EMBnet.journal

Volume 21
2015

- Longevity of Biological Databases
- CexoR: an R/Bioconductor package
- 2015 Annual General Meeting

and more...

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 two interesting technical notes, and a report on the last EMBnet Annual General Meeting (AGM) and workshop, hosted in Oeiras (PT), from 10-12 June 2015. At this meeting, EMBnet delineated and agreed a new investment strategy, to be realised through a series of initiatives, including new projects to restyle *EMBnet.journal* and to support GOBLET in the development of high-quality training materials. The research article by Attwood *et al.* deserves special mention, as it touches on one of the most crucial issues of modern science-life research – that of the fragility of the databases in which we 'preserve' the fruits of our scientific endeavours, and the threat their ephemeral nature poses to data diversity and scientific reproducibility (a challenge at the heart of major infrastructure initiatives like ELIXIR and BD2K).

Volume 21 also produced a Supplement disseminating the results of the last conference organised by the SeqAhead COST Action, dedicated to issues related to Next Generation Sequencing (NGS) data analysis and storage.

This closing editorial of volume 21 looks forward to a future new-look volume 22, with a fully revised, enhanced layout. The editorial and technical teams behind the Journal have worked hard both on the layout design and on establishing new administrative routines, and will soon launch a PR campaign to expose more authors and readers to the ameliorated journal.

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

Contents

2015 Annual General Meeting – Executive Board Report




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

[1]University of Manchester, Manchester, United Kingdom
[2]International Institute of Tropical Agriculture, Ibadan and CNR, Institute for Biomedical Technologies, Bari, Italy
[3]Kenya Medical Research Institute (KEMRI), Kenya
[4]Swedish University of Agricultural Sciences, Uppsala, Sweden

Received 9 November 2015; **Published** 21 December 2015

Attwood TK *et al.* (2015) *EMBnet.journal* **21**, e855. http://dx.doi.org/10.14806/ej.21.0.855

During the last year, the Executive Board (EB) endeavoured to meet on a regular basis, and to hold meetings with the Operational Board (OB); fewer meetings were held with the full EMBnet constituency. Aside from the usual issues involved in organising large meetings online (regardless of the chosen technology), the principal obstacle to holding more frequent meetings related to personnel: above all, i) Andreas Gisel (EB Secretary) was seconded to a post in Nigeria, where his Internet connection proved to be extremely unreliable, and maintaining contact with him was often very difficult; and ii) Goran Neshich stepped down as a member of the EB on 15 December 2014. Despite these hurdles, it's been another busy year, and we've worked hard to try to keep the momentum going. This report gives a brief overview of our efforts to move EMBnet, and our affiliated projects and initiatives, forward.

Specifically, since the 2014 Annual General Meeting (AGM) in Lyon (26-30 May 2014), working closely with our affiliates, members of the EB participated in, and/or helped to organise, a range of research and educational meetings, including conferences, hackathons, workshops, courses and tutorials – some of these are summarised in the table below.

In addition to this full programme of meetings, the EB worked closely with the Publicity & Public Relations Project Committee (P&PR PC) to sponsor relevant conferences: in particular, the SAGS-SASBi Joint Congress, Kwalata Game Ranch (ZA)[1], 23-26 September 2014, and the Joint NETTAB 2014 Workshop, Turin (IT)[2], 15-17 October 2014. We also worked with the P&PR PC both to create and disseminate the monthly *EMBnet.digest*[3], and to develop and publish four new *QuickGuides*[4].

EMBnet's training strategy continued to focus on our leadership of the Global Organisation for Bioinformatics Learning, Education and Training (GOBLET)[5], which now has around 40 organisational and individual members. Interaction and cooperation with a range of major international societies and networks has been facilitated through GOBLET, significantly increasing EMBnet's visibility. Amongst others, notable achievements in the last year include publications in *Bioinformatics*[6], *EMBnet.journal*[7] and *PLoS CB*[8]; development of a joint training strategy[9] with ELIXIR; running education and training workshops in Manchester, Boston and Toronto; launching an open, global survey of bioinformatics training needs; and working with the ISCB to launch the Computational Biology Education (CoBE) Community of Special Interest (COSI)[10], to harmonise the ISCB and GOBLET training communities. EMBnet can be rightly proud to have spearheaded this highly successful initiative – for further details, please refer to the May 2015 edition of *EMBnet.digest*[11].

1 www.sasbi-sags2014.org.za/
2 www.igst.it/nettab/2014/
3 www.embnet.org/embnet-digest
4 www.embnet.org/embnet-quickguides
5 www.mygoblet.org/
6 www.ncbi.nlm.nih.gov/pubmed/25189782
7 journal.embnet.org/index.php/embnetjournal/article/view/751/1092
8 journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004143
9 https://www.elixir-europe.org/news/elixir-and-goblet-publish-joint-training-strategy
10 cosi.iscb.org/wiki/CoBE:Home
11 www.embnet.org/sites/default/files/digest/EMBnet.digest-05-May.205.pdf

Sponsor / Project	Title	Format	Location	Date
AllBio	Broadening the Bioinformatics Infrastructure to Unicellular, Animal & Plant Science	Congress	Florence (IT)	11-13 Jun 2014
AllBio/SeqAhead	Epigenetics, coding & non-coding RNA	Workshop	Bari (IT),	25-26 Jun 2014
AllBio/SeqAhead	RNA-seq data analysis	Tutorial	Bari (IT)	27 Jun 2014
GOBLET/SEB	Bioinformatics	Workshop	Manchester (UK)	4 Jul 2014
AllBio/ISBE	Managing Big Data	Workshop	Berlin (DE)	9-11 Jul 2014
ISCB/ISMB/GOBLET	The online world of bioinformatics education	Workshop	Boston (USA)	11-15 Jul 2014
AllBio	Open Science & Reproducibility	Workshop	Norwich (UK)	15-17 Sep 2014
AllBio	Final Consortium Meeting	Meeting	Amsterdam (NL)	29-30 Sep 2014
ELIXIR/BioMed-Bridges	A Common Vocabulary to Classify Resources in the Life Science Domain	Workshop	Brussels (BE)	16 Oct 2014
GOBLET	Train the High-School Teacher Outreach	Workshop	Toronto (CA)	14 Nov 2014
GOBLET	3rd AGM	Business Meeting	Toronto (CA)	15-16 Nov 2014
SeqAhead	e-Infrastructures for Massively Parallel Sequencing	Workshop	Uppsala (SE)	19-20 Jan 2015
SeqAhead	Next Generation Sequencing: a look into the future	Conference	Bratislava (SK)	16-17 Mar 2015
ELIXIR	AllHands	Conference	Hinxton (UK)	30 Mar-1 Apr 2015

Throughout the year, we have described these and our other activities in EMBnet.digest and EMBnet.journal. For example, we made a special report on EMBnet's 26th AGM in the May 2014 digest[12]; we reviewed highlights of ISMB 2014, especially the creation of the ISCB/GOBLET CoBE COSI, in the July issue[13]; we provided a round-up of the year's activities in December's digest[14]; we said farewell to AllBio in the February issue[15]; and, as mentioned above, we reviewed some of GOBLET's major achievements since its establishment in November 2012 in May's digest.

For the Journal, having moved to an instant-access model, articles are now published immediately on completion of peer-review and layout, and are collated into volumes only once a year. During 2014, the bulk of the journal work involved preparation of volume 20 (which included reports on GOBLET, AllBio and EMBnet's 2014 AGM), and volume 20, supplement A, containing proceedings of the *NGS Data After the Gold Rush* conference, held 6-8 May in Norwich (UK). Volume 21 is now open for submissions, and kicks off with our article reviewing the Longevity of Biological Databases[16].

One of the biggest challenges for the EB this year was the loss of our primary contact with the local organising committee for the planned AGM in Serbia, following Goran Neshich's decision to step down from the EB in December. We therefore held an extraordinary meeting of the OB, 18-20 February 2015, in Amsterdam (NL). Here, in an attempt to invigorate EMBnet, we formulated a new investment strategy; this will be presented at the forthcoming AGM, which Pedro Fernandes kindly volunteered to host in Oeiras (PT). The AGM will provide an opportunity for open discussion, both to suggest improvements and/or extensions to the strategy, and to commit to implementing it. This year, the terms of three members of the EB will end: Terri Attwood, Andreas Gisel and Etienne de Villiers. However, no candidacies for these positions were received. Our recommendation is therefore that an Interim Board should remain in place to oversee implementation of the rejuvenation strategy, the success of which should be evaluated at the 2016 AGM.

This year, we'd like to give special thanks to Axel Thieffry and Domenica D'Elia, who've worked tirelessly to sustain and coordinate the many activities of the P&PR PC; and Lubos Klucar for his proficient work in managing the production of EMBnet.journal. As always, there's still a lot to do. We therefore encourage you all to engage with, and contribute to, EMBnet, to ensure that EMBnet can continue to live up to its name: the Global Bioinformatics Network.

Chair: T.K. Attwood
Secretary: A. Gisel; **Treasurer:** E. de Villiers;
Member: E. Bongcam-Rudloff
3 June, 2015

12 www.embnet.org/sites/default/files/digest/EMBnetDigest_2014-05.pdf
13 www.embnet.org/sites/default/files/digest/EMBnetDigest_2014-07 (1)_0.pdf
14 www.embnet.org/sites/default/files/digest/EMBnet.digest-Dec-2014.pdf
15 www.embnet.org/sites/default/files/digest/EMBnet.digest-02-Feb.2015.pdf
16 journal.embnet.org/index.php/embnetjournal/article/view/803/1209

2015 EMBnet Annual General Meeting – Publicity and Public Relations Project Committee report




Domenica D'Elia[1]✉, Lubos Klucar[2], Rafael Jimenez[3], Axel Thieffry[4]

[1]CNR, Institute for Biomedical Technologies, Bari, Italy
[2]Institute of Molecular Biology, Slovak Academy of Sciences, Slovakia
[3]ELIXIR Chief Technical Officer – ELIXIR directorate, Hinxton, Cambridge, United Kingdom
[4]Section for Computational and RNA Biology & Biotech Research and Innovation Center (BRIC), Department of Biology, University of Copenhagen, Denmark

Received 1 December 2015; **Published** 21 December 2015

D'Elia D *et al.* (2015) *EMBnet.journal* **21**, e857. http://dx.doi.org/10.14806/ej.21.0.857

This report provides a short summary of the work and achievements of the EMBnet Publicity and Public Relations Project Committee (P&PR PC) from June 2014 to May 2015. In particular, we give a brief overview of our activities, whose priorities were agreed according to the needs and requests both of the Executive Board (EB) and of the EMBnet community.

The main focal points were: i) management of the EMBnet website and content moderation; ii) coordination and monthly drafting, release and advertisement of *EMBnet.digest*; iii) assistance in the production and release of new EMBnet *QuickGuides*; iv) support in the production and advertisement of *EMBnet.journal*; v) management of relationships within EMBnet's communities and related networks/societies; vi) support in the organisation of the 2015 EMBnet AGM and of a series of workshops and tutorial events, mostly relating to EMBnet members' involvement in affiliated projects.

We also managed EMBnet sponsorships of large conferences, such as the SAGS-SASBi Joint Congress[1], held on 23-26 September 2014 in Kwalata Game Ranch (ZA), and the Joint NETTAB-IB 2014 Workshop[2], Turin (IT), 15-17 October 2014. The P&PR PC produced for the SAGS-SASBi Joint Congress a new EMBnet presentation, given by Alan Christoffels, and new promotional material that was included in the conference pack. Judit Kumuthini was responsible for the EMBnet exhibition desk. EMBnet received acknowledgement, had its logo on all printed and digital conference material and in the SAGS/SASBi website. As for the NETTAB-IB workshop, the Chair of the P&PR PC participated in the workshop, presented the EMBnet poster, and gave a 5-minute presentation of EMBnet and its membership benefits. The EMBnet leaflet and a leaflet on EMBnet membership were distributed to participants (about 130 researchers from Italy and many other European countries). A snapshot of the new EMBnet presentation is shown in Figure 1.

The P&PR PC had regular meetings, although their frequency was lowered at the beginning of the New Year, owing to some drastic changes, such as the sudden stalling of our plans to organise the 2015 AGM in Serbia, following the resignation of a member of the EB. These events partially froze new initiatives that the P&PR PC had planned for the year, but did not prevent us from accomplishing our normal duties and tasks. This year, the PR&PR PC has supported the EB and the EMBnet community with the release of 12 *EMBnet.digests*[3], the publication of four new *QuickGuides*[4] and of *EMBnet.journal* Vol. 20 (2014 release)[5] and the journal Supplement A[6], on the conference *"NGS Data after the Gold Rush – COST Action BM1006"*, held in Norwich (UK), 6-8 May 2014.

We assisted EMBnet members with any type of support requested; managed and answered contacts' requests posted in our website; informed the community about job opportuni-

1 www.sasbi-sags2014.org.za/
2 www.igst.it/nettab/2014/
3 www.embnet.org/embnet-digest
4 www.embnet.org/embnet-quickguides
5 journal.embnet.org/index.php/embnetjournal/issue/view/77
6 journal.embnet.org/index.php/embnetjournal/issue/view/79/showToc

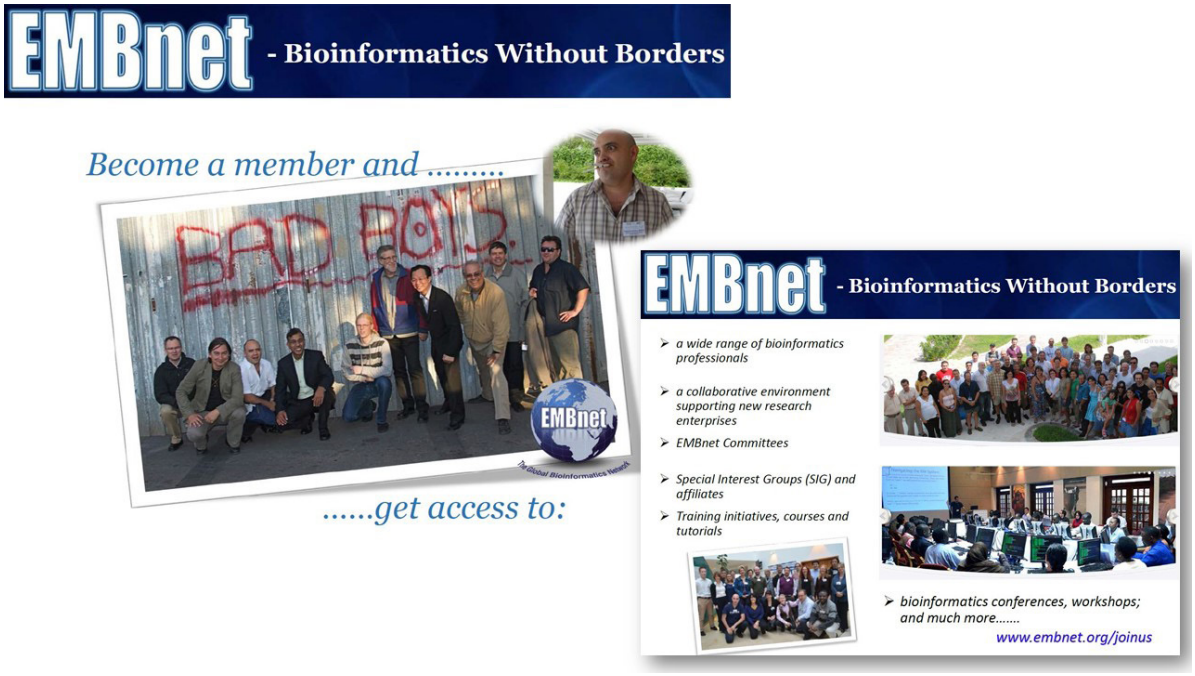


Figure 1. Snapshot (two slides) of the EMBnet presentation, re-designed in 2014 by the P&PR PC.

ties, *EMBnet.digest* and *EMBnet.journal* releases in a timely fashion, and advertised them on the EMBnet website[7] and in the LinkedIn EMBnet Group[8].

The Chair of the PC attended all Operational Board (OB) and EMBnet virtual general meetings, and collaborated in the activities of the OB in order to support the EB in guiding EMBnet toward the objectives agreed with the Board at the 2014 AGM, hosted in Lyon (FR), 26-30 May (Attwood, 2014).

A complete and exhaustive summary of EMBnet's achievements from June 2014 to May 2015, and to which the P&PR PC is proud to have contributed, is provided by the EB in the related article by Attwood *et al.*, "2015 Annual General Meeting – Executive Board Report"[9], in this issue.

Acknowledgements

The Chair of the P&PR PC personally thanks all the Committee's members; without their support, many of the activities and achievements reported above would have been impossible to realise. In particular, special thanks go to Axel Thieffry for his effective and enthusiastic contribution. Finally, and as always, I would like to thank the EB, which has continually supported the P&PR PC with great effectiveness.

