has grown exponentially and been stored in databases. The simple piling up of data, however, is of little help not only to researchers but also to computers. To be useful, they need to be sorted some way or another. Such a step is easily performed by specialized software. But as for many things, without a human touch something lacks. Swiss-Prot is a protein sequence database that sprung into existence 30 years ago when protein sequences were still trickling in. In those days, every sequence could be nursed. Today, however, millions of protein sequences are produced on a monthly basis. How does Swiss-Prot cope? Thanks to its biocurators.

Photo: ©Vivienne Baillie

The life of a biocurator may sound simple. But it is not. Dozens of articles have been published to define this new species of life scientist. Biocurators are professional scientists who are trained to collect, sort, synthetize, organise and validate biological information which is then disseminated via databases. The nature of their job demands the patience and thoroughness of a librarian and they have been referred to – whether endearingly or not – as 'museum cataloguers of the internet age', 'those who prefer computers to pipettes', 'self-confessed bookworms', or 'monk copyists'. A somewhat narrow-minded and outdated view.

Biocuration is not only a science in itself but, like the technology that nourishes it, it is continuously evolving. Biocurators have to adapt fast as they face the unending production of huge amounts of data and have to deal with ever-changing biological knowledge. In Swiss-Prot, for instance, although the number of protein sequences does not exceed the half million mark, it provides data that have been checked and improved by its biocurators. This upgraded information is in turn (re)used for automatic annotation of new incoming sequences. As such, biocurators have become an essential and central piece of the life science puzzle.

About 70 biocurators currently work for the SwissProt database – which is now part of the UniProtKB knowledgebase – in Switzerland (SIB), England (EBI) and the US (PIR), most of whom (50) work at SIB. They are biologists, biochemists and chemists with a strong background in wet-lab research, and generally hold PhD degrees. Their job consists in reviewing experimental and predicted data for each and every protein with a sharp critical eye, as well as verifying in detail every protein sequence. In this way, they provide a complete overview of any information there exists on a given protein.

The information is extracted from various sources; most of it regards experimental data which is drawn from the literature. Biocurators reconcile any conflicting results and then compile them into a concise, comprehensive report – both in free text and structured format – with controlled vocabularies that can be read by a machine. The process of expert biocuration adds a wealth of knowledge to UniProtKB/Swiss-Prot records, and includes information related to a protein's function, structure and subcellular location but also a wide range of sequence features such as active sites or

post-translational modifications, besides the protein's interactions with other proteins.

Huge amounts of data means tons of articles. Does a biocurator read every single publication there is on a given protein? The answer is no. A crucial step in expert curation is knowing how to identify a representative subset of publications that will provide a complete overview of the information that is available at the time. Any information added manually to Swiss-Prot is linked to its source – in this way, users can trace each piece of information to its origin. For maximum efficiency, Swiss-Prot biocurators generally deal with groups of related proteins – such as proteins that belong to the same family or are found in different species – as the background knowledge already exists within the database. Swiss-Prot biocurators also collaborate and leverage the work of complementary curated resources to ease consistency and data exchange, thereby ensuring that biocuration efforts are not duplicated.

But this is all very theoretical. How exactly do Swiss-Prot biocurators deal with the knowledge that is pouring in? The recent characterization of the Notum protein in humans and the common fruit fly, Drosophila melanogaster, provides an excellent example. One Swiss-Prot biocurator waded through over 100 articles that were linked to the term "Notum" – which also happens to be the name given to the dorsal portion of an insect's thoracic segment. According to well-established criteria, two papers were sufficient to extract the information needed to acquire a comprehensive overview of the protein.

Notum was initially characterized in D.melanogaster as a protein that had a vital role in developmental morphogenesis by inhibiting the Wnt signalling pathway – a pathway that passes signals into a cell via cell surface receptors. For over 20 years, scientists thought the pathway was inhibited by Notum via the hydrolysis of specific proteoglycans known as glypicans. However, the two selected articles contradicted these results: inhibition occurs via serine depalmitoleylation of the Wnt proteins themselves. Thus the role of Notum as an inhibitor of the pathway is confirmed, but the mechanism is different.