References

Attwood TK (2014) EMBnet, the Global Bioinformatics Network: a report on the workshop and 26th AGM, Lyon, May 2014. *EMBnet.journal* **20**, e786. http://dx.doi.org/10.14806/ej.20.0.786

Chair: Domenica D'Elia
Secretary: Lubos Klucar;
Member: Axel Thieffry and Rafael Jimenez
3 June, 2015

7 www.embnet.org/
8 www.linkedin.com/groups/922107
9 journal.embnet.org/index.php/embnetjournal/article/view/855

An Active Investment Strategy for EMBnet - AGM workshop report, Oeiras, June 2015

Teresa K. Attwood

University of Manchester, Manchester, United Kingdom

Received 23 March 2016; **Published** 1 April 2016

Attwood TK (2015) *EMBnet.journal* **21**, e867. http://dx.doi.org/10.14806/ej.21.0.867

Introduction

EMBnet's 2015 AGM and associated activities were hosted in Oeiras (PT), from 10 to 12 June. Marking the 27th formal meeting of EMBnet, the event was an opportunity to take stock and carefully consider EMBnet's future. Earlier in the year, the Operational Board (OB) had agreed that EMBnet would benefit from active investment in specific projects to galvanise its members and drive the community forward. Accordingly, this year, the AGM activities included a full-day OB meeting as a prelude to a one-day workshop for the new strategy to be discussed and refined, followed by the traditional business meeting, where the investment plans would be presented for formal endorsement by the Board.

The purpose of this report is to set out the background to the workshop, to explain the motivation for creating the new investment strategy, to detail its principal themes, and to summarise the workshop's main conclusions.

Overview of the current status

Motivation for the investment strategy: the need for action

One year ago, we discussed some of the changes that would need to take place at the 2015 AGM. In particular, three members of the Executive Board (EB) would be obliged to step down, and a future new – yet to be identified – leadership team would have to assume responsibility for taking EMBnet forward. Early planning was deemed essential to help manage this major change efficiently and effectively. In consequence, an extraordinary Face-To-Face (F2F) meeting of the OB was arranged to discuss i) the location for the AGM; ii) how to manage the forthcoming leadership changes; and iii) how to re-invigorate EMBnet. Members of the OB were asked to submit written proposals in advance of the meeting, outlining future programmes of work to help address these issues.

During the meeting (which, for convenience, took place from 18-20 February 2015, at Schiphol), Pedro Fernandes volunteered to host the AGM at the Instituto Gulbenkian de Ciência, Oeiras. It was agreed to hold the meeting in June.

In terms of EMBnet's leadership changes, it was agreed that a call for new EB candidates should be made as soon as possible, including a deadline for response. It was recognised, however, that this call might not be successful in eliciting candidates; in the worst-case scenario, it might also prove difficult to find volunteers during the AGM itself. Regardless of the outcome, given the scale of the change (*i.e.*, the simultaneous loss of the Chair, Secretary and Treasurer), the OB recommended that, during the AGM, an Interim Board (IB) should be put in place to mentor any new EB candidates, to oversee implementation of the new investment strategy, and to help evaluate its success at the 2016 AGM.

The final component of this extraordinary OB meeting was a lengthy and vigorous debate around each of the submitted proposals for rejuvenating EMBnet. All agreed that a significant investment should be made to stimulate focused activities at the heart of EMBnet. Two main

Figure 1. Terri Attwood presents the agenda of the EMBnet AGM 2015 workshop.

themes eventually emerged as candidates for funding: a) supporting *EMBnet.journal*; and b) strengthening the link with GOBLET.

After the meeting, a Virtual General Meeting (VGM) was called to share and discuss this outcome; however, aside from the OB itself, only three members of EMBnet attended. The 2015 AGM workshop was therefore an important opportunity for wider discussion, to solicit feedback on each theme of the strategy (including suggestions for improvements, extensions and/or other modifications), and to commit to implementing them.

Introducing theme 1: supporting EMBnet.journal

The leap from *EMBnet.news* to *EMBnet.journal* had been enormous, and managing the change had been challenging. Journal production is almost entirely dependent on Lubos Klucar; some members of the Executive Editorial Board also give support with reviewing, copy-editing, writing editorials, *etc.* The number of articles is also a concern, but until the Journal has acquired an impact factor, attracting articles is likely to remain problematic.

The proposed solution was to hire an Editorial Assistant for one year to help with journal publicity, solicitation of new articles, gaining sponsors, and so on. The idea would be to closely monitor progress, and to evaluate the impact at the next AGM: on this basis, the decision would finally be taken whether to continue producing the journal, or perhaps to revert back to *EMBnet.news* (EMBnet's formerly successful newsletter and progenitor of the journal).

Introducing theme 2: strengthening the link with GOBLET

As part of the 2012 AGM, EMBnet had invited leaders of nine other organisations to a workshop to discuss global challenges in bioinformatics training. Each of these organisations had some sort of Education and Training (E&T) initiative, each with similar aims, and each with the same problem: how to deliver tangible benefits to their communities with limited funds and just a handful of time-pressed volunteers. The outcome of that meeting was an agreement to create a Global Organisation for Bioinformatics Learning, Education and Training – GOBLET.

Six months later, following the model of EMBnet, GOBLET was formally established as a Stichting. The GOBLET Foundation uses a similar fee structure to EMBnet, but with two more tiers: for organisations, the tiers are bronze (€250), silver (€500), gold (€1,000) and platinum (€2,500). EMBnet is a Gold Member.

Three years on, with ~40 organisational and individual members, GOBLET is one of the most significant, most visible and most successful of EMBnet's recent initiatives: it is supported by publications; it is promoted at ISMB; it is recognised worldwide; it is also recognised as a partner of ELIXIR. Having spear-headed the formation of GOBLET, there is a golden opportunity here for EMBnet to make an even stronger statement, to strengthen our commitment to GOBLET.

The proposal was to make a ring-fenced donation to GOBLET in order to hire an assistant for two years, to professionally develop branded training materials, to disseminate these via GOBLET's Training Portal, to advertise them on EMBnet's website, and publish them in *EMBnet.journal*, where appropriate. As with the Editorial Assistant, the idea would be to monitor progress closely, and evaluate the impact of this dedicated role at the next AGM.

The Investment Strategy

Supporting EMBnet.journal

The set-up supporting the operation of *EMBnet.journal* is complicated. The technical infrastructure – the server – is hosted in Sweden; the administration is performed in Slovakia. Day-to-day reviewing, copy-editing, editorial writing, and so on, is done by members of the full Editorial Board, some of whom also endeavour to promote the journal. Overall, this process runs fairly smoothly, if rather slowly at times.

One of the main impediments to the success of the journal is that it lacks an Impact Factor (IF), which inhibits researchers from submitting their articles. To move forward, the journal needs to be indexed in PubMed. Of course, this is not all – all members of EMBnet can help by submitting articles to the journal, and encouraging others to do so. But this is not enough: ideally, for the journal to operate on a more professional footing, it needs the support of an assistant to help with:

- email correspondence;
- announcements via different media;
- advertisement (both acquiring advertisers and promoting the journal); and

- creating Special Editions (inviting guest editors in specific subjects/areas).

It had therefore been proposed to hire an assistant for one year to help with such routine tasks and to implement new promotional strategies, to improve and support the production of the journal: €1,000 per month would be available to support this post. The successful candidate would be expected to work closely with the journal's Executive Editorial Board to prioritise the tasks, to report on a monthly basis to the EB, and to report in person at the AGM, where the impact of the work would be evaluated. Here, measures of success might include:

- tripling the number of peer-reviewed articles published;
- tripling the readership (measured in different ways); and
- attracting new income from advertisements.

It was recognised that €1,000 per month posed a possible threat to the project, as it might be insufficient to attract candidates of a suitable calibre. However, despite the acknowledged risk, the proposal was considered essential to help progress the journal. Overall, then, this component of the investment strategy was supported by all workshop participants.

Strengthening the link with GOBLET

EMBnet was created ~27 years ago as a government-mandated, government-supported infrastructure to provide 'bioinformatics' services to European national communities – in particular, to distribute the EMBL Data Library (this was before the term bioinformatics had been coined in this field, and hence EMBnet was the European Molecular Biology network not the European Bioinformatics network). However, with the advent of easy-to-use browsers, which revolutionised access to the Internet, the world changed, and this once-essential service role of EMBnet was largely superseded. Eventually, EMBL Bank and its associated services found a sustainable home at the European Bioinformatics Institute (EBI), and a much larger initiative was ultimately created, not just to sustain EMBL Bank, but to support all EBI core resources and services – this was ELIXIR, a new, more ambitious, more substantially funded sort of 'EMBnet', based on the principle of nationally-funded Nodes mandated to provide bioinformatics services to their communities.

Despite these radical changes in the European bioinformatics landscape, EMBnet continued to function – often, arguably, without much real sense of purpose. The last major funding that EMBnet received for its activities was via the European project, EMBCORE. However, this was at a time when EMBnet had expanded its borders to include a range of developing countries. These could not benefit from European grants, thereby creating an awkward division in EMBnet's funding stream that was hard to manage. Since then, no funds (from Europe or elsewhere) have been sought to unify all EMBnet members around a common project.

Many former members of EMBnet have since found a new focus in ELIXIR, and several current members now play active roles there. Some of those have been particularly active in the realms of bioinformatics E&T – indeed, it was partly because of this that EMBnet took the lead in creating GOBLET, as this was an opportunity to put EMBnet's E&T activities on the map globally. With this in mind, in 2014, EMBnet's Publicity and Public Relations (P&PR) Project Committee (PC) undertook a survey to try to gain a view of what members felt were EMBnet's core values. The survey identified various strong points: these included *EMBnet.journal*; *EMBnet.digest*; EMBnet QuickGuides; E&T activities (GOBLET); its worldwide network (albeit not including North America or Japan); the network's expertise; and its LinkedIn Group (which provides greater visibility). On the flip side, the survey also highlighted that EMBnet is weak in terms of research and new tool innovation, that it offers few services online (given the number of groups the network encompasses), that there are few interactions between members outside AGMs, and there is no common project to unify members.

Overall, the survey concluded that EMBnet's major core values are in bioinformatics E&T, capacity building and networking. To build on its strengths, it was recommended that modern E&T initiatives should be developed, reinforcing our alliances with other societies and networks that are proficient in this area, and that new initiatives should be launched, including projects to galvanise new activities and recruit younger members. Some of the recommendations were very specific and gave considerable food for thought. In particular, Laurent Falquet suggested that *"we should orient ourselves towards a bioinformat-*

ics education community, by publishing more QuickGuides, online courses, and perhaps focus EMBnet.journal on publishing more education-related articles", that EMBnet "*should drop the old trend of providing cutting-edge services because we are not strong in research anymore*", and finally, that EMBnet should "*become one of the most important GOBLET members, by contributing as much as possible.*" He argued that "*Providing well-maintained services useful for teaching purposes would be a definitive strength (e.g., a Galaxy/Chipster server, with specific small databases and well-documented training exercises.)*"

Shortly after this rather provocative survey outcome, the 2014 AGM took place in Lyon. This event was notable for a number of reasons, not least because EMBnet's Secretary provided a critical review of EMBnet, concluding that active steps need to be taken to help drive EMBnet forward (Attwood TK, 2014).

During the OB meeting, taking into account the survey results, and reflecting on the AGM critique, a strong recommendation was made to review EMBnet's position with respect to E&T. Overall, it was suggested to refocus EMBnet's activities and, as part of a new investment strategy, to make a public move to become a 'life-time' member of GOBLET, making a substantial donation – €25K – to match. Such a move would allow an E&T assistant to be hired for two years, and would have a number of clear benefits. Specifically, it would:

- highlight EMBnet's commitment to E&T;
- provide opportunities to create high-quality, jointly-branded materials (tutorials, exercises, manuals, updated/revitalised/professionalised QuickGuides, *etc.*) for advertisement on EMBnet's website, publication, where appropriate, through *EMBnet.journal* and dissemination via GOBLET's Training Portal;
- help to focus EMBnet's work and demonstrate a clear vision and purpose;
- allow concentration on the core values identified by the survey –bioinformatics E&T, capacity building and networking – but in partnership with a burgeoning and active training organisation;
- allow the EMBnet brand to persist through GOBLET;
- encourage more EMBnet members to engage with GOBLET.

The candidate would be expected to prioritise a set of tasks (working closely with relevant groups in GOBLET and EMBnet), to report monthly to the organisations' respective EBs, and to report in person at their AGMs. As with the Editorial Assistant, however, it was recognised that €1,000 per month might not attract candidates of a suitable calibre. Although acknowledged to be risky, particularly in terms of the size of the investment, the proposal was nevertheless considered strategically important for EMBnet, and highly valuable for GOBLET. Consequently, there was warm support from all workshop participants.

eBioKit Tutorial

Erik presented the latest version of the eBioKit, describing its content, functionality and usefulness, not just for teaching activities but also for research purposes in countries where Internet connections are either very slow or absent. He also described how the Kit is distributed, upgraded and updated, and presented courses and past experiences that have demonstrated its efficacy. Workshop participants were invited to use the eBioKit, to suggest further improvements, and to collaborate on the implementation of additional tools, courses and resources that are not currently present within the system.

Discussion and conclusions

The final session of the workshop had been intended to allow a more general discussion of the investment strategy and to solicit possible improvements, extensions and/or additions to its two main strands. However, as both had been enthusiastically endorsed earlier in the workshop, further discussion was unfruitful.

EMBnet Fellowships

Moving on, the workshop focused on the final component of the overall investment plan. During the 2014 AGM, a new 'Fellowship' scheme had been approved as a possible method for getting new blood into EMBnet. At that time, it had been agreed that two Fellowships, each amounting to €2,000, would be awarded during 2014-2015 – initially, the term was set at 18 months, but this was subsequently revised to 1-year terms. For a variety of reasons, however, the initiative had not progressed. The workshop participants were therefore invited to revisit the Fellowship scheme,

and to suggest possible EMBnet-related projects that could be advertised on the website. Possible activities included:

- development of a QuickGuide app for mobile devices;
- development of a camera-ready template for QuickGuides;
- website development, including new contribute content, RSS feeds, *etc.*;
- development of a Utopia Documents plugin for *EMBnet.journal*;
- development of training materials for the eBioKit; and
- creating a Linux-based eBioKit VM.

It was once more agreed that the Fellowship scheme should proceed.

Re-design of EMBnet.journal

An additional project discussed during this session concerned the need to professionalise the look-and-feel of *EMBnet.journal*. This had originally been considered a potential Fellowship project. However, it was recognised that this work is intimately tied up with the success of strand 1 of the investment strategy – hence, there would be little point hiring an assistant to help invigorate and promote the journal, if the journal's style was not re-worked at the same time. Further, this task could not sensibly be part of the Editorial Assistant's role, because the skill-sets required for these positions are quite different. It was generally felt that this was both important and urgent, and that €2K should be made available specifically to engage an individual to re-design the journal pages. The workshop participants agreed that this was a sensible way forward.

Communications and Community

During the general discussions, a proposal was made to automate the process by which EMBnet collects and disseminates news. While *EMBnet.digest* has attempted to fill the 'news' gap created by the transition of *EMBnet.news* to *EMBnet.journal*, the digest is time-consuming to produce, and selective in content (this is deliberate, as the digest is intended to reflect news from the EMBnet community rather than generic bioinformatics information that pervades the Internet). However, there are tools available that facilitate more automated approaches for scraping and collating news from the Internet – *e.g.*, Scoop.it and Paper.li. Together, such tools could be

Figure 2. Group picture of the EMBnet 2015 AGM workshop participants, Oeiras (PT).

used to provide an alternative, more 'lightweight' method of news generation, creating a kind of 'new look' *EMBnet.digest*. It was agreed that, with Pedro's guidance, Axel Thieffry would look into this more closely.

IT for Life Scientists

George Magklaras mentioned that he had been contacted by members of USENIX (the Advanced Computing Systems Association/ USENIX is a brand name in the field of computing professionals) to organise a conference track dedicated to life-science computing needs, focusing especially on how to train IT professionals to cater for life scientists. It had been proposed to arrange a call with them (including himself, Pedro and Terri Attwood), within the next couple of weeks. George was convinced that organising conferences with other bodies to educate IT professionals about the IT needs of life scientists was valuable for EMBnet to do on behalf of the community, and hence he would follow this up and set a date for the meeting.

Another idea that emerged from this general discussion was that EMBnet could publish screencasts. Given the right subject, these can be very popular, they can help drive traffic to websites, and are relatively straightforward to produce. As a 'proof of concept', George suggested that he could create some short screen-casts, for example, to supplement existing QuickGuides.

Close of meeting

This had been a very constructive and productive workshop. Terri thanked all for their participation, for their suggestions of new activities and for their positive support of the investment proposals, which would be taken forward for endorsement at the business meeting the next day; with that, the meeting closed.

References

Attwood TK (2014) EMBnet, the Global Bioinformatics Network: a report on the workshop and 26th AGM, Lyon, May 2014. *EMBnet.journal* **20**, e786. http://dx.doi.org/10.14806/ej.20.0.786

CexoR: an R/Bioconductor package to uncover high-resolution protein-DNA interactions in ChIP-exo replicates

Pedro Madrigal

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom

Received 26 April 2015; **Accepted** 11 June 2015; **Published** 30 July 2015

Madrigal P (2015) *EMBnet.journal* **21**, e837. http://dx.doi.org/10.14806/ej.21.0.837

Competing Interests: none

Abstract

For its unprecedented level of spatial resolution, chromatin immunoprecipitation combined with λ exonuclease digestion followed by high-throughput sequencing (ChIP-exo) has the potential to replace ChIP-seq as the standard approach for genome-wide mapping of protein-DNA interactions. In this assay, the midpoint between the strand-specific paired peaks, formed in the forward and reverse strands, is typically delimited by the exonuclease stop-sites, within which the protein-binding events are located. Although numerous algorithms have been developed for peak-calling in ChIP-seq data, none of them is fully adjusted for the analysis of ChIP-exo. This is because those statistical models do not make use of ChIP-exo's strand-specificity for the identification of protein-DNA binding sites. Here, we present the CexoR algorithm, which aims to ease the analysis of replicated ChIP-exo data in BAM alignment format. The detection algorithm relies on the Skellam distribution (cross-correlation of two Poisson distributions) to calculate probabilities of consecutive punctate-sources of read-enrichment located nearby at Watson-and-Crick strands. ChIP-exo peak-pairs are identified and ranked by their irreproducible discovery rate estimated across biological replicates, and finally reported in BED format files. CexoR can potentially be applied to other ChIP-exo-based protocols, such as ChIP-nexus.

Availability and implementation: CexoR has been implemented in R, and is freely available at http://bioconductor.org.

Introduction

Precisely mapping protein-DNA binding to genomic sites is a pivotal task in order to better understand gene regulation. Chromatin ImmunoPrecipitation (ChIP) followed by microarray hybridisation (ChIP-chip) or sequencing (ChIP-seq) have been extensively used to create maps of Transcription Factor (TF)-binding sites, comparing ChIP-seq favourably with respect to ChIP-chip in terms of resolution and signal-to-noise ratio (Ho *et al.*, 2011). Although ChIP-seq remains the standard, most-used methodology (Furey, 2012), λ exonuclease digestion followed by high-throughput sequencing (ChIP-exo) has recently emerged as a powerful and promising technique able to substitute ChIP-seq, and to circumvent its limitations (Rhee and Pugh, 2011; Mendenhall and Bernstein, 2012). In this protocol, the distribution of ChIP-exo reads is characterised by pairs of two distinct peaks, one at each DNA strand, centred at the λ exonuclease borders and separated frequently at fixed distances. Importantly, the improved resolution of ChIP-exo can provide new insights into protein-DNA interactions (Rhee and Pugh, 2011; Serandour *et al.*, 2013). Furthermore, ChIP-exo allows distinguishing weaker peaks more confidently, and also closely-located binding events that in ChIP-seq are generally deconvolved through computational approaches (*e.g.*, Guo *et al.* (2012)).

Numerous algorithms enable ChIP-seq peak-finding in biological samples considered individually (Bailey *et al.*, 2013). The peak-calling process involves the detection of single regions of significant tag enrichment. However, as underlined in Guo *et al.* (2012), common ChIP-seq peak-finders may fail to identify ChIP-exo single-base-resolution binding if the model they build is not adjusted to the actual distribution of the reads produced by this sequencing technology.

Notably, the offset of top- and bottom-strand reads observed in ChIP-seq is not present in ChIP-exo, and therefore it is not necessary to estimate insert sizes and adjust the positive- and negative-strand reads accordingly (Serandour *et al.*, 2013). For example, some ChIP-seq peak-callers do not account for strand-specific information, while others just compute strand cross-correlation to estimate the fragment length, afterwards shifting the reads with respect to the other strand (Bailey *et al.*, 2013). Software tools like GeneTrack (Albert *et al.*, 2008), GPS-GEM (Guo *et al.*, 2012), peakzilla (Bardet *et al.*, 2013) and MACS (Feng *et al.*, 2012) have been used for peak-calling in ChIP-exo data-sets. However, GeneTrack was designed with ChIP-chip and ChIP-seq in mind, thus requiring manual matching of ChIP-exo peak-pairs located nearby on opposed DNA strands (Rhee and Pugh, 2011). GEM achieved an impressive performance using positional priors based on sequence information. Nevertheless, the presence of a recognisable motif does not guarantee the true discovery of protein-DNA interactions (Bonocora *et al.*, 2013), and these priors should not be used when this premise is not valid. Therefore, non-canonical sites should not be discarded during peak calling, but after, if required for specific downstream analyses, as they might represent cooperativity of the ChIP-ed TFs with other DNA-binding proteins. Furthermore, unlike ChIP-exo, most ChIP-seq peak-calling tools are based on a comparison between a treatment sample and a negative control (which is not available for most ChIP-exo data-sets). Based on this comparison, some of them are able to provide statistical assessments in the form of *p*-values or False Discovery Rates (FDRs) based on different statistical models. As a consequence, default peak-caller stringency cut-offs can generate unreliable FDR estimations (Li *et al.*, 2011; Bailey *et al.*, 2013). Only GEM, mentioned above, and MACE (Wang *et al.*, 2014) have dedicated functionality for ChIP-exo (Zentner and Henikoff, 2014).

To address these inconveniences and allow **C**hIP-**exo** data analysis in **R**, we have developed the Bioconductor package CexoR, which searches peak boundaries in the forward and reverse strands (peak-pairs) rather than strand-agnostic regions for significant enrichment of a treatment compared to a paired negative control. These boundaries are located at the 5′ ends of the ChIP-exo aligned reads, and indicate the location of the λ exonuclease stop-sites (see graphical abstract Figure in Rhee and Pugh (2011)). CexoR is the first R package focusing exclusively on ChIP-exo peak-pair calling, including assessment of reproducibility between biological replicates, and it works without the presence of a control sample. The Irreproducible Discovery Rate (IDR) (Li *et al.*, 2011) analysis, included in the package, has been extensively used in ChIP-seq and RNA-seq data generated by the ENCODE Project (Landt *et al.*, 2012), and it is a recommended approach during ChIP-seq data analysis (Bailey *et al.*, 2013). The analysis of ChIP-exo data is very straightforward, as it only requires a single execution of the function `cexor`.

Implementation

Statistical model

The workflow of CexoR is illustrated in Figure 1.

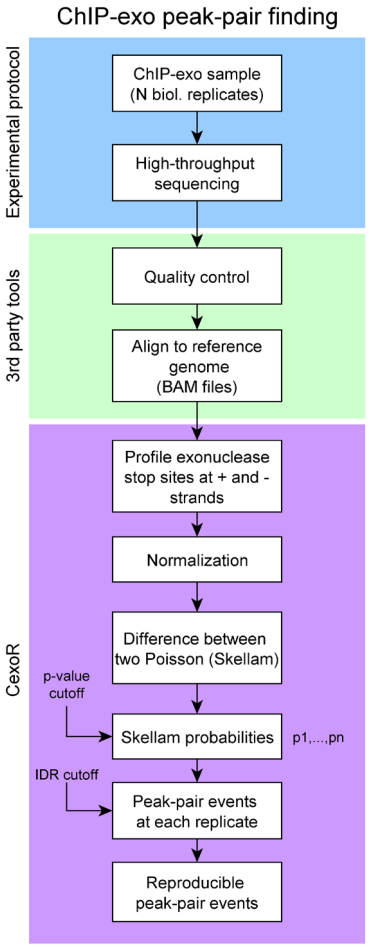


Figure 1. Workflow of ChIP-exo data analysis in R using CexoR.

λ exonuclease stop-site (5'-end of the reads) counts are calculated separately for both DNA strands from the alignment files in BAM format using the Bioconductor package Rsamtools. Counts are then normalised using linear scaling to the sample depth of the smaller data-set. Using the Skellam distribution (Skellam, 1946), CexoR models, at each nucleotide position, the discrete signed difference of two Poisson counts with expected values μ_+ and μ_- in forward and reverse strands. We model the count difference $n_1 - n_2$ at each nucleotide of two statistically independent random variables N_1 (stop-sites in '+' strand) and N_2 (stop-sites in '-' strand), each having Poisson distribution with expected values μ_1 and μ_2. The probability mass function for the Skellam distribution for a count difference $k = n_1 - n_2$ of two Poisson-distributed variables with means μ_1 and μ_2 is given by:

$$f(k;\mu_1,\mu_2) = e^{-(\mu_1+\mu_2)}\left(\frac{\mu_1}{\mu_2}\right)^{k/2} I_k(2\sqrt{\mu_1\mu_2})$$

where k=…,-1,0,1,…, and $I_k(z)$ is the modified Bessel function of the first kind,

$$I_k(z) = \left(\frac{z}{2}\right)^k \sum_{j=0}^{\infty} \frac{\left(\frac{z^2}{4}\right)^j}{j!\Gamma(k+j+1)}$$

where $\Gamma(a)$ is the gamma function. This is done under the assumption that the λ exonuclease digests each DNA strand independently, and that digested DNA sites are random (Rhee and Pugh, 2011). Then, detecting adjacent significant count differences of opposed sign (peak-pairs) in both strands, CexoR delimits the flanks of the protein-binding events at base-pair (bp) resolution (Figure 2). The range of distances allowed between peak-pairs located in opposed strands in a replicate is user settable (parameter `dpeaks`). A one-sided *p*-value is obtained for each peak using the complementary cumulative Skellam distribution function, and a conservative *p*-value for the peak-pair (default cut-off $p \leq$ 1E-12) is reported as the sum of the two *p*-values. Then, peak-pairs across replicates, whose midpoint is located at a user-defined maximum distance (parameter `dpairs`), are selected for further analysis (Figure 2).

It is extremely important to select the parameters `dpeaks` and `dpairs` carefully, for example taking into account the expected length of the footprint of the ChIPed TF, or if the binding events typically cluster nearby along the genome. To account for the reproducibility of signal values of replicated peak-pairs, $\log_{10}$ *p*-values of each replicate are submitted for IDR analysis (Li *et al.*, 2011). Finally, the locations of reproducible binding events formed within peak-pairs are reported, as well as their midpoints. Additionally, Stouffer's and Fisher's combined *p*-values are given for the final peak-pair calls.

Installation

To install CexoR, start R and enter:

```
R> source("http://bioconductor.org/biocLite.R")
R> biocLite("CexoR")
```

Example of use

We downloaded three replicates of human CCCTC-binding factor (CTCF) ChIP-exo data from NCBI Short Read Archive accession number SRA044886 (Rhee and Pugh, 2011), and

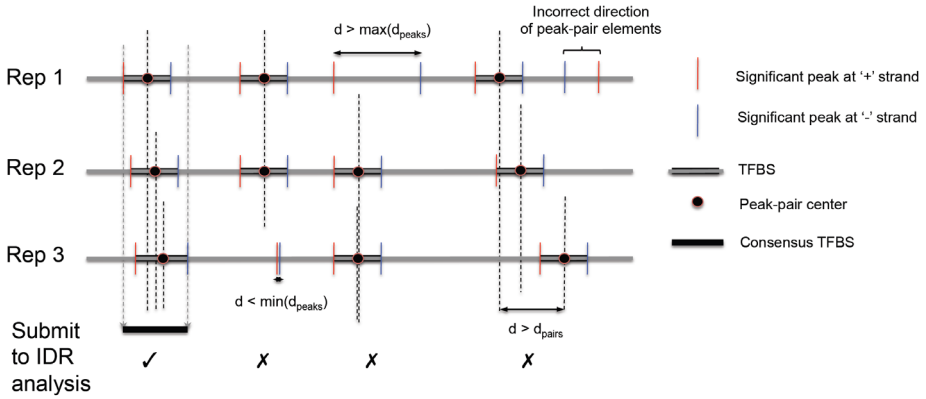


Figure 2. Illustration of the definition of ChIP-exo peak-pairs and overlap criteria between replicates.

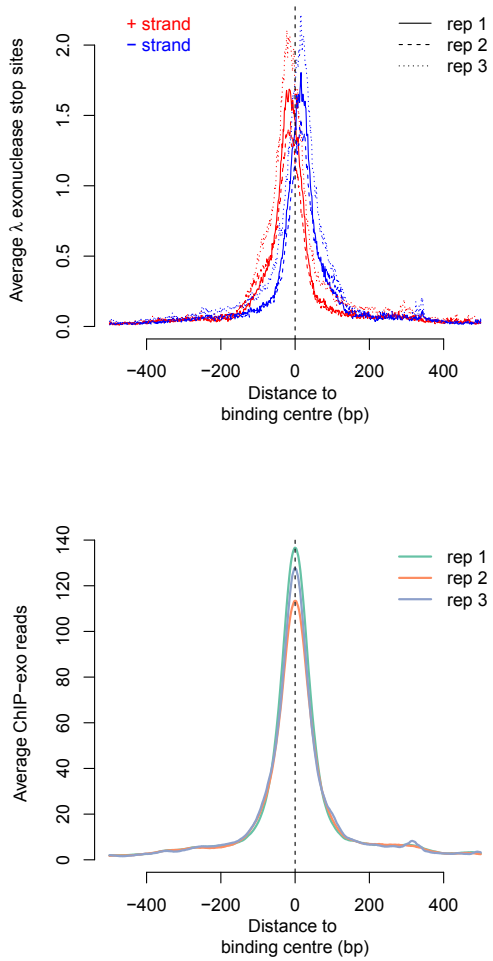


Figure 3. Graphical output of the example code used to identify 2,200 CTCF-binding sites in chromosome 1. (Top) Average λ exonuclease stop-sites. (Bottom) Average ChIP-exo profile of mapped reads. Only the final 2,200 regions are considered in the plots.

aligned the reads to the human reference genome (hg19) using Bowtie 1.0.0 (Langmead *et al.*, 2009). Reads not mapping uniquely were discarded. ChIP-exo data analysis in CexoR is straightforward, as it only requires a single execution of the function `cexor`. For example, to find TF-binding sites in the first chromosome, we run:

```
R> library(CexoR)
R> system("wget http://genome.ucsc.edu/
goldenpath/help/hg19.chrom.sizes")
R> genome <- read.table("hg19.chrom.sizes",
head=F)
R> chipexo <- cexor(bam=c('CTCF_rep1.bam',
'CTCF_rep2.bam', 'CTCF_rep3.bam'), chrN=as.
character(genome$V1[1]), chrL= genome$V2[1],
idr=0.01, p=1e-12, dpeaks=c(5,100),
dpairs=50, bedfile=TRUE)
```

We find >16,000 peak-pairs for each replicate, but only 2,200 reproducible TF-binding events after IDR analysis (p-value<1e-12; IDR<0.01)

```
R> for(i in 1:3){print(length(chipexo$paired
PeaksRepl[[i]]))}
[1] 18624
[1] 16188
[1] 20394
R> length(chipexo$bindingEvents)
[1] 2200
```

We can now plot the mean profile of λ exonuclease stop-sites and reads, 500 bp around the central position of reproducible peak-pair locations, by running the function `plotcexor`

```
R> plotcexor(bam= c('CTCF_rep1.bam', 'CTCF_
rep2.bam', 'CTCF_rep3.bam'), peaks=chipexo,
EXT=500)
```

The output is shown in Figure 3.

These visualisation plots are obtained using the Bioconductor package genomation (Akalin *et al.*, 2015).

Full details and examples are given in the manual and vignette of the package, release version 1.6.[1] and devel version 1.7.2[2].

Conclusions

Here, we present a new software package to analyse ChIP-exo data-sets. This is an alternative to the recently developed model-based analysis of ChIP-exo (MACE) (Wang *et al.*, 2014). The major differences between MACE and CexoR are: i) MACE detects peak-pairs using the Chebyshev inequality for outlier detection, making no assumption about the distribution of the coverage signal, while CexoR considers the cross-correlation of two Poisson distributions at each DNA strand; ii) MACE matches the borders using the Gale-Shapley stable matching algorithm, which performs an optimisation procedure to estimate border pair sizes, while CexoR uses a 'closest principle' to match peak-pairs within an allowed distance between significant peaks located in opposed strands in a replicate; iii) MACE incorporates an optional step of sequence-bias correction, which shows very little improvement when applied; and iv) MACE computes Shannon's entropy before border detection to consolidate a signal across multiple replicates, while CexoR runs IDR analysis across previously detected

1 bioconductor.org/packages/release/bioc/html/CexoR.html
2 bioconductor.org/packages/devel/bioc/html/CexoR.html

peak-pairs whose central positions are located at close distances in the replicates. Paired-end read information is not used by any of the packages.

In summary, the Bioconductor package CexoR is able to locate reproducible protein-DNA interactions in ChIP-exo data-sets with no need for genome sequence information, manual matching of peak-pairs, paired control data (inputs), or downstream assessment of replicate reproducibility. In addition, the R statistical environment allows integration with other pipelines and downstream analyses via other R and Bioconductor packages. We hope that our software tool will speed up the analysis of forthcoming ChIP-exo data-sets.

If the assumptions are valid (imbalance of forward- and reverse-read distribution in peak-pairs at the boundaries of a TF-binding site), the package can also be used with other next-generation sequencing data, such as ChIP-nexus (He *et al.*, 2015). It is important to note that CexoR can only be used with $\geq$ 2 samples. Further validation and benchmarks of advanced peak-detection methods will be necessary in the new generation of protocols profiling TF binding at high resolution.

Key Points

- ChIP-exo data analysis involves more complex bioinformatics than standard ChIP-seq.
- CexoR (ChIP-exo data analysis in R) is a new Bioconductor package able to locate reproducible protein-DNA interactions in ChIP-exo data-sets.
- CexoR is among the first bioinformatics tools allowing peak-pair calling in ChIP-exo, and the algorithm considers a cross-correlation of two Poisson distributions.
- CexoR could potentially be used with other sequencing data-sets, such as ChIP-nexus, and it includes functionality for the visualisation of the results.