Thanks to this new experimental data, an update of the function – and the protein's name – was echoed across 10 different species in UniProtKB/SwissProt. But the biocuration did not stop there... New gene ontology (GO) terms were created, a new enzyme commission (EC) number was issued, the positions of the enzyme's active sites were annotated from its 3D structure in the Protein Data Bank (PDB), and the modification of the Wnt protein by depalmitoleylation and its consequences have been annotated using controlled vocabulary defined by the RESID database of Protein Modifications. And, so as not to lose track of any information even if it has become redundant, the former function of the Notum protein is described in detail under a 'CAUTION' section.

This example illustrates how essential expert biocuration is, and how the correct identification of only a few targeted publications can give rise to information that is not only vital but has a chain reaction effect. Thanks to manual biocuration, new biological functions continue to be unveiled within what are thought to be well characterized protein families. Correct and up-to-date information is spread across other databases in addition to being fed back into the automatic annotation and function prediction systems loop. Existing close and mutually beneficial collaborations between different resources are also heightened, demonstrating yet again how essential expert biocuration is in maintaining biological knowledge.

Cross-references to UniProt

Palmitoleoyl-protein carboxylesterase NOTUM, *Homo sapiens* (Human): Q6P988
Palmitoleoyl-protein carboxylesterase NOTUM, *Drosophila melanogaster* (Fruit fly): Q9VUX3

proteinspotlight
> ONE MONTH, ONE PROTEIN <
Protein Spotlight (ISSN 1424-4721) is a monthly review written by the **Swiss-Prot** team of the **SIB Swiss Institute of Bioinformatics**. Spotlight articles descrive a specific protein or family of proteins on an informal tone.
http://web.expasy.org/spotlight
SIB Swiss Institute of Bioinformatics

EMBnet.journal
Bioinformatics in Action

On releasing tension

Vivienne Baillie Gerritsen

Like life, cells are subject to continuous change. Nothing in the vicinity of a cell remains still-unless death has interrupted its course. And the same goes for the inside of each cell. All sorts of molecules are being shuttled from one part to another, after having been created or on their way to being degraded. The cell membrane is also a very dynamic and supple structure, with molecules wandering through it constantly. One means of transport for shifting molecules around are known as endosomes. Endosomes are formed by the invagination of one part of the cell membrane – like a bubble budding towards the inside of the cell – which is then cleaved and able to float free in the cell's cytoplasm. Invagination then occurs on the surface of the endosomes themselves to form even smaller bubbles – or vesicles – that are in turn also set free. Like endosomes, these inner vesicles are a means of molecular transport too. It's a sort of Russian doll experience… The art of invagination per se may sound simple but it involves a lot of imagination on behalf of biology. Recently, scientists discovered a protein coined Snf7 that is providing hints as to how endosome invagination may occur, and the subsequent creation of vesicles.

by Zaq Guimarães, 2016. Courtesy of the artist

In the same way we have buses, trains and planes to travel around the world, molecules have different ways of moving around a cell, or indeed from one cell to another. Endosomes are just one means of transport inside a cell and are made out of cell membrane, from where they bud. Imagine a cell, and that you're inside it. Now imagine a bubble budding from the membrane towards you, as though some invisible finger were pushing from the outside towards the inside. Now watch the bubble being severed from the membrane and left to float on its own. This is what an endosome is. And inside, it is carrying all the molecules – or cargo – that were part of the cell's membrane as it was being formed. Now, let's take it a little further: imagine sitting inside an endosome. Exactly the same chain of events occurs at the level of the endosome's membrane: a bubble buds, is severed, and floats off on its own. Bubbles that bud off the inside of endosomes are known as vesicles. And, like endosomes, they also transport molecules.

All kinds of proteins and protein complexes are involved in the making of endosomes and vesicles. Vesicle formation is driven by what are known as "endosomal sorting complexes required for transport", or ESCRTs. There are three ESCRTs (I, II and III) which, between them, relay cargo to where invagination will take place, promote invagination and then set the nascent vesicle free by severing it from the endosomal membrane. The ESCRT machinery and its role in budding were first observed in yeast for transporting molecules destined for degradation. However, it is now known that such budding events are similar to those used for HIV release and daughter cell abscission.

Snf7 is the most abundant protein in the ESCRT-III complex and seems to be by Zaq Guimarães, 2016 Courtesy of the artist particularly involved in the budding process itself, i.e. membrane invagination. Snf7 has a highly structured core domain of four alpha-helices, with a C-terminal alpha-helix that folds back onto the core domain. When the Cterminal end unfolds, Snf7 adopts an elongated "open" form that promotes protein-protein interactions; this is believed to be at the heart of membrane budding. What happens is Snf7 monomers can spontaneously fold into a single closed ring – or nucleation ring – which will, at one point, open. The "open form" induces Snf7b polymerisation; concomitantly, Snf7 is bent into a curve which is not "natural". When a second Snf7 monomer is added, it is also bent into an unnatural curve and so on. As the filament grows, the geometrical shape that

eventually emerges is a two-dimensional spiral filament of Snf7 monomers on the cytoplasmic side of the endosome. Moreover, since the curvature of each open Snf7 monomer is unnatural, there is a mechanical stress which is heightened as the spiral filament elongates. And mechanical stress spells locked-up energy – in this case elastic energy.

There comes a point when the spiral stops growing – which seems to happen once the filament's length has reached 260nm. Why, no one really knows. Perhaps other proteins form a sort of wall in which it is confined, perhaps it reaches an energy threshold. Nevertheless, literally bursting with elastic energy, the spiral is thought to release its tension by lunging towards the inside of the endosome – in the manner of a jack in the box – forming in its wake a superhelix which drags the membrane along with it, shaping the nascent vesicle. In this way, and to cut a long story short, Snf7 has the power to deform membranes by compressing elastic energy in a two-dimensional spring which ultimately expands.

It took molecular biology, mathematics, physics, chemistry and structural biology to imagine and test such a hypothesis, which seems to have convinced the scientific community. Snf7 filaments do indeed form spiral springs on endosome membranes and, upon some form of external cue, they are expected to release their stored elastic energy to form the budding vesicles. This demonstrates that a protein – depending on its conformation – has the ability to distort part of a membrane. Growing spirals no doubt speckle the endosome surface ultimately giving rise to vesicles, while similar invaginations occur at the level of the cell membrane. In this way, molecules are entering, travelling within and leaving the cell continuously. However, though Snf7 is proving to be essential for membrane invagination, it is also involved in membrane scission with the rest of the ESCRT complex – and membrane scission is an activity scientists are also very eager to understand at the molecular level.

Cross-references to UniProt

Vacuolar-sorting protein SNF7, *Saccharomyces* cerevisiae (Yeast) : P39929
Palmitoleoyl-protein carboxylesterase NOTUM, *Drosophila melanogaster* (Fruit fly): Q9VUX3

References

1. Chiaruttini N., Redondo-Morata L., Colom A., Humbert F., Lenz M., Scheuring S., Roux A.
Relaxation of loaded ESCRT-III spiral springs drives membrane deformation
Cell 163:866-879(2015)
PMID: 26522593
2. Carlson L.-A., Shen Q.-T., Pavlin M.R., Hurley J.H.
ESCRT filaments as spiral springs
Developmental Cell 35:397-398(2015)
PMID: 26609952

proteinspotlight
> ONE MONTH, ONE PROTEIN <
Protein Spotlight (ISSN 1424-4721) is a monthly review written by the **Swiss-Prot** team of the **SIB Swiss Institute of Bioinformatics**. Spotlight articles descrive a specific protein or family of proteins on an informal tone.
http://web.expasy.org/spotlight
Swiss Institute of Bioinformatics

You want it darker*

Vivienne Baillie Gerritsen

We are highly adaptable. We have been for the past few million years, and continue to be so on a daily basis. Whichever way you look at it, the art of adaptation really is just a way of preserving your integrity – physical or psychological – and coping the best way possible with the environment you are evolving in. Throughout the animal world and over the aeons, the capacity to adapt has always been Nature's answer to predators and hostile physical, geographical or climatic conditions. In short, adaptation is the best way to survive and Charles Darwin was the first to explain animal diversity in this way in his Origin of Species. Ever since, the study of fossils or more recently genomes is a constant support to Darwin's theory of what was then coined 'natural selection.' But it all remained very theoretical; it is difficult to observe animal adaptation within a man's lifetime when it occurs over thousands or even hundreds of years. However, there is a moth in Great Britain, known as the Peppered Moth which, over a relatively short period of time, adapted to the effects of pollution resulting from the Industrial Revolution by changing the colour of its body and wings. The protein involved in this change was recently discovered and named 'the cortex protein.'