Acknowledgements

The author would like to thank the two anonymous reviewers for their helpful comments.

References

Akalin A, Franke V, Vlahoviček K, Mason CE, Schübeler D. Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics* **31**, 1127-1129. http://dx.doi.org/10.1093/bioinformatics/btu775

Albert I, Wachi S, Jiang C, Pugh BF (2008) GeneTrack-a genomic data processing and visualization framework. *Bioinformatics* **24**, 1305–1306. http://dx.doi.org/10.1093/bioinformatics/btn119

Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q *et al.* (2013) Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput Biol* **9**, e1003326. http://dx.doi.org/10.1371/journal.pcbi.1003326

Bardet AF, Steinmann J, Bafna S, Knoblich JA, Zeitlinger J *et al.* (2013) Identification of transcription factor binding sites from ChIP-seq data at high resolution. *Bioinformatics* **29**, 2705–2713. http://dx.doi.org/10.1093/bioinformatics/btt470

Bonocora RP, Fitzgerald DM, Stringer AM, Wade JT (2013) Non-canonical protein-DNA interaction identified by ChIP are not artifacts. *BMC Genomics* **14**, 254. http://dx.doi.org/10.1186/1471-2164-14-254

Feng J, Liu T, Qin B, Zhang Y, Liu XS (2012) Identifying ChIP-seq enrichment using MACS. *Nat Protoc* **7**, 1728–1740. http://dx.doi.org/10.1038/nprot.2012.101

Furey TS (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet* **13**, 840–852. http://dx.doi.org/10.1038/nrg3306

Guo Y, Mahony S, Gifford DK (2012) High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol* **8**, e1002638. http://dx.doi.org/10.1371/journal.pcbi.1002638

He Q, Johnston J, Zeitlinger J (2015) ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature Biotech* **33**, 395-401. http://dx.doi.org/10.1038/nbt.3121

Ho JW, Bishop E, Karchenko PV, Nègre N, White KP *et al.* (2011) ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics* **12**, 134. http://dx.doi.org/10.1186/1471-2164-12-134

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**, 1813-1831. http://dx.doi.org/10.1101/gr.136184.111

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25. http://dx.doi.org/10.1186/gb-2009-10-3-r25

Li Q, Brown JB, Huang H, Bickel PJ (2011) Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* **5**, 1751–1779. http://dx.doi.org/10.1214/11-AOAS466

Mendenhall EM, Bernstein BE (2012) DNA-protein interactions in high definition. *Genome Biol* **13**, 139. http://dx.doi.org/10.1186/gb-2012-13-1-139

Rhee HS, Pugh BF (2011) Comprehensive genome-wide protein-DNA interactions at single-nucleotide resolution. *Cell* **147**, 1408–1419. http://dx.doi.org/10.1016/j.cell.2011.11.013

Serandour AA, Brown GD, Cohen JD, Carroll JS (2013) Development of an Illumina-based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties. *Genome Biol* **14**, R147. http://dx.doi.org/10.1186/gb-2013-14-12-r147

Skellam JG (1946) The frequency distribution of the difference between two Poisson variates belonging to different populations. *J R Stat Soc Ser A* **109**, 296.

Wang L, Chen J, Wang C, Uusküla-Reimand L, Chen K (2014) MACE: model based analysis of ChIP-exo, *Nucleic Acids Res* **42**, e156. http://dx.doi.org/10.1093/nar/gku846

Zentner GE, Henikoff S (2014) High-resolution digital profiling of the epigenome. *Nat Rev Genet* **15**, 814-827. http://dx.doi.org/10.1038/nrg3798

Renewing bioinformatics workflow systems by using a Web 2.0 approach





Roberto Colella[1]✉, Bachir Balech[2], Antonella Vaccina[3], Pietro Leo[3], Gaetano Scioscia[3]

[1]Istituto di Studi sui Sistemi Intelligenti per l'Automazione - Consiglio Nazionale delle Ricerche, Bari, Italy
[2]Istituto di Biomembrane e Bioenergetica - Consiglio Nazionale delle Ricerche, Bari, Italy
[3]GBS BAO Advanced Analytics Services and MBLab, IBM Italia S.p.A., Italy

Received 14 May 2015; **Accepted** 27 July 2015; **Published** 6 August 2015

Colella *et al.* (2015) *EMBnet.journal* **21**, e840. http://dx.doi.org/10.14806/ej.21.0.840

Competing Interests: none

Abstract

The use of "mashups" is expanding considerably in the business environment. Business mashups are usually adopted within integrating business and data-service frameworks to provide the ability to develop new integrated services quickly. Typically, mashups provide organisations with a pronounced and flexible commodity to combine internal with external services in order to create new services, usually accessed through user-friendly Web-browser interfaces. In this study, a Web 2.0 technology was adopted to promote a key field of bioinformatics research through the management and automation of bioinformatics workflows. Consumables (widgets and services) have been developed using the Lotus Widget Factory, an Eclipse plug-in providing an easy-to-use development environment enabling developers of all skill levels to create dynamic widgets rapidly. A workflow built from widgets works as follows: the core widget receives data from one or more widgets, invokes a generic Web service, performing iteration and/or recursion, and sends the results to all other connected widgets. The number of iterations and recursions depends on the input data-set dimension and user-defined parameter values related to each specific application. Some prototype workflows have been assembled and tested with a number of widgets created with algorithms from the European Molecular Biology Open Software Suite (EMBOSS), exposed as Web services. The adoption of recent Web 2.0 technologies, such as mashup platforms, has enabled rapid generation, sharing and discovery of reusable application building-blocks (widgets, feeds, mashups), and has shown to be a plausible alternative environment for supporting bioinformatics workflow design, management and execution.

Introduction

The execution of complex bioinformatics workflows is becoming increasingly important for advanced scientific research, given the huge amount of data output by next-generation sequencing technologies (*e.g.*, Marguiles *et al.*, 2005; Bentley *et al.*, 2008). In this context, analysis workflows are becoming more complex to build, requiring advanced technical skills, which end-users may not have.

Several software and platform products have been proposed to meet the typical needs of bioinformaticians. Examples include the integration of multiple data sources (*e.g.*, data stored on local file systems, query results from public or private databases, feeds), the availability of computational tools necessary to achieve specific research results, and workflow storage for reproducibility. Amongst the most frequently used tools for managing and executing workflows are Taverna (Wolstencroft *et al.*, 2013), Bioextract (Lushbough *et al.*, 2010) and Galaxy (Goecks *et al.*, 2010). The first of these requires installation as a standalone workbench, allowing workflow design and building, plus the ability to execute them on the cloud and share them on a public website[1]. The second is a Web-based application, which does not allow use of available operators to assemble workflows. Bioextract's main functionality is based on recording actions performed by users, and not on "drag and drop" propositions. Galaxy offers efficient online reuse of previously implemented applications, but its difficulty resides in the need for advanced programming skills to build new workflows. The main

1 www.myexperiment.org

goal of this article is to evaluate the usability of a widget-based tool that facilitates and speeds up the development of bioinformatics workflows. To that end, we provide a complete description of this workflow management and storage system, presented as a prototype and tested on two locally assembled bioinformatics workflows in a mashup framework.

The mashup framework

Mashups are applications that integrate information from different data sources into a single new service. Data from different sources need to be represented in such a way that users can understand and analyse them. In enterprise IT management, there is an opportunity to mash up data from various products, keeping intact data behaviour and data flow, to provide new insights (Fichter, 2010).

Mashup techniques have been successfully adopted in several business areas. For instance, Boeing (Ayhan *et al.*, 2009), Wells Fargo, the UK's Kent County, AMEC Paragon and the New York State Department of Labor (Sezici, 2009) are examples of the use of mashups for fast application delivery and improved decision-making. Recently, the mashup approach has also been suggested for use in bioinformatics (Gong, 2013; Hogan *et al.*, 2011; Cheung *et al.*, 2008), but to our knowledge a bio-mashup editor is still lacking. The kind of issues cited above suggested the adoption of IBM Mashup Center, an end-to-end enterprise mashup platform supporting rapid assembly of widgets, which are dynamic miniature Web applications embedded within HTML pages. This tool includes a Mashup Builder, a widget-based browser interface that contains all the necessary components for creating, assembling, configuring and designing objects, such as widgets, mashup pages and spaces. Moreover, it provides a set of out-of-the-box, business-ready widgets, which jump-start mashup creation and enhance information visualisation options, such as charting.

The uniqueness of this system lies in the simplicity of extending the mashup environment by incorporating custom IT widgets from the IBM Mashup Catalog, or widgets from external Web resources, including any of the thousands of Google Gadgets. Furthermore, Mashup Center allows bioinformaticians to work with feeds, which can be mixed and transformed into new feeds, also known as data mashups. Using the Data Mashup builder, a visual browser-based tool, information and business analysts can re-mix, merge, group, sort, annotate, filter and transform feeds in a variety of ways, creating a single view of disparate sets of information in a very short time. Once a mashup is assembled, it can be easily shared and, by means of some embedded visual tools, the workflow owner can define users or groups of users who can view or edit their various pages. Additionally, with just a few clicks, Mashup Center allows users to customise widgets and pages, and then copy-and-paste the scripts behind them into a Web page, all without writing additional code. Mashups can also be published to the *Mashuphub* catalogue, a shared environment where other users can easily reuse them. Figure 1 shows the context diagram of the adopted platform.

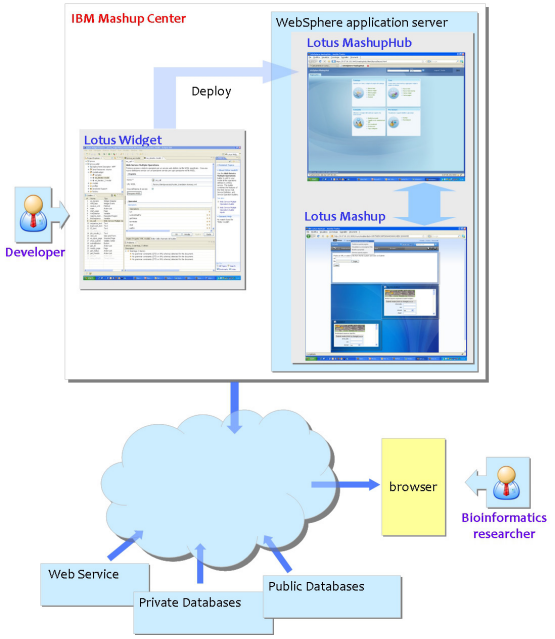


Figure 1. Context diagram of the mashup platform, showing the developer and the bioinformatics researcher interaction with the mashup framework. At the top, the developer deploys the widget(s) on the WebSphere application server to become available to the bioinformatician (bottom) to create a custom workflow.

Implementation

The development of widgets has been carried out using the Lotus Widget Factory[2]. This is an Eclipse plug-in based on the concept of models that a developer assembles from basic bricks

2 https://www.ibm.com/developerworks/lotus/documentation/widgetfactory

called builders. The builders are generic components that encapsulate a given capability. Lotus Widget Factory comes with a large number of predefined builders, ranging from user-interface components, such as buttons, to components responsible for fetching data from remote Web services. A user-friendly wizard interface is associated with each builder, and lets the developer specify its characteristics, such as input data. Once the development has been completed, widgets are deployed as ".WAR" (Web-application ARrchives) in Lotus MashupHub and can be added to a mashup page. Our goal was to obtain detailed insights into the usability of this framework for the assembly, execution and management of bioinformatics workflows. To this end, we implemented separate widgets for some bioinformatics algorithms in order to offer users easy assembly of their own workflows. In addition, we used these widgets to assemble some prototype workflows. In the following paragraphs, we detail the widgets implemented, covering generic and/or specific user-defined requisites for DNA/protein sequence analysis.

Data Source

The Data-Source widget allows selection of an input file from a local file system or a URL invoking a REST service, and then parses the fetched data. The parser can interpret different file formats (EMBL, FastA, *etc.*) to extract all the contained DNA/protein sequences and display them in a tree view. At this step, users can choose which of the sequences will be sent to the next workflow block. Note that the data are converted to FastA format and then arranged in an XML structure, which facilitates communication amongst the consecutive widgets embedded in the workflow.

Merge and Split

The Merge and Split widgets operate on the XML data-flow between widgets. The Merge widget converts input data (sequences, matrices, *etc*...) into a unique output data-flow. It is useful to create a single XML file containing sequences from many files (*i.e.*, coming from a Data-Source widget). In contrast, the Split widget is used to separate the elements of the XML data according to a given regular expression, to facilitate recursive usage of the subsequent workflow widgets.

WSDL-described Web-service widget

The WSDL-described (Web-Service Description Language) Web-service widget is the core application of mashup techniques for bioinformatics workflow building. Its main aim is the execution of an algorithm remotely exposed as a Web service, implemented by means of the following Widget Factory builders:

- *Web service multiple operation and HTML page* builder invoke the Web service, get the available operations and create the user interface;
- *widget event* builder, together with a data-decoding Java method, receive and parse XML data from the previous widget;
- *repeated region* builder iterates over the XML structure and enables recursive invocation varying according to user-defined parameter values;
- *action list* builder executes the "run" Web-service action for all the items found in the XML input data, getting a job ID for each of them;
- *HTML page* builder creates a results page and invokes the "waitfor" and "getResults" Web-service operations (action list);
- *another action list* builder stores the results of the executed jobs in an XML output;
- *widget event* builder sends the XML structure to all the widgets connected to it.

The Web-service widget has been implemented to accept the WSDL file describing EMBOSS[3] (Rice *et al.*, 2000) bioinformatics tools exposed as Web services. With very few customisations, mainly regarding variable names and eventual multiple inputs, it was possible to create ~200 widgets corresponding to applications in the entire EMBOSS suite.

REST Web-service widget

The widget executes a REST service call, stores the results in the XML format output (action list builder), and sends it to all the widgets connected to it (*widget event* builder).

Recursion widget

The Recursion widget can be wired to a Web-service widget, and can collect all the parameters from it. This widget subsequently displays a menu with all the relevant application-specific parameters, allowing users to set their corresponding values for execution during the recursion.

3 emboss.sourceforge.net

Weblogo widget

Weblogo[4] is a Web application that can be used if a graphical representation is needed to summarise one or more sequence alignments obtained by a given algorithm. The application can be installed locally or exposed as a REST service.

In summary, the result of creating the widgets described above is that users can assemble their own workflows by choosing widgets from a drop-down menu and dragging them onto the application page of the mashup editor and connecting them. They can also choose which workflow steps are to be executed automatically, simply by checking a box on each user interface. Another important aspect of this system is the ability to inspect intermediate results, as each widget included in the workflow shows the results it has produced. This can be useful for trouble-shooting and further adapting bioinformatics workflows.

Results and Discussion

The main result of the solution described here is the availability of a prototype workbench system to develop and build either classical analysis workflows or more complex ones. Our experience in building this prototype has shown that bioinformatics researchers can easily design and develop their own workflows and application pages using different tools and data sources. Apart from the existing default widgets, including those mentioned above, a palette of widgets providing the EMBOSS suite applications has been added. In addition, the system flexibility allows advanced users to add new applications, and therefore create new widgets. To validate the functionality of this system, two workflow case-studies are presented in the following paragraphs: i) a phylogenetic inference workflow, and ii) a universal primer-design and validation workflow.

Phylogenetic inference workflow

Our first example of a workflow assembled using the Mashup Center is phylogenetic inference using the neighbour-joining (Saitou and Nei, 1987; St John *et al.*, 2003) or UPGMA (Reguant and Bordons, 2003) methods, commonly used in molecular-evolution studies. The workflow constructs a consensus phylogenetic tree (Figure 2 shows the first steps of the workflow), starting from a set of DNA sequences, and assigns a

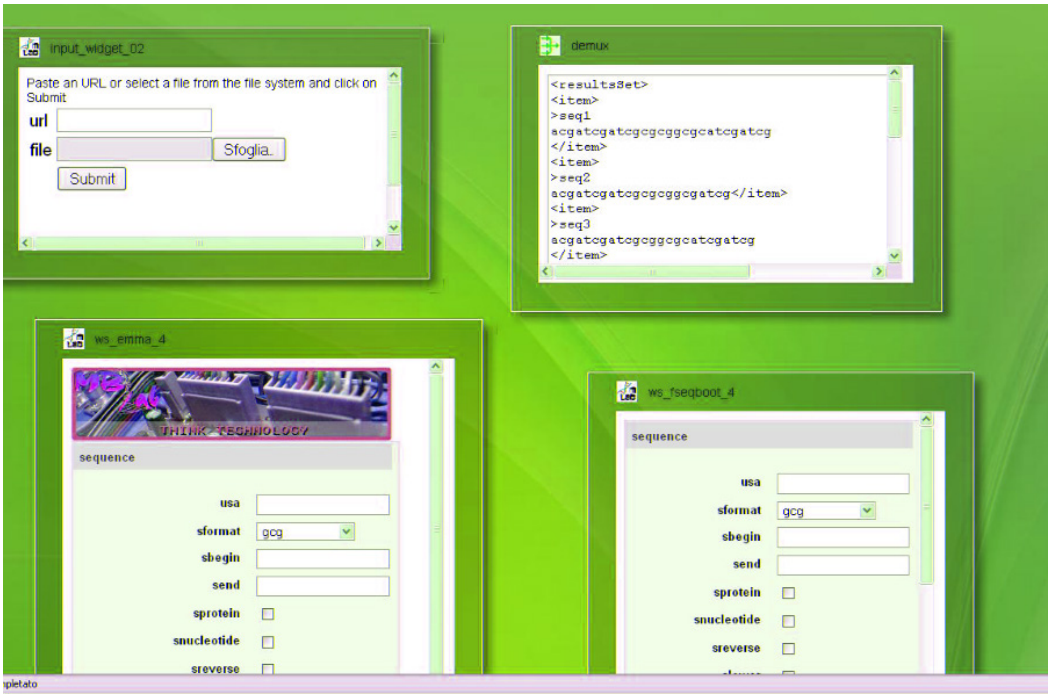


Figure 2. Partial representation of the phylogenetic inference workflow in the mashup editor. Input and split widgets are shown at the top, while the bottom ones correspond to emma and eseqboot, the first two steps of the workflow.