* Title taken from Leonard Cohen's latest album, "You Want It Darker"

"Standing figures and telegraph poles"
by Theodore Major (1908-1999).

By the early 20th century, air pollution caused by the Industrial Revolution had begun to hit parts of Europe and, in particular, parts of Great Britain quite hard. Following decades of manufacturing, soot had settled on buildings, walls, fences and trees and gradually painted towns and the nearby countryside a dull grey. Needless to say, it had a profound effect on plant and animal life living close to industrial centres. During this period, amateur naturalists and moth collectors couldn't help but notice that a moth known as *Biston betalaria*, or more commonly the peppered moth, was changing colour: the original black-speckled white moth was gradually becoming only black. It wasn't due to the deposit of soot on their wings, but seemed to be an actual modification in the colour of their wings and bodies. The phenomenon was termed 'industrial melanism' to describe the blackening of the moths' colour – as opposed to melanochroism which is the darkening of any given colour. Many theories attempting to explain the process emerged, while in the wake of Charles Darwin's Origin of Species published in 1859, evolutionists saw the change as an example of natural selection taking place before their very eyes. For camouflage purposes, the peppered moths were slowly adopting a darker colour, less conspicuous when resting against the bark of a tree or on a soot-darkened wall. However, their theory needed evidence.rest of the ESCRT complex – and membrane scission is an activity scientists are also very eager to understand at the molecular level.

The English entomologist J.W.Tutt (1858-1911) was one of the first to describe the colour change in peppered moths. He saw it as a form of crypsis, i.e. an animal's ability to make itself discreet – invisible to predators for instance – by means of camouflage or mimicry. The normally occurring light-coloured lichen on trees was gradually being killed off by soot and leaving the bark of trees, bare and dark. In the presence of predators, the original peppered moth would have been very visible on such a surface, while the all black version – named *carbonaria* today – would be unnoticed. In the 1950s, the British geneticist and lepidopterist H.B.D. Kettlewell (1907- 1979) carried out a few elegant investigations which by Theodore Major (1908-1999) "standing figures and telegraph poles" showed that the change of colour in the moths was not a case of crypsis but instead a case of natural selection where the original light-coloured peppered moth had been gradually replaced by carbonaria in industrial areas. He demonstrated this by placing both types of moth first on dark and then on light surfaces. In each experiment, most of the moths he managed to recapture were those whose colour matched the colour of the backgrounds, thus supporting the theory of natural selection. On a dark background, predators went straight for the original peppered moths, while on a light background their black kin were caught.

What was happening on the molecular level? The wing patterns of butterflies and moths (*Lepidoptera*) – of which there are hundreds of thousands of different species worldwide – are, in fact, a striking example of diversity brought about by natural selection. The different colours observed are a combination of the quantity of melanin in the cells – known as scale cells – that form the patterns seen on lepidopteran wings, and the cells' optics. The cortex protein, or simply cortex, has a direct role in the amount of melanin present in scale cells. Cortex belongs to a fast-evolving subfamily of cell-cycle regulators, and may well regulate pigmentation during early wing disc development, perhaps by regulating scale cell development itself. As such, cortex will have become a major target for natural selection involved in pigmentation; pattern variation in *Lepidoptera*, and its expression does indeed vary significantly between butterflies and moths with different wing patterns. The first reported sighting of *carbonaria* is said to have occurred in 1848 in Manchester, though previous sightings seem to have been made. This would imply that the mutation linked to industrial pollution would have appeared shortly before soot levels rose – although it could also have existed undetected at a low frequency for hundreds of years...

Cortex controls melanism in the peppered moth. But what is the exact nature of the mutation? Breeding experiments in the early 20th century had already suggested that industrial melanism was the consequence of an inherited form of a single dominant or semi-dominant gene. It wasn't before the 21st century that the molecular identity of the mutation was unveiled, and came as a bit of a surprise. It turned out to be the doings of a large tandemly repeated transposable element which inserts itself into the non-coding part of the cortex gene. Statistical inference based on the distribution of *carbonaria* haplotypes has even pinpointed the transposition event to the year 1819 – which would be consistent with the historical record. As for the effects of the mutation on cortex itself, the protein is expressed more abundantly in *carbonaria* than it is in the light-coloured peppered moth.