4 weblogo.berkeley.edu

bootstrap value to each node of the tree. It has been tested on a data-set comprising 600 DNA sequences (600 bp long) of the cytochrome oxidase subunit-one (COI) mitochondrial gene (Janzen *et al.*, 2005) belonging to organisms of the *Hesperiidae* family. Our workflow comprises a Data-Source widget and five WSDL Web-service widgets, each invoking one EMBOSS application:

- **emma** executes a multiple alignment across DNA sequences provided in FastA format;
- **eseqboot** generates multiple data-sets (alignments), which are resampled versions of the input data-set, necessary to compute the statistical significance of the final output phylogenetic tree;
- **ednadist** computes the distance matrix corresponding to the input alignment;
- **eneighbor** estimates phylogenies from distance-matrix data using the neighbour-joining or the UPGMA clustering methods;
- **econsense** returns the consensus phylogenetic tree.

Universal primer-design and validation workflow

In order to implement and accomplish the universal primer-design workflow (in Figure 3), we combined several bioinformatics tools able, on the one hand, to design universal primer-sets based on multiple DNA sequence data and, on the other, to validate the primer pairs obtained on the starting data-set. Primer universality is a crucial step in environmental sequencing studies, as the maximum number of organisms is targeted during PCR enrichment prior to sequencing. The workflow has been tested on a data-set of 64 DNA sequences, corresponding to the gene *ITS-1* of *Pucciniastraceae*, extracted from ITSoneDB (Santamaria *et al.*, 2012). A detailed description of the workflow steps is provided below:

- **emma** aligns the initial DNA sequence data-set;
- **cons** defines a consensus sequence corresponding to the multiple alignment;
- **einverted** controls inverted repeats on the consensus sequence;
- **extractseq** extracts a new consensus sequence free from repeated patterns, and keeps its length intact;
- **eprimer3** performs primer design, taking the newly obtained consensus sequence as template, and outputs a number of primer pairs having different characteristics (*e.g.*, GC content, linguistic complexity, PCR product length, *etc.*). At this step, users can choose the best primer pairs that fit their experiment;
- a final universality validation step is performed on the initial data-set by *in silico* PCR using the **primersearch** program. It is important to note that, in this last step, the mis-match percentage value can be changed according to experimental needs.

IBM Mashup Center is a flexible platform that can readily resolve bioinformatics issues. It can be seen as a collection of different tools and sources expressed as Java code, Web services, databases, Web applications and portals to fulfil the typical needs of bioinformatics researchers. The main benefit of the proposed platform

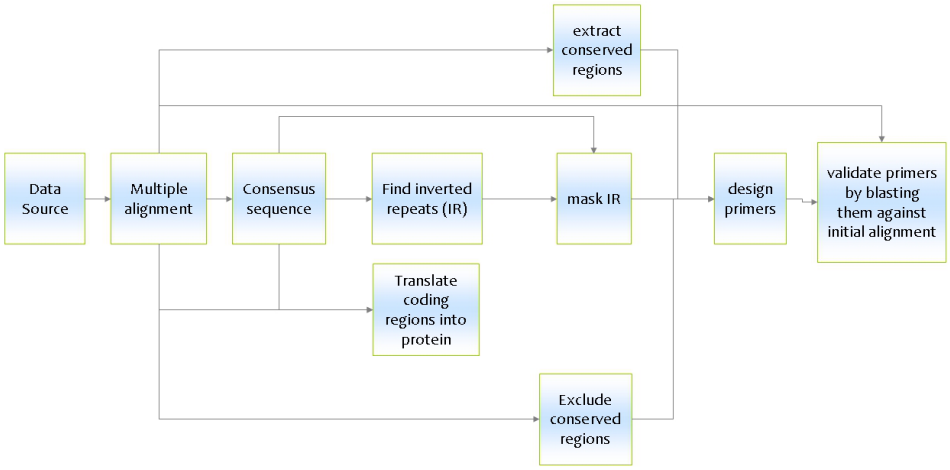


Figure 3. Schematic representation of the Primer Design Workflow, illustrating the basic actions computed by the workflow.

is its user-friendly interface to rapidly assemble tools and sources into a single workflow, and as an interface to different features provided by the MashupHub catalogue. Currently ongoing enhancements include optimisation of the implemented widgets by improving their performance, and the development of new widgets. In addition, the user interface will provide, in future, the possibility of easily creating user-defined Web services. This would facilitate the assembly of complex workflows completely tailored to users' needs.

Availability and requirements

The system was tested locally and is currently still a prototype. It will be released with its complete documentation and requirements once the above-mentioned optimisations have been achieved.

Key Points

- Bioinformatics workflows are built from different tools, each executing their own bio-computational tasks, working together in a standardised manner.
- Bioinformatics widgets are core dynamic elements of graphical user interfaces that contain embedded bioinformatics applications.
- Mashups are applications that integrate information from different data sources into a single new service.
- Bioinformatics widgets can be connected within a mashup framework to form a bioinformatics workflow.

Acknowledgements

The authors acknowledge support of the Ministero dell'Università e della Ricerca (MIUR), under the project DM19410 "Laboratorio di Bioinformatica per la Biodiversità Molecolare".

References

Ayhan S, Comitz P, Stemkovski V (2009) "Aviation Mashups" Digital Avionics Systems Conference. DASC '09. IEEE/AIAA 28th, 6.D.5-1, 6.D.5-9. http://dx.doi.org/10.1109/DASC.2009.5347436

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**(7218), 53-59. http://dx.doi.org/10.1038/nature07517

Cheung KH, Yip KY, Townsend JP, Scotch M (2008) HCLS 2.0/3.0: Health care and life sciences data mashup using Web 2.0/3.0. *J Biomed Inform*, **41**(5), 694-705. http://dx.doi.org/10.1016/j.jbi.2008.04.001

Fichter D (2009) "What is a Mashup." In: Engard N (Ed.) *Library Mashups. Exploring new ways to deliver library data*. Medford, N.J: Information Today, Inc.

Goecks J, Nekrutenko A, Taylor J and The Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*. **11**(8), R86. http://dx.doi.org/10.1186/gb-2010-11-8-r86

Gong P. (2013). Dynamic integration of biological data sources using the data concierge. *Health Inf Sci Syst*, **1**(1), 1-19. http://dx.doi.org/10.1186/2047-2501-1-7

Hogan JM, Sumitomo J, Roe P, Newell F (2011). Biomashups: the new world of exploratory bioinformatics? *Concurr Comput*, **23**(11), 1169-1178. http://dx.doi.org/10.1109/eScience.2008.92

Janzen DH, Hajibabaei M, Burns JM, Hallwachs W, Remigio E *et al.* (2005) Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. *Philos Trans R Soc Lond B Biol Sci* **360**(1462),1835-1845. http://dx.doi.org/10.1098/rstb.2005.1715

Lushbough C, Bergman MK, Lawrence CJ, Jennewein D, Brendel V (2010) BioExtract server--an integrated workflow-enabling system to access and analyze heterogeneous, distributed biomolecular data. *IEEE/ACM Trans Comput Biol Bioinform* **7**(1), 12-24. http://dx.doi.org/10.1109/TCBB.2008.98

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**(7057), 376-380. http://dx.doi.org/10.1038/nature03959

Reguant C, Bordons A (2003) Typification of Oenococcus oeni strains by multiplex RAPD-PCR and study of population dynamics during malolactic fermentation. *J Appl Microbiol* **95**(2), 344-353. http://dx.doi.org/10.1046/j.1365-2672.2003.01985.x

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16(6), 276-277. http://dx.doi.org/10.1016/S0168-9525(00)02024-2

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**(4), 406-425.

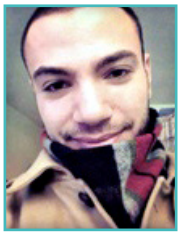
Sezici E (2009) New IBM Mashup Capabilities Bring Business Analytics to the Desktop. SYS-CON Media. http://sap.sys-con.com/node/1160750 (accessed 7 May 2015).

Santamaria M, Fosso B, Consiglio A, De Caro G, Grillo G et al (2012) Reference databases for taxonomic assignment in metagenomics. *Brief Bioinform* **13**(6):682-695. http://dx.doi.org/10.1093/bib/bbs036

St John K, Warnow T, Moret BME, Vawter L (2003) Performance study of phylogenetic methods: (unweighted) quartet methods and neighbor-joining. *J Algorithm* **48**(1), 173-193. http://dx.doi.org/10.1016/S0196-6774(03)00049-X

Wolstencroft K, Haines R, Fellows D, Williams A, Withers D *et al.* (2013) The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res* **41**(Web Server issue), W557-561. http://dx.doi.org/10.1093/nar/gkt328

Longevity of Biological Databases


Teresa K. Attwood[1]✉, **Bora Agit**[1], **Lynda B.M. Ellis**[2]

[1]University of Manchester, Manchester, United Kingdom
[2]University of Minnesota, Minnesota, United States

Received 23 January 2015; **Accepted** 13 March 2015; **Published** 4 May 2015

Attwood TK *et al.* (2015) *EMBnet.journal* **21**, e803. http://dx.doi.org/10.14806/ej.21.0.803

Competing interests: TKA currently serves on the editorial board of EMBnet.journal; BA none; LBME none

Abstract

Public Web-based databases are essential for present-day biological research: they i) store the results of past laboratory experiments; ii) guide the focus of future ones; and, iii) allow all to benefit from the wealth of information they contain. Many new databases are born each year; but how long do they live? This study looked at the 18-year survival of 326 databases. Over 60% were dead within that time period, and a further 14% were archived, no longer updated. Those that survived were, for the most part, important to their institution's main focus, and had core institutional support. Database longevity depends on the existence of infrastructures that are underpinned by long-term financial strategies. Researchers and funders need to consider the ramifications for the security of their data, and of the financial investments in them, if they choose to create new databases independently of core infrastructures.

Introduction

During the last 30 years, since the first public release of resources like the EMBL Data Library (Hamm and Cameron, 1986) and GenBank (Burks *et al.*, 1985), databases have become an indispensable part of the tool-kit of modern biological research: we depend on them to store experimental data of all kinds, to inform our research, and to share the fruits of our collective knowledge with the scientific community. Back in the 1980s, when the field of bioinformatics was just emerging, there was an unwritten rule that biological databases (and their associated analysis software) should be made freely available. In consequence, they became a side-effect of research projects and, each year, many new databases were born and distributed to a voracious community. Indeed, they became such a familiar part of the research landscape that an entire issue of a prestigious journal (*Nucleic Acids Research*) was formed to alert the community to updates and modifications to existing resources and to the appearance of new ones, and a Web-based database was created to catalogue them – DBcat (Discala *et al.*, 1999).

Superficially, this is a success story – life scientists took little persuading that their data benefitted from proper management and analysis. However, no overarching financial strategy underpinned this database revolution – once created, therefore, many struggled to survive. So the question is, how long do they live in reality? In 1998, Ellis and Kalumbi surveyed maintainers of public biological databases listed in DBcat. This survey found that more than two-thirds (68%) of the 153 databases for which information was received (48% response rate) had uncertain near futures (1-5 year funding) (Ellis and Kalumbi, 1998). We, and others, have commented on this shaky future, arguing that a viable, sustainable framework for long-term data stewardship is sorely needed (Ellis and Kalumbi, 1999; Ellis and Attwood, 2001; Abbott, 2009; Bastow and Leonelli, 2010; Baker, 2012; Hayden, 2013).

Fifteen years beyond the original survey, we were curious to know which of those biological databases that were alive at the end of the 20th century had managed to persist into the 21st? In particular, we were keen to understand what distinguishes the survivors from the rest. In an at-

tempt to answer these questions, we planned to return to the DBcat listing that had underpinned the 1998 survey.

Methods

Ironically, the DBcat database itself died in 2006. However it – and much of the older Web – lives on in the Internet Archive[1]. DBcat was first archived in May 1997, when its home page reported it contained information on 383 databases. This archive[2] was used in the present study. The full data-set used in this study is presented in the spreadsheet, Supplementary File 2[3]; an explanation of the contents of each Sheet in the spreadsheet can be found in Supplementary File 1[4].

DBcat records were examined for each database entry. Eight were duplicates, leaving 375 databases (see Sheet 1 in Supplementary File 2)[3]. Each of these was examined in turn to determine: i) whether it was indeed a public Web-based database; ii) if so, whether it was still 'alive' in the first half of 2015; and, iii) if alive, when it was last updated.

What is a public web database? Information in DBcat was, for the most part, entered by the database maintainers themselves. We eliminated five as commercial, two as links to a research group or research centre, and 31 as lists of information lacking even a search function or in other ways not a Web database. Four others, freely available initially but commercial upon re-study, were also eliminated. Some databases might disappear, and their name could be used, knowingly or unknowingly, for a newer database in the same field. In nine situations, we could not determine whether or not this occurred; we classed the state of these databases as **unclear** and removed them from the set (see Sheet 3 in Supplementary File 2[3] for a list of all excluded entries). This left 326 entries (see Sheet 2 in Supplementary File 2)[3].

What does 'life' mean for a Web-based database? Determining what constitutes 'life' or 'death' for a Web-based database is non-trivial – answers to the question are not black or white. If the data in a database had been transferred to another, different database, such as the transformation of the collection of 'Modules in Extracellular Proteins', which was published as SMART (Shultz *et al.*, 1998), we classed the original database as **alive–rebranded**.

Some live databases contain notices stating, for example, that they are no longer updated (*e.g.*, the Blocks Database (Henikoff *et al.*, 2000)), or their database history shows that to be the case. Databases that had not been updated since 2012 or earlier but were still functional and searchable, even if only in mirrors, we considered to have been **archived**.

We also found databases whose search function had either disappeared or was non-functional or had lost other key functionality. We counted these as **dead** even if they still existed on the Web. If a mirror site was being updated (*e.g.*, SCOPe at the University of California Berkley (Fox *et al.*, 2014)), the database was classed as **alive**, even if the parent was dead or archived.

In an attempt to gain insight into the relative 'health' of some of these resources, we looked more closely at the 46 databases from the DBcat DNA category included in our analysis. The approach was purely qualitative: databases maintained by large groups or consortia at institutes or organisations whose main mission was service provision at some level, or that were funded privately, we considered to have strong financial support; those that appeared to be maintained by individuals, especially those in academic environments, we considered to have weaker financial support.

The status of these databases changes as we speak: their URLs change; they change their names; their data move. If alive at one moment, they may be archived at the next; if archived, they are eventually likely to die; dead databases might even return to life. Our data and analyses are hence a snapshot of a moving target, and should consequently be read in that light.

Results

As shown in Table 1 and illustrated in Figure 1, of the 326 entries investigated, we classed 53 as alive, 23 as alive-rebranded, and 47 as archived; according to our criteria, a total of 203 (62%) were dead (see Sheet 2 in Supplementary File 2)[3].

Of the 46 entries in the DBcat DNA category, 21 were alive or alive-rebranded, three were ar-

1 archive.org/
2 web.archive.org/web/19970502044745/http://www.infobiogen.fr/services/dbcat/
3 http://journal.embnet.org/index.php/embnetjournal/article/downloadSuppFile/803/1096
4 http://journal.embnet.org/index.php/embnetjournal/article/downloadSuppFile/803/1095

Table 1. 18-year survival status of 326 databases from the May 1997 DBcat listing.

Category	N	Percent
Alive	53	16.3%
Alive - rebranded	23	7.0%
Archived	47	14.4%
Dead	203	62.3%
TOTAL	**326**	**100%**

Figure 1. Illustration of the data listed in Table 1, showing the proportions of databases that were alive, dead (or becoming so) after a period of 18 years.

chived, 22 were dead, six were excluded, and one was unclear.

Of the 21 alive or alive-rebranded databases, 17 (81%) were supported by stronger financial infrastructures than the others. Of the 22 dead databases, most (73%) appeared to have had weaker financial support, in the sense of originating from academic environments, or research institutes whose core mission was not service provision (see Sheet 4 in Supplementary File 2)[5].

Discussion

Classification

We classified databases as alive according to whether they were updated in 2013 or more recently, and as archived if they were not. We accept that this is an arbitrary cut-off date, but while some archived databases may simply be 'resting' during funding droughts and may resume updates when funds begin to flow again, equally, those that are currently alive may cease to do so if they hit funding deserts – the likelihood is that these numbers will balance.

We excluded databases that were commercial in the original 1997 DBcat data, as they were never public databases. We also excluded those that became commercial after that time. We could have instead classed those as 'dead', as they are no longer public databases.