Driven by the intense industrialisation of a nation, over a relatively short period of time the colour of the peppered moth shifted from black-speckled to black. The inserted transposable element increases the abundance of cortex in scale cells causing them to darken in the process, though how exactly remains a mystery. Cortex has thus proved to have an important role in a spectacular story of rapid adaptation in the peppered moth. However, it is rare that a protein acts on its own. Industrialisation didn't happen on the same scale everywhere either. Up until the 18th century, London was the most industrialised and polluted city in Britain. Later, other places were hit although a few have since stepped back and are considered rural nowadays. The atmospheric pollution in these areas has also varied, as did the darkening of surfaces, food abundance and predators. Migration too can influence frequency changes… So, it is difficult to explain industrial melanism on the sole basis of the cortex protein. What can be said with certainty, however, is that it did play a major role, and what happened to the peppered moth offers an elegant and classical example of natural selection induced by human activity and driven by selective predation, over a brief and observable period of time.

Cross-references to UniProt
Protein cortex, *Biston betularia* (peppered moth) : P0DOC0

References

1. Van't Hof A.E., Campagne P., Rigden D.J., Yung C.J., Lingley J., Quail M.A., Hall N., Darby A.C., Saccheri I.J.
The industrial melanism mutation in British peppered moths is a transposable element
Nature 534:102-105(2016)
PMID: 27251284
2. Nadeau N.J., Pardo-Diaz C., Whibley A., Supple M.A., Saenko S.V., Wallbank R.W.R., Wu G.C., Maroja L., Ferguson L., Hanly J.J., Hines H., Salazar C., Merrill R.M., Dowling A.J., French-Constant R.H., Llaurens V., Joron M., McMillan W.O., Jiggins C.D.
The gene cortex controls mimicry and crypsis in butterflies and moths
Nature 534:106-110(2016)
PMID: 27251285

proteinspotlight
> ONE MONTH, ONE PROTEIN <
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 descrive a specific protein or family of proteins on an informal tone.
http://web.expasy.org/spotlight

A walk on the rough side

Vivienne Baillie Gerritsen

Life can be hard. There are times when you find yourself in the most unfriendly circumstances and, more often than not, the best way to deal with the situation is to find your own solution and wriggle your own way out. Living species are the most imaginative of beings when it comes to designing defence mechanisms. Some release nasty smells to ward off predators, or melt into the landscape and become invisible to them. While others live in protective shells, inject toxins that paralyse, or are simply wise enough to walk away from danger. Many species have also chosen to live in places no one would ever dream of settling down in and, over time, have developed ways to flourish in severe temperatures, faced with complete dehydration and under crushing pressures. Among these are tardigrades, also known as water bears or moss piglets. One tardigrade in particular, Ramazzottius varieornatus, lives in very harsh conditions and even seems to have found a way to survive harmful radiation. Thanks to a protein which has been dubbed damage suppressor protein.

Waterbear, by Thomas Shahan. Courtesy of the artist.

The first tardigrade was described in 1773 by the German zoologist Johann August Ephraim Goeze (1731-1793), and the name *tardigrada* – literally meaning *slow stepper* – was coined three years later by the Italian priest and biologist Lazzaro Spallanzani (1729-1799). Tardigrades are tiny cylindrical eight-legged invertebrates that measure up to 2.1mm in length. They are divided into five segments. The first is the head, followed by four others from which protrude a pair of legs frequently punctuated by a set of claws. Despite its littleness, a tardigrade's internal structure is surprisingly complex: it has a complete digestive system and a well-developed nervous system consisting of rings around a mouth. About 1,200 different species are known to date, though it is thought that there are many more. As for their lifespan, adult tardigrades live an average of several months.

Water bears are scattered across the globe where they live in freshwater environments, marine environments but also terrestrial habitats such as in moss. Like many other multicellular extremophiles, they can survive harsh conditions – very high pressures, austere temperatures and lack of water – by slowing down their physiology and almost bringing to a standstill their vital organs to preserve them. In this dehydrated state, tardigrades literally shrivel up and decrease in size, and can remain in this state for as much as 20 years. Tardigrades are particular, however, in that they can survive extreme conditions both in their dehydrated (inactive) and active forms. And *R.varierornatus* seems to be fitted out with the best survival kit of them all.