Database Longevity

Database longevity depends on finding a continuous funding source. This is possible, say, for a database that supports the main focus of its host institution: for example, the Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (DSMZ) hosts one of the largest microbial culture collections worldwide. Its free Web catalogues[6] will be around as long as the DSMZ exists; they are funded, and updated, as a key part of their institution's mission.

It is sobering how many of the 326 databases were found to be dead (62%) or to exist in an archived state (14%) – the situation may actually be worse than this, as the authors have personal communications of funding problems for some of the databases classed as alive. Regardless, the figures are consistent with the results of the 1998 survey, in which 68% of responding database curators claimed uncertain 1-5 year financial futures for their resources.

Economic models

Previous work listed several economic models that are, or could be, used for the support of biological databases. We looked at public funding, asymmetric pricing, advertising, deal-making, direct sales and hybrids (Ellis and Kalumbi, 1999); a decade later, several of these models, and their inherent complexities, were also reviewed by Bastow and Leonelli (2010).

Some databases evolve to include more than their database functions, including income-producing endeavours (direct sales), which may help fund database costs: for example, an important focus of the DSMZ catalogues is listing the price of their cultures and how to order them.

Public funding remains the most frequently used financial model, with well-known problems when such funding ceases: for example, in 2009, The Arabidopsis Information Resource (TAIR, (Lamesch *et al.*, 2012)) lost its public funding, generating a relatively large amount of publicity for its plight (Abbott, 2009). Other databases in the alive and archive categories face, or have faced, similar problems (Baker, 2012).

5 http://journal.embnet.org/index.php/embnetjournal/article/downloadSuppFile/803/1096

6 www.dsmz.de/catalogues.html

Asymmetric pricing – charging some users more than others – is less frequently used. TAIR, for example, is now funded by subscriptions, charging commercial organisations more than educational institutions or non-profits. It is not yet clear how successful this strategy may be (Hayden, 2013). Other databases may offer some content free and the complete version for a license fee: for example, Transfac (Matys *et al.*, 2006) has a free public version that is more than 10 years older than its commercial version. Commercialisation is only viable for those databases with a sufficiently large subset of users who are willing and able to pay for commercial versions.

Advertising is not used, in part because advertisers are unwilling to pay for display on the relatively low-traffic Web pages of most biological databases. Corporate sponsorship is part advertising and part deal-making: the corporation pays to help support a database that provides value to its potential customers, who may see its logo and a link to its website listed under 'Sponsors', and gains good-will. No biological database has gained appreciable funding through such sponsorship.

What distinguishes survivors?

It is interesting to reflect on the enormous investment that has been made during the last 20-30 years to establish and sustain so many biological databases, and the energy – the human cost – it has taken to maintain them. More than 60% of Web-based databases available in DBcat in 1997 have died – a significant waste of investment. The persistence of Web-based resources is a known problem: *e.g.*, Hennessey and Ge (2013) found that the median lifespan of Web pages referenced in article abstracts from the Web of Science citation index, published between 1996 and 2010, was around nine years, 62% of them being archived. Similarly, our analysis has shown that while a small number of the 1997 DBcat databases have been able to persist through rebranding exercises, many others are only now accessible in some archived form (in which their value, and future accessibility, is likely to erode further with time). Less than 20% are still actively maintained.

Case studies

Those databases that do persist today have clearly had winning survival strategies. Many have experienced funding crises, and have had to be rescued from the brink of extinction. Swiss-Prot is a case in point (Bairoch *et al.*, 2004; Bairoch, 2000). In 1996, Swiss-Prot hit a problem: an application for renewal of a grant from the Swiss National Science Foundation (SNSF) was turned down, because the database was being widely used outside Switzerland, and SNSF funds were intended to support primarily national, rather than international, projects; at the same time, an application to the EU was declined, because its infrastructure grants were intended to complement existing local funding, which the SNSF had just declined to provide. To alert users to the problem – at this point, funds existed only for two further months of the biocurators' salaries – an Internet appeal was launched, announcing that Swiss-Prot would disappear on 30 June 1996 if no solution could be found. The *"Internet storm of protest"* that followed did not go unheeded: the Swiss scientific funding agencies recommended that a stable, long-term funding mechanism be sought to sustain the database (Bairoch, 2000). Interim funding was provided on a short-term basis, from 1997-1999; during this time, Bairoch and his colleagues were involved in high-level talks that led to the creation, in 1998, of the SIB Swiss Institute of Bioinformatics as a non-profit foundation, providing the database with a 'permanent' home (Bairoch, 2000).

One consequence of this was that, by Swiss law, the government could only fund up to 50% of the budget of such an institution, the remainder having to be found via other avenues, preferably commercial. Accordingly, a new company – GeneBio – was established as the commercial arm of the SIB. The licensing strategy adopted by GeneBio was, perhaps, unusual. The company's founders wanted to ensure that the methods by which academic and commercial users accessed Swiss-Prot would not change – it was therefore based on trust, relying on commercial users contacting the company to pay an annual licence fee. This system was very successful for several years; however, it was not the end of the story. Additional funding subsequently acquired from the National Institutes of Health (NIH) stipulated that access to the database must be free – Swiss-Prot could therefore no longer be sold commercially.

With this NIH funding, Swiss-Prot was subsumed into UniProtKB (Apweiler *et al.*, 2004; Bairoch *et al.*,

2004)), along with TrEMBL (Bairoch and Apweiler, 1996) and the Protein Information Resource Protein Sequence Database (PIR-PSD) (George *et al.*, 1986). Today, UniProtKB is managed by the European Bioinformatics Institute (EBI), the SIB Swiss Institute of Bioinformatics and the PIR – the UniProt Consortium – and falls under the protective umbrella of Europe's distributed infrastructure for life-science information, ELIXIR (Crosswell and Thornton, 2012). The PIR-PSD's role in this story is interesting, not least because it had competed with Swiss-Prot for many years. In principle, it gained a new lease of life through the creation of UniProtKB. However, for most users, the resource has become largely invisible, archived in UniParc and not overtly visible in UniProtKB except via given entries' database cross-references.

Probably the oldest biological database still in use is the Protein Data Bank (PDB), first launched in 1971 (Anonymous, 1971). Inevitably, during its more than 40-year history, the PDB has faced its share of funding struggles – not least, in the late 1990s, when the funding agencies invited researchers to submit competitive grant proposals in a bid to stabilise the resource and improve its efficiency. This eventually led to a new consortium approach to its management – the so-called Research Collaboratory for Structural Bioinformatics (RCSB) – and with it, a move, in 1999, from its location at the Brookhaven National Laboratories to Rutgers, The State University of New Jersey (Berman *et al.*, 2000), where it remains today.

Aside from UniProtKB and the PDB, amongst the strongest surviving databases are EMBL (now part of ENA (Cochrane *et al.*, 2013)), GenBank (Benson *et al.*, 2014), DDBJ (Kosuge *et al.*, 2014), Ensembl (Flicek *et al.*, 2014) and InterPro (Mitchell *et al.*, 2014). Several of these will benefit from being part of ELIXIR, in which they are 'named services' that may ultimately qualify for core support, whether at the EBI or at designated ELIXIR Nodes across Europe as their host countries ratify ELIXIR's Consortium Agreement. ELIXIR is a pan-European, inter-governmental initiative seeded by the European Strategy Forum on Research Infrastructures (ESFRI), which, in 2002, set out to support the long-term needs of European research communities.

Of course, originating at an institute, organisation or Node with strong financial support is not a guarantee of strong *database* support, and is hence not in itself a guarantee of longevity, especially if the host institution loses its core funding and closes, or undergoes rebranding and mission evolution, or if the key author leaves. For example, of the databases observed to be dead in the DBcat DNA category, ALU (DBC0002) was developed at the NCBI by an individual who moved elsewhere; Genexpress (DBC00007) was developed at Infobiogen, which closed down; the HGMP Primers Database (DBC00280) was developed at the Human Genome Mapping Project Resource Centre, a UK Research Council-funded institute that closed down; and TIGR-AT (DBC00133), EGAD (DBC00197) and HCD (DBC00202) were developed at The Institute for Genome Research (TIGR), which rebranded as the J. Craig Venter Institute (JCVI), and no longer maintains or supports many of TIGR's databases (these databases are marked with an S* comment value in Sheet 4 in Supplementary File 2)[7].

Against this background, recognising the increasing importance of data, or rather, of 'big data', in underpinning advances in biomedicine, a trans-US-NIH initiative – Big Data to Knowledge (BD2K) – was recently launched in the United States (Margolis *et al.*, 2014). BD2K will facilitate biomedical research, in part by supporting a 'data ecosystem' that is able to accelerate knowledge discovery. Discussions of possible interactions between ELIXIR and BD2K are in their infancy, and it will be interesting to see what concerted plans, if any, may emerge for sustaining a data ecosystem globally. Meanwhile, it's clear that European databases that do not belong to ELIXIR Nodes will face much stiffer competition for funds in future, as governments divert resources to sustain their central Nodes. Whether this will be an affordable model remains to be seen. ELIXIR may seem like a light at the end of a long and dark funding tunnel for some databases, but may ultimately cause the lives of many more to be extinguished.

Access to data in perpetuity?

The last point brings us to the issue of 'biodiversity'. Diverting funds primarily to large, successful databases threatens the existence of smaller but nonetheless valuable resources. Consider, for example, InterPro, which integrates around 12 different databases (including PROSITE (Sigrist

7 http://journal.embnet.org/index.php/embnetjournal/article/downloadSuppFile/803/1096

et al., 2013), PRINTS (Attwood *et al.*, 2012), and Pfam (Finn *et al.*, 2014)) and was developed as a key tool for automatic annotation of TrEMBL entries (Apweiler *et al.*, 2001; Mitchell *et al.*, 2014). By virtue of being housed at the EBI, InterPro may achieve some future measure of protection under ELIXIR; however, its source databases that are not maintained at the EBI – most of them – will not. InterPro is thus in danger of losing many of its partners and, with them, much of its diagnostic strength and richness. Ultimately, it is in danger of becoming a mere HMM-based resource, its 'biodatabase biodiversity' completely lost.

Another interesting issue that has emerged in recent years has been the drive to create 'open data repositories'. Just as the Open Access movement drove the creation of Institutional Repositories to archive research papers, similar arguments are pressuring universities into establishing their own research data repositories; there are also moves afoot to create citable 'data papers', to incentivise (rather than mandate) scientists to deposit their data. How this will work in practice is unclear.

One of the drivers behind initiatives like this is the desire to improve research communication by coupling scientific articles more strongly with their research data (Bourne *et al.*, 2011). This will require the research community to *"develop best practices for depositing research data-sets in repositories that enable linking to relevant documents, and that have high compliance levels driven by appropriate incentives, resources and policies."* This vision takes us beyond the problems of how to maintain a few hundred biological databanks, into a world in which we will have to figure out how to archive all published research data such that they will be accessible and searchable for all time. Even if we accept that a static data archive is different from a functional (and evolving) database, if we have not yet solved the sustainability problems for biological databases, it will be interesting to see how archives for *all* research data will be managed in perpetuity.

Regardless, the good news is that, at least at some level, the scientific community and the bodies that fund scientific research have woken up to the importance of organising and archiving research data. Whether this will help to address some of the meatier issues of long-term database maintenance is moot. What remains clear is that this is still very much an unsolved problem, one that the International Society for Biocuration (ISB) is beginning to consider very seriously. The Society has observed that, while research infrastructures are becoming more widespread, securing funding for database maintenance is still problematic, even for well-established databases – although funders are generally keen to support projects that generate yet more data, there is still insufficient recognition of the importance of data curation. This motivated the ISB to launch a survey in order to gain an overview of the financial situation of databases managed by its current members. The results of the survey will be shared at a workshop (*Money for biocuration: strategies, ideas & funding*) to be held at the 8th International Biocuration Conference in Beijing, 23-16 April 2015, in which participants will have the opportunity to discuss what the ISB, and biocurators in general, can do to help. We look forward, with great interest, to the outcomes.

Conclusion

Much has changed since the 1998 database survey, but there are also several constants. Biological databases are expensive to create and maintain; nevertheless, databases continue to be created afresh each year. Far from stemming the tide of new repositories, some funding bodies are requesting researchers to elaborate 'data management plans' as part of their research proposals. Compelling scientists to explain how their data will be archived and made accessible seems like an important step forward, especially as responsibility for their financial future is being pushed onto institutions. Nevertheless, initiatives like this will not guarantee the long-term sustainability of databases, whose value to the community depends on active update and maintenance schedules rather than passive archiving.

Despite past funding issues, some of the most successful databases have survived by being integrated into larger database federations (ENA, UniProt, InterPro for example). Above all, however, it is clear that institutional support is a key feature in the precarious ups-and-downs of the database-funding landscape. Regardless of their sustainability strategy, databases require the input of skilled biocurators and bioinformaticians, and their ongoing commitment will continue to be costly to support in the long term.

As larger databases battle for their futures, many more smaller, specialist databases are being lost along the way. European infrastructures like ELIXIR and funding initiatives like BD2K will certainly have a significant role to play in securing the long-term future of some key databases, and of the biocurators and bioinformaticians required to manage them. It is too early to tell what the data ecosystem of tomorrow will look like; nevertheless, it is probably safe to say that it will be dominated by many of the most successful databases of today.

Key Points

- Active maintenance and update of public, Web-based biological databases is time-consuming and costly.
- Without financial sustainability plans, most databases created as outputs of research projects consequently die or are archived within 10-15 years.
- Longevity is a function of core institutional support.
- Researchers should understand these facts before creating a database independently of core infrastructures underpinned by long-term financial strategies.

Acknowledgements

The authors wish to thank the legions of scientists and educators who have developed and maintained biological Web-based databases, past and present.

References

Abbott A (2009) Plant genetics database at risk as funds run dry. Nature 462, 258-259. http://dx.doi.org/10.1038/462258b

Anonymous (1971) Protein Data Bank. *Nature New Biology* **233**, 223. http://dx.doi.org/10.1038/newbio233223a0

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29**(1), 37-40. http://dx.doi.org/10.1093/nar/29.1.37

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B *et al.*. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **32**(Database issue), D115-119. http://dx.doi.org/10.1093/nar/gkh131

Attwood TK, Coletta A, Muirhead G, Pavlopoulou A, Philippou PB *et al.* (2012) The PRINTS database: a fine-grained protein sequence annotation and analysis resource – its status in 2012. *Database* (Oxford) 2012:bas019. http://dx.doi.org/10.1093/database/bas019

Bairoch A (2000) Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times! *Bioinformatics* **16**(1), 48-64. http://dx.doi.org/10.1093/bioinformatics/16.1.48

Bairoch A, Apweiler R (1996) The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res.* 24(1), 21-25. http://dx.doi.org/10.1093/nar/24.1.21

Bairoch A, Boeckmann B, Ferro S, Gasteiger E (2004) Swiss-Prot: juggling between evolution and stability. *Brief. Bioinform.* **5**, 39-55. http://dx.doi.org/10.1093/bib/5.1.39

Baker M (2012) Databases fight funding cuts. *Nature* **489**, 19. http://dx.doi.org/10.1038/489019a

Bastow R, Leonelli S (2010) Sustainable digital infrastructure. *EMBO Rep.* **11**(10), 730-734. http://dx.doi.org/10.1038/embor.2010.145

Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J *et al.* (2014) GenBank. *Nucleic Acids Res.* **42**(D1), D32-D37. http://dx.doi.org/10.1093/nar/gku1216

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**(1), 235-242. http://dx.doi.org/10.1093/nar/28.1.235

Bourne P, Clark T, Dale R, de Waard A, Herman I *et al.* (eds.) (2011) The Force11 White Paper: Improving Future Research Communication and e-Scholarship. A publication resulting from the Schloss Dagstuhl Perspectives Workshop: The Future of Research Communication, 15-18 Aug 2011. http://www.force11.org/white_paper

Burks, C., Fickett, J.W., Goad, W.B., Kanehisa, M., Lewitter, F.I., Rindone, W.P., Swindell, C.D., Tung, C.S. and Bilofsky, H.S. (1985) The GenBank nucleic acid sequence database. *Comput. Appl. Biosci.* **1**(4), 225-233. http://dx.doi.org/10.1093/bioinformatics/1.4.225

Cochrane G, Alako B, Amid C, Bower L, Cerdeño-Tárraga A *et al.* (2013) Facing growth in the European Nucleotide Archive. *Nucleic Acids Res.* **41**, D30-D35. http://dx.doi.org/10.1093/nar/gks1175

Crosswell LC1, Thornton JM. (2012) ELIXIR: a distributed infrastructure for European biological data. *Trends Biotechnol.* **30**(5), 241-242. http://dx.doi.org/10.1016/j.tibtech.2012.02.002

Discala C, Ninnin M, Achard F, Barillot E, Vaysseix G (1999) DBcat: a catalog of biological databases. *Nucleic Acids Res.* **27**, 10-11. http://dx.doi.org/10.1093/nar/27.1.10

Ellis LBM, Kalumbi D (1998) The demise of public data on the web? *Nature Biotechnology* **16**, 1323-1324. http://dx.doi.org/10.1038/4296

Ellis LBM, Kalumbi D (1999) Financing a Future for Public Biological Data. *Bioinformatics* **15**, 717-722. http://dx.doi.org/10.1093/bioinformatics/15.9.717

Ellis LBM, Attwood TK (2001) Molecular Biology Databases: Today and Tomorrow. Drug Discovery Today **6**, 509-513. http://dx.doi.org/10.1016/S1359-6446(01)01802-5

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.* **42** (D1), D222-D230. http://dx.doi.org/10.1093/nar/gkt1223

Flicek P, Ridwan Amode M, Barrell D, Beal K, Billis K, *et al.* (2014) Ensembl 2014. *Nucl. Acids Res.* **42** (D1), D749-D755. http://dx.doi.org/10.1093/nar/gkt1196

Fox NK, Brenner SE, Chandonia JM (2014) SCOPe: Structural Classification of Proteins – extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **42** (D1), D304-D309. http://dx.doi.org/10.1093/nar/gkt1240

George DG, Barker WC, Hunt LT (1986) The protein identification resource (PIR). *Nucleic Acids Res.* **14**(1), 11-15. http://dx.doi.org/10.1093/nar/14.1.11

Hamm GH, Cameron GN (1986) The EMBL data library. *Nucleic Acids Res.* **14**(1), 5-9. http://dx.doi.org/10.1093/nar/14.1.5

Hayden EC (2013) Popular plant database set to charge users. *Nature News* (31 August 2013) http://dx.doi.org/10.1038/nature.2013.13642

Henikoff JG, Greene EA, Pietrokovski S, Henikoff S (2000) Increased coverage of protein families with the Blocks Database servers. *Nucleic Acids Res.* **28**, 228-230. http://dx.doi.org/10.1093/nar/28.1.228

Kosuge T, Mashima J, Kodama Y, Fujisawa T, Kaminuma E *et al.* (2014) DDBJ progress report: a new submission system for leading to a correct annotation. *Nucleic Acids Res.* **42** (D1), D44-D49. http://dx.doi.org/10.1093/nar/gkt1066

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40** (D1), D202-D210. http://dx.doi.org/10.1093/nar/gkr1090

Margolis R, Derr L, Dunn M, Huerta M, Larkin J *et al.* (2014) The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J Am Med Inform Assoc.* **21**, 957-958. http://dx.doi.org/10.1136/amiajnl-2014-002974

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108-D110. http://dx.doi.org/10.1093/nar/gkj143

Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S *et al.* (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* **43** (D1), D213-D221. http://dx.doi.org/10.1093/nar/gku1243

Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N *et al.* (2013) New and continuing developments at PROSITE. *Nucleic Acids Res.* **41**, D344-D347. http://dx.doi.org/10.1093/nar/gks1067

Schultz J, Milpetz F, Bork P, Ponting CP (1998). SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 5857-5864.

1000 Databases for the Bioinformatician

by W. Touw, E. Krieger, G. Vriend
CMBI, Radboudumc, Nijmegen (NL)

Very (too) many projects in bioinformatics are directed towards protein secondary structure prediction, and we have provided -for longer than the existence of the Internet - free access to the DSSP software and databases needed for these projects.

The CMBI protein structures facilities further include HSSP, WHAT IF and its derived WHAT_CHECK and PDBREPORT, the PDBFINDER, several improved PDB variants (PDB_REDO, BDB), and a few more. Some of these facilities are also available in the eBioKit (Figure 1), either through MRS or via YASARA. WHAT IF can now calculate a wide variety of features for the whole (useful subset of the) PDB.

Together with the lists of sequence unique chains in the PDB (PDB_SELECT), these data could potentially spur a flurry of prediction software activities. Currently, 20 datasets are available at swift.cmbi.ru.nl/gv/lists/, but the potential for new datasets is unlimited.

These first 20 datasets fall in five main categories: 1) elementary geometric aspects, such as bond, torsion angles and surface areas; 2) amino acid contact prediction- this got a big boost recently, but research has focused on the reduction of false-positive prediction, rather than the equally important definition of what is a contact.

The second group of sets therefore deals in many ways with amino acid contacts in proteins; 3) symmetry contacts, contacts with ions, and salt-bridges; 4) sets of augmented PDB files in which, for example, symmetry calculations have been worked out; 5) 'other' datasets.

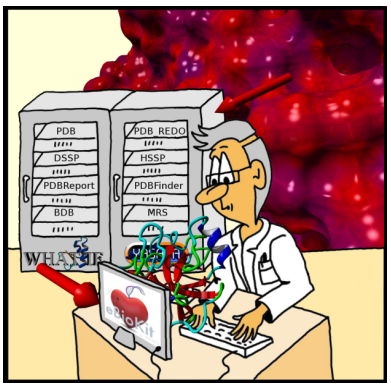


Figure 1. Andreas working with the eBioKit.

Users can download individual files or entire datasets. New datasets will be made on request, provided that WHAT IF can produce the requested data with existing options. We encourage people to ask for novel datasets, because this can only stimulate the important field of protein structure bioinformatics.

A series of PDB-related databanks for everyday needs. Touw WG, Baakman C, Black J, Te Beek TA, Krieger E, Joosten RP, Vriend G. Nucleic Acids Res. 2015 Jan 28;43 (Database issue):D364-368. doi: 10.1093/nar/gku1028. PMID: 25352545

Broadening the bioinformatics infrastructure to unicellular, animal and plant science

by E. Bongcam-Rudloff and A. Gisel

AllBio was the product of responding to a former KBBE call entitled '*Supporting the development of Bioinformatics Infrastructures for the effective exploitation of genomic data: Beyond health applications*'. A group of 'old' EMBRACE members - most of them members of EMBnet - accepted the challenge, and formulated a proposal, applying the experience that this group had acquired and developed in past projects.

The project – AllBio – was awarded to a consortium of 10 partners from 8 countries, giving them the opportunity to increase bioinformatics awareness and to spot the bioinformatics needs and bottlenecks in non-human-health fields, such as animal, microbiology and plant science. The project was based on so-called 'test-cases', in which the aforementioned biological communities formulated data-analysis problems they faced but for which they did not know the right approaches or have the right tools to solve them.

AllBio collected more than 60 test-cases from across Europe, and selected 15 that represented generally encountered bioinformatics problems. The AllBio partners, together with specifically selected bioinformatics specialists, organised several events between bioinformaticians and biologists to discuss and try to solve these problems. The most successful events were problem-specific hack-a-thons based on a series of community-building

workshops, where several teams physically sat together for several days and, on the spot, produced successful software solutions. Some of these new software packages have been published and made available for other end-users.

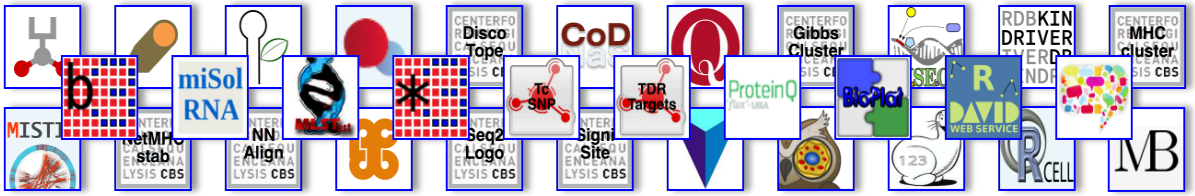
As a Coordination Support Action, AllBio was very successful in a) demonstrating how, in a very cost-efficient but productive way, to solve such data analysis problems, bringing together specialists from diverse disciplines, and b) delivering some outstanding solutions by applying these strategies. For more information visit www.allbioinformatics.eu

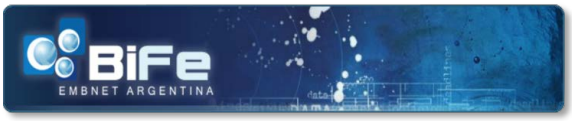
EMBnet.digest

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.

Bioinformática Federal (BiFe) 2015 Activities

by Ignacio Enrique Sánchez, EMBnet Argentina





Bioinformática Federal, **BiFe**, is the Argentine Node of EMBnet. The site groups bioinformatics and genomics resources (fully or partially) developed in Argentina. BiFe was devised to contribute to the accessibility of research results, and strengthen Argentine bioinformatics and genomics communities.

BiFe hosts, or links to, 32 bioinformatics resources, whose logos are shown above. Of these servers and databases, 18 are dedicated to the analysis of protein and nucleic acid sequence and structure, 10 to genomics and drug design, three to image analysis, while one is a database of biological databases.

Two application servers hosted by BiFe have been developed using our **open-source toolkit**, which is publicly available on our website, under "*Build your own server*", together with a tutorial.

The general-purpose toolkit can help to build an application server by providing easy implementation of fully customisable input and output forms, file uploading, automatic FTP file retrieval, dynamic application loading, tables, graphics and protein structure representation.

Moreover, our **database of 30 genomes** showcases the results of genome projects dealing with Argentine organisms and carried out (totally or partially) in Argentina. Of these genome projects, 15 deal with bacteria relevant to human health, the food industry and agriculture, nine with bacteria from extreme environments - such as Antarctica and high-altitude Andean lakes - and six with eukaryotes.

BiFe is currently located at the Protein Physiology Laboratory, Departamento de Química Biológica and IQUIBICEN-CONICET, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina.

The current EMBnet staff include Dr. Adrián G. Turjanski and M.Sc. Leandro G. Radusky. We acknowledge funding from Ministerio Argentino de Ciencia, Tecnología e Innovación Productiva (MINCYT) and Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).

IMGT®: Past, Present, Future

by Marie-Paule Lefranc & Sofia Kossida , IMGT, Montpellier, France.

IMGT®, the international ImMunoGeneTics information system®, was created in 1989 by Marie-Paule Lefranc at Montpellier, France. The founding of IMGT® marked the advent of immunoinformatics, a new science, which emerged at the interface between immunogenetics and bioinformatics.

For the first time, ImmunoGlobulin (IG) or antibody and T cell Receptor (TR) Variable (V), Diversity (D), Joining (J) and Constant (C) genes were officially recognised as 'genes' alongside conventional genes. This major breakthrough allowed genes and data of complex and highly diversified adaptive immune responses to be managed in genomic databases and tools.

IMGT® comprises seven databases, 15,000 pages of Web resources and 17 tools, and provides a high-quality and integrated system for the analysis of genomic and expressed IG and TR repertoires of adaptive immune responses. These tools and databases are used in basic, veterinary and medical research, in clinical applications and in therapeutic antibody engineering and humanisation.

IMGT® has been built on the IMGT-ONTOLOGY axioms and concepts, which bridges the gap between genes, sequences and three-dimensional (3D) structures.

The IMGT® standards are used in clinical applications and in therapeutic antibody engineering. Thus, IMGT/V-QUEST is frequently used by clinicians for the analysis of IG somatic hypermutations in leukemia, lymphoma and myeloma, and, more particularly, in Chronic Lymphocytic Leukemia (CLL), in which a low percentage of mutations of the rearranged IGHV gene in the VH domain of the leukemic clone have a poor prognostic value for patients.

A new era is opening with the use of Next Generation Sequencing (NGS), and the IMGT/HighV-QUEST Web portal has become the paradigm for the analysis of the adaptive immune repertoire in normal (vaccination) and pathological situations (infectious disease, …).

IMGT® standards use is also more needed than ever before in therapeutic antibody humanisation and engineering, as demonstrated by the IMGT/DomainGapAlign tool and the IMGT/2Dstructure-DB, IMGT/3Dstructure-DB and IMGT/mAb-DB databases.

GOBLET : Major Achievements from 2012 to 2015

By Teresa K. Attwood

Under the auspices of its 24th AGM, held in Uppsala (SE) in June 2012, EMBnet invited leaders of nine bioinformatics, biocuration and computational biology societies and networks to a workshop to discuss global bioinformatics training initiatives.

Like EMBnet, each of these organisations included some kind of education and training committee or programme, each with similar aims, and each with the same problem: how to deliver tangible benefits to their communities with limited funds and a mere handful of time-pressed volunteers.

The participants concluded that it would be useful to establish an umbrella organisation to coordinate bioinformatics training activities world-wide: to share, not duplicate, effort; to share, not duplicate, cost; to work together towards common solutions and a sustainable future. Thus was born GOBLET, the Global Organisation for Bioinformatics Learning, Education and Training.

GOBLET was subsequently established, in November 2012, as a legal foundation: its mission, to provide a global, sustainable support structure for bioinformatics trainers and trainees; to facilitate bioinformatics capacity development in all countries; to develop standards and guidelines for bioinformatics education and training; to act as a hub for fund gathering; to reach out, amongst others, to high-school teachers, to bridge the gap to the next generation of bioinformaticians; and to foster international communities of bioinformatics trainers.

With membership having grown more than threefold, GOBLET is going from strength to strength.

Notable highlights include:

- Three publications (in Bioinformatics, EMBnet.journal & PLoS CB),
- Development of a joint training strategy with ELIXIR,
- Holding education & training workshops (in Manchester, Boston & Toronto),
- Running global surveys of bioinformatics training needs,
- With the ISCB, establishing an education track for posters at ISMB conferences,
- Again working with the ISCB, launching the Computational Biology Education (CoBE) Community of Special Interest (COSI), to harmonise the ISCB education & GOBLET training communities,
- Winning a grant from the Canadian Institutes of Health Research to support the 2014 AGM,
- Planning the fourth AGM in Cape Town (ZA), November 18-20.

EMBnet can be proud to have spearheaded this highly successful initiative. To get involved, please contact us at www.mygoblet.org.

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 AGM2015
A new start

by Teresa K. Attwood

The successful events of the 2015 AGM were hosted by Pedro Fernandes at the Instituto Gulbenkian de Ciencia, Oeiras (PT), 10-12 June. This was one of our most important meetings in recent years because i) having been in post for six years, most members of the Executive Board (EB) were obliged to step down, and ii) the assembled Board ratified a major new initiative to invest strategically in EMBnet's future.

The events included a full-day Operational Board (OB) meeting (10 June), the "*Active Investment Strategy*" workshop (11 June), and finally, the formal business meeting (12 June) – here, of the 25 paid-up members, 19 were present or represented by proxy.

Crucially, during the AGM, three nominations were made, and accepted, for new members to step up to the EB. In consequence, we warmly welcomed a new leadership team: Domenica D'Elia, Lubos Klucar, Emiliano Barreto and Erik Bongcam-Rudloff (whose term will end at the 2016 AGM). Given the scale of this change, it was agreed that Etienne de Villiers and Terri Attwood would remain affiliated with this team as part of an interim 'extended EB', to mentor the new

members and help implement and evaluate the success of the investment strategy.

In addition, as they have not functioned optimally in recent years, it was agreed that the Committees would not persist in their current guise; however, their interests will continue to be championed through Special Interest Groups (SIGs) led by Axel Thieffry, George Magklaras and Pedro Fernandes, representing EMBnet's publicity & public relations, technical and education/training needs respectively.

Importantly, the Board also unanimously endorsed the OB's proposals to i) hire an editorial assistant for one year to support the work of *EMBnet.journal*, and ii) make a substantial donation to GOBLET in order to hire an assistant for two years, tasked with producing educational materials jointly branded and endorsed by GOBLET.

The progress made as a result of these investments will be reported monthly to the extended EB, and evaluated by the full Board at the next AGM. Collectively, these are very exciting steps forward, and we look forward to seeing how the journey develops in the years ahead.

Recent highlights from the CNR Institute for Biomedical Technologies: EMBnet Italy

by Domenica D'Elia

The CNR-ITB research team in Bari, Italy, was among the pioneers of bioinformatics research in Italy. As a founder member of EMBnet, it served as the national bioinformatics service hub from 1989, and has continued to do so for more than 20 years.

Alongside this primary role, the Institute was also the first in Italy equipped with a 454 ROCHE Next Generation Sequencing (NGS) platform. The advent of NGS technologies has completely transformed how research is performed, requiring multidisciplinary collaborations and establishment of strong interconnections between wet-lab and bioinformatics researchers.

The Italian EMBnet team comprises biologists (specialising in functional genomics and transcriptomics, cancer research, neurological diseases, organism evolution and phylogeny) and computer scientists, statisticians and bioinformaticians (specialising in the development of analysis algorithms, databases and platforms for integrative bioinformatics).

Thanks to this wide variety of competences, the Institute is able to carry out research in:
- biomedicine and biotechnology applications for cancer research, and identification of diagnostic and prognostic biomarkers for multiple sclerosis, Parkinson and Alzheimer syndromes;
- functional genomics and transcriptomics, including the most recent focus on expression profiling of non-coding RNAs for functional studies on their role in human gene expression, and genome epigenetic modifications in virus- and viroid-infected plants;
- biodiversity, involving development of bioinformatics tools for rapid identification of species and metagenomics studies.

Notable recent highlights come from two research lines, one focused on a strategy for p53 re-activation in chemo-resistant clear cell Renal Cell Carcinoma (RCC), the other focusing on the study of cross-kingdom interactions between miRNAs from edible plants, and human gene targets involved in cancer and age-related diseases. These projects are revealing interesting insights that could have significant impacts on the clinical treatment of RCC, and on the development of biotechnology applications in nutraceutical research for preventive medicine.

We welcome collaboration with members of EMBnet; to learn more, please feel free to contact Domenica D'Elia at domenica.delia@ba.tib.cnr.it.

Elixir-Denmark: First Annual Danish Bioinformatics Conference

by Axel Thieffry

The First Annual Danish Bioinformatics Conference was held on 27-28 August 2015, in Odense, Denmark. Organised by ELIXIR Denmark, this 200+ participant conference was also supported by the University of Copenhagen (UCPH), the Technical University of Denmark (DTU), the University of Southern Denmark (SDU) and Aarhus University (AU).