How did *R.varierornatus* acquire such skills? Some claim that, over millions of years, tolerance-specific genes have been pumped into the tardigrade's genome by a process known as horizontal gene transfer, or HGT. So much so that up to 17.5% of its genes would be of foreign origin. This theory has met with much controversy, however. The genome of *R.varierornatus* is certainly expected to carry waterbear, by Thomas Shahan Courtesy of the artist stress-related genes if it is to survive in the environments it does, however its genome does not seem to carry more foreign genes than would be expected – i.e. about 1.2% of the total genome. What is more, besides acquiring tolerance-specific genes, *R.varierornatus* would have lost a few metabolic pathways that respond to stress – which is an unexpected state of affairs but must also have its say in protecting the tardigrade.

Fierce dehydration or radiation can do a lot of harm to molecules: they can, for example, shred DNA apart. R.*varierornatus* seems to be particularly protected against high-dose radiation. How? This faculty could simply be a side-product of the tardigrade's capacity to tolerate complete dehydration. Nevertheless, the recently characterized damage suppressor protein, or Dsup, may well have a central role in protecting DNA. When inserted into genetically engineered human cells, Dsup improved their radiotolerance by a staggering 40 to 50%. Dsup is a nuclear protein and expressed abundantly during the tardigrade's embryonic stage – precisely when DNA is being replicated and is hence vulnerable. So far, the protein has been found in no other living organism.

How exactly, though, does Dsup protect the tardigrade's DNA? Its middle region is long and alpha-helical; its C-terminal is needed to locate the protein to the nucleus, and associate with nuclear DNA. It is thought that Dsup literally huddles around the tardigrade's DNA thus creating a sort of protective shield against radiation. In most instances, when a protein clings onto DNA there is a great chance that it will inhibit, or at least interfere with, DNA replication and transcription. But this doesn't seem to be the case with Dsup. It could be that the amino-terminal and middle regions of Dsup both have roles in somehow relieving the adverse effects that are expected when Dsup associates with DNA.

Unique creatures such as tardigrades, which live under unique circumstances, are bound to develop unique ways to survive, and hence acquire molecules and pathways that are also unique to them. This makes them a potential source of genes and mechanisms involved in protection that could be of interest to scientists. Compared with control cells, Dsup-expressing human cultured cells saw a reduction of up to 50% in DNA damage caused by X-rays. Transferring Dsup into genetically-engineered animals should therefore increase resistance to radiation damage. By extrapolation, patients undergoing radiotherapy during cancer treatment or indeed people who work in radiation-rich environments could benefit from the doings of a protein such as Dsup – though it is always far more tricky to know in advance how a whole organism will react. What is more, there are undoubtedly other systems in R.*varierornatus* that contribute to its radiotolerance – one of which could be a DNA repair system for instance. Needless to say, astrobiologists also have a keen interest in tardigrades, and have had so since the early 1960s. If these invertebrates can put up with such hostile conditions, could they not survive in outer space? And if so, would they not give us clues as to what may already be living out there? An intriguing thought.

Cross-references to UniProt

Damage suppressor protein Dsup, *Ramazzottius varieornatus* (Tardigrade) : P0DOW4

References

1. Chiaruttini N., R Hashimoto T., Horikawa D.D., Saito Y., Kuwahara H., Kozuka-Hata H., Shin-I T., Minakuchi Y., Ohishi D., Motoyama A., Aizu T., Enomoto A., Kondo K., Tanaka S., Hara Y., Koshikawa S., Sagara H., Miura T., Yokobori S.-I., Miyagawa K., Suzuki Y., Kubo T., Oyama M., Kohara Y., Fujiyama A., Arakawa K., Katayama T., Toyoda A., Kunieda T.
Extremotolerant tardigrade genome and improved radiotolerance of human cultured cells by tardigrade-unique protein
Nature Communications DOI: 10.1038/ncomms12808
PMID: 27649274

proteinspotlight
> ONE MONTH, ONE PROTEIN <
Protein Spotlight (ISSN 1424-4721) is a monthly review written by the **Swiss-Prot** team of the **SIB Swiss Institute of Bioinformatics**. Spotlight articles descrive a specific protein or family of proteins on an informal tone.
http://web.expasy.org/spotlight
SIB Swiss Institute of Bioinformatics

EMBnet.journal
Bioinformatics in Action

ISSN 1023-4144

DEAR READER,

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

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

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

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

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