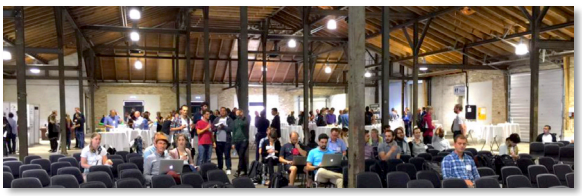
As one of the Nodes of ELIXIR (the pan-European initiative to build a sustainable infrastructure for biological information to support life-science research), the event (#elixirdk) provided the scientific community with an overview of the current state-of-the-art of bioinformatics and systems biology in Denmark.

The conference was organised around five main themes:

- Systems Biology and Medical Informatics,
- Proteomics Informatics,
- RNA Bioinformatics,
- Population Genetics, and
- Medical Genomics.

Picture @anttipursula

Poster sessions were organised to showcase the latest achievements in Danish bioinformatics research areas and tools development. Students and early-stage researchers were also invited to join CBIO Vikings, an ISCB Regional Student Group, to promote scientific interactions, collaborations and social networking.



Picture @vignirisberg

Of particular note, in an effort to bridge the gap between academia and industry, several private companies (Novo Nordisk, Exiqon and Novozymes) discussed the industrial view of Danish Bioinformatics, its challenges and opportunities.

Picture @axthief

Following the conclusion of this successful first conference, the second annual event has been scheduled to take place 25-26 August 2016.

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 contribution to Bioinformatics in Sri Lanka

by Kanchana Senanayake,
EMBnet Sri Lankka

The IBMBB (Institute of Biochemistry, Molecular Biology and Biotechnology, University of Colombo) became the Sri Lankan National node of EMBnet in 2007, when it was unanimously voted in at the AGM in Málaga (ES). The aim of this InFocus is to thank EMBnet for its contribution to our Bioinformatics MSc course, the first cohort of which has now successfully graduated. Needless to say, EMBnet made a significant contribution to developing the field of bioinformatics in Sri Lanka.

Dr. Erik Bongcam-Rudloff, current Chair of EMBnet, visited IBMBB in 2006, responding to an invitation made by SIDA and UPPMAX, University of Uppsala (SE). He convinced us to join EMBnet, highlighting the benefits we could receive as a member of this bioinformatics community. Prior to his visit, we could only organise 1- or 2-day workshops when experts in the field visited Sri Lanka. In 2009, supported by EMBnet, IBMBB initiated discussions with the University of Colombo School of Computing (UCSC) to start a joint MSc programme in bioinformatics. Guided by EMBnet, IBMBB and UCSC began formulating the course structure and syllabus. As a result, the first Bioinformatics MSc programme in Sri Lanka commenced in May 2012. This is a 4-semester, full-time course, with a research project that begins half-way through the third semester course-work and runs throughout the fourth semester.

Along the way, we continued to receive invaluable support from EMBnet. Dr. Jose Valverde (EMBnet Spain), ran a 4-day course on Structural & Functional Analysis of Proteins, and supervised an MSc dissertation. He generously held discussions with our researchers to solve their problems and provide new insights into their research projects. Drs Goran Neshich and Emiliano Barreto (EMBnet Brazil and EMBnet Colombia respectively) extended their support by evaluating MSc dissertations.

We have now produced five Bioinformatics MSc graduates. Three students did not want to proceed with a research project to earn the Master degree, so they will graduate with Postgraduate Diplomas in Bioinformatics. We hope EMBnet will continue to extend its support towards further development of bioinformatics in Sri Lanka.

NETTAB & IB 2015 "Two-Day Hands-on Tutorial"

by Domenica D'Elia, EMBnet Italy

The NETTAB & Integrative Bioinformatics *Two-Day Hands-on Tutorial: Bioinformatics Analysis of Omics Data* was held 12-13 October 2015, in Bari (IT) as a satellite event of the NETTAB & IB 2015 Joint Symposium. Organised by Domenica D'Elia (EMBnet Italy) and Paolo Romano (IRCCS San Martino IST, Genoa (IT)), it received generous support from EMBnet (which granted 10 free tutorial registrations), from the InterOmics project and from the IT resources of the ReCaS project (Programma Operativo Nazionale Ricerca e Competitività 2007-2013 – 4).

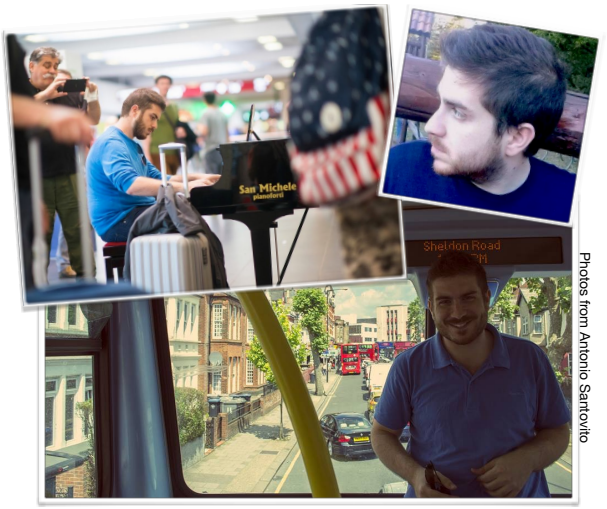
The tutorials (four in total) were held at the University of Bari's Department of Computer Science and organised in full- and half-day parallel sessions, the last replicated on both days. The aim of the programme was to provide researchers and students opportunities to learn about best practices in challenging bioinformatics tasks. Specifically, a full-day tutorial was dedicated to the analysis of re-sequencing data for detecting genomic variants in human diseases, such as somatic single nucleotide variants (SNVs) in cancer (run by Fabio Iannelli, IFOM, Milan, IT) and germline SNVs in rare Mendelian disorders (run by Anna De Grassi, Professor at the University of Bari).

The other full-day tutorial, run by Ioannis Vlachos (from the DIANA-Lab, Pasteur Institute of Athens and the Department of Computer & Communication Engineering of the University of Thessaly (GR)), covered the design and analysis of small-RNA-Seq experiments, methods for genome alignment and/or microRNA expression estimation, and the state-of-the-art of online tools for microRNA functional analysis. Emek Demir, from the Sloan Kettering Institute, New York (USA), ran a half-day tutorial on the use of the Pathway Commons Web service and BioPAX, a standard, community-developed language for the description of biological pathways. Pasqualina D'Ursi, from the CNR-ITB in Milan (IT), focused her tutorial on applications of fast and inexpensive docking protocols, combined with accurate molecular dynamics techniques to predict protein-ligand complexes.

I would like to thank first of all EMBnet, InterOmics, the ReCaS project and the Department of Computer Science for their precious support, and all the teachers and students both for their dedication and for the enthusiasm they expressed for this initiative.

"Hello, nice to meet you"

by Antonio Santovito,
EMBnet Editorial Assistant



Photos from Antonio Santovito

It all began this summer when my sister sent me a link, "Why don't you apply?" Months later, an invite for a Skype interview rattled my inbox. "Go to lunch, we want you at your best", someone wrote me before that call started on time that afternoon. "Hello, nice to meet you", said a voice from my computer..

Two months have passed since the day I met the EMBnet Editorial Board for the first time, and I am now the new Editorial Assistant. I am an Italian communicator, a media studies student and the owner of Creactive, my communication and marketing business. After a decade as a Web designer, almost two years as a journalist, and three as a contributor for an advertising agency, my experience with EMBnet began and it has been engaging since the first day. Working with a non-profit is an honour and also a pleasure for me because it's related to another of my favourite topics: science.

I had confirmation that this would be an amazing job when I met Domenica D'Elia in Bari (IT). She welcomed me in the best way, and we began working together with some media content. Meanwhile, the meetings with the Editorial Board were becoming more and more interesting. All is proceeding with one eye on the past and the other to the future, solving some technical issues with old publications, while making plans for the journal's promotion, including a new logo and a new layout, both for print and Web.

I know the EMBnet Journal Editorial Board is giving me a very important opportunity. Honestly, it daily fills my inbox with hundreds of conversations with scientists from all over the world, each with his personality, opinions and questions to monitor, manage and answer. I have to admit that sometimes these interactions can be very funny - my expectations have been more than satisfied. I will do my best to keep working with such a great network. So it's now my turn to say to you, "Hello, nice to meet you".

EMBnet.digest

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

a question of perspective

Vivienne Baillie Gerritsen

Paradigms are meant to be broken. In the 1980s, biology students were taught "the one gene = one protein" dogma which has since stepped down from its pedestal, as we now know that one gene, by way of any number of post-translational modifications on the protein sequence, can actually give rise to more than one protein. Or what would be more correct: to more than one function. In the same way, structural biologists are beginning to realise that proteins are not always stable bu t that a significant amount exist in particularly unstable forms – which has given them the name "disordered proteins". Until recently, proteins were thought to fold up into thermodynamically stable forms before getting on with what they had to do. Now we know that it is not necessarily the case. Eukaryotic translation initiation factor 4E-binding protein 2, for instance, is one such disordered protein whose lack of stability gives rise to a new kind of biological regulation.

by Jodee Knowles

Courtesy of the artist

Strangely enough, until the early 1950s, scientists believed that proteins were particularly malleable entities. Then along came the pioneering work on the necessity for two fragments of a ribonuclease to bind tightly for it to be effective, and scientists began to produce the first crystal structures of proteins. Ever since, a stable 3D conformation was considered to be the ideal state for a protein to function. However, in 1986 already, a handful of scientists were beginning to realise that perhaps a few proteins carried out their existence a little differently, and were somewhat unsure as to the stable conformation they wished to adopt. So didn't really adopt any at all.

But it took a further 20 years for such a notion to become popular. This is because of the angle from which structural biologists have been observing proteins. It is not an easy task to predict the kinetics and thermodynamics underlying the conformational states of a protein – not to mention those driving its catalytic reactions and binding properties. So, as is the case in scientific research, biologists set a basis from which they can make powerful correlations. In this case, low energy states and a limited number of combinations of macromolecules which provided links between the 3D conformation of proteins and their functions. However, it is becoming increasingly apparent that proteins carry out their business at higher energy states. Which is beginning to push the initial dogma over the cliff.

This just goes to show how paradigms – though necessary – can impede scientific progress by keeping the understanding of some phenomena within certain limits until other parameters, which researchers cannot ignore anymore, emerge and the paradigm is interrupted and takes a jump forward. The low energy paradigm gave huge insight – and over a long period – into the biological function of proteins, but it slowed the understanding, or acceptance, of highly dynamic states.

Proteins that lead a life in highly dynamic states are what has been coined "intrinsically disordered", because they do not adopt one sole three-dimensional conformation and stick to it, but rather

they embrace a series of different conformations – although, from a purely thermodynamic point of view, they remain stable. Contrary to expectations, disordered proteins are not a rare event; current predictions estimate that 15% of the proteome is, quote, fully disordered! How do we know, you may ask? Thanks to the field of computational biology and algorithms that are able to predict disorder…

It is hardly surprising that disordered proteins represent a significant challenge to structural biologists. To complicate matters further they are not to be considered only at the level of monomers… Disordered proteins lack perhaps a stable tertiary structure but they are able to carry out numerous biological functions, especially those associated with signalling, transcription regulation, cell division and differentiation. And, as for the more popular stable proteins, post-translational modifications (PTMs) of disordered proteins are a source of additional functions. As an example, disordered protein Eukaryotic translation initiation factor 4E-binding protein 2 (EIF4EBP2) is involved in the suppression of cap-dependent translation initiation, which is brought about by multiple phosphorylation of EIF4EBP2.

EIF4EBP2 is the major neural isoform of a family of proteins that bind to a translation initiation factor eIF4E – so long as another initiation factor known as eIF4G hasn't got there first! Binding or not binding to eIF4E all has to do with the conformation of EIF4EBP2 which depends on its phosphorylation; phosphorylation can occur at multiple sites. When EIF4EBP2 is highly phosphorylated, it is unable to interact with eIF4E, thus leaving the way open for eIF4G. When EIF4EBP2 is weakly phosphorylated or not at all, it binds to eIF4E very tightly and translation initiation is suppressed.

This is the first time researchers have discovered that translation initiation can actually be regulated via the structural polymorphism of a protein, itself mediated by phosphorylation – a novel mode of biological modulation led by intrinsically disordered proteins. Disorder to order (and vice versa) involves large conformational changes in a protein – as opposed to those which occur when the more "common" ligands bind to their target proteins for instance. EIF4EBP2 is the first protein to have been discovered which undergoes multiple phosphorylation bringing about an important conformational change.

Disordered proteins are shedding light on an entire new domain of biology and concomitantly shattering a long-standing dogma. They are able to carry out multiple functions by way of conformational plasticity, which itself depends on the proteins' state of phosphorylation. What is more, given regions within a disordered protein are – depending on their conformation – able to interact with different target proteins, thus lending the protein multispecificity. Biologists are also beginning to realise that disordered proteins are probably at the heart of evolution since they offer rapid regulatory complexity. This could explain the preponderance of disordered proteins in signalling networks within higher eukaryotes, and scientists expect them to be involved in various pathologies, especially those characterized by loss of biological reaction such as cancer. There is still much to learn about disordered protein PTM-induced folding but there is little doubt that these novel findings will have an important therapeutic impact.

Cross-references to UniProt

Eukaryotic translation initiation factor 4E-binding protein 2, *Homo sapiens* (Human) : Q13542

References

1. Bah A., Vernon R.M., Siddiqui Z., Krzeminski M., Muhandiram R., Zhao C.
Folding of an intrinsically disordered protein by phosphorylation as a regulatory switch
Nature 519:106-109 (2015)
PMID: 25533957

2. Forman-Kay J.D., Mittag T.
From sequence and forces to structure, function, and evolution of intrinsically disordered proteins
Structure 21:1492-1499 (2013)
PMID: 24010708

proteinspotlight

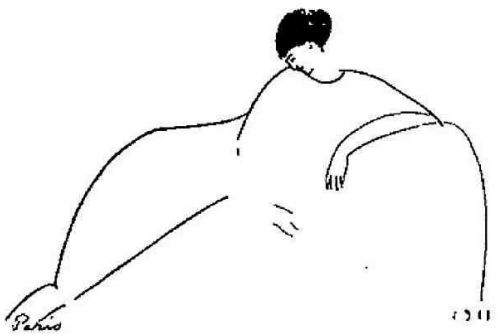
Swiss Institute of Bioinformatics

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

the length of things

Vivienne Baillie Gerritsen

It is important to know when to stop. A cell has to know when to stop expanding. A flower's pistil and its stamen when to stop elongating. And a flagellum to stop extending. Because there is a fair chance that without this knowledge, it would be difficult to keep organisms alive. But how do all these various parts of living matter know when the time has come to stop growing? There must be a mechanism of some kind. A sort of molecular device which holds up a STOP sign, or acts as a means of measure when something has reached the required shape or length. Recently, such a scheme was discovered in the flagella of the alga *Chlamydomonas reinhardtii*; a ruler of sorts that defines not only the length of the units which make up the axoneme but also the nature of the flagellum's structure. This molecular yardstick is a protein known as coiled-coil domain-containing protein 40 – or CCDC40.



Russian poet Anna Akhmatova, by Amadeo Modigliani

Source: Wikipedia

All things have ideal lengths. Too long or too short is rarely a good option. This holds for people's height, the length of their tongue and of their toes. And, on the molecular level, for the width of a cell and the length of a domain within a protein complex. That is why it is necessary to have systems inside us that act as surveyors making sure that everything is kept within acceptable norms.

Flagella and cilia are found on the surface of many cells – or one-cell organisms – and are either used to move through fluids or to make fluids move past them, besides frequently showing acute sensory roles. While cilia have both motile and fluid-circulation functions, flagella are only used for cell motility. Both flagella and cilia form slim elongated structures that protrude from the main body of a cell but are still enclosed within its membrane, which thus forms a sort of sheath around them. Their molecular structure is identical and includes what is known as an axoneme as well as many accessory matrix proteins involved in assembling or disassembling the axoneme – a very dynamic structure. In all, it is thought that about 300 proteins are engaged in keeping flagella functional.

Characteristically, an axoneme is a long cylindrical structure made up of nine long outer filaments, or microtubules, which line its circumference, and one long central filament. The central filament is composed of two adjoining microtubules. The nine outer filaments are equally composed of two adjoining microtubules but are also flanked by what are known as dynein arms and radial spokes. It is not difficult to understand that such a complex structure demands a scaffold; something that makes sure that the different

parts are not only bound to each other but that it also happens in an orderly fashion.

CCDC40 is part of the scaffold. If CCDC40 is lacking or defective, the parts making up the axoneme lose their sense of proportion and direction, and the structure is disorganised and badly assembled. This results in a flagellum, or cilium, that is short and unable to beat. Such an occurrence in sperm brings about male sterility for example. So it is important for the axoneme to piece together properly. How then does CCDC40 manage this?

A closer look at the axoneme shows that its various components are arranged within a 96 nanometre longitudinal repeat whose extremities are marked by the positions of successive radial spokes. This 96 nanometre periodicity is broken if CCDC40 is defective. Why? CCDC40 stretches out along the axis of each microtubule, spanning exactly 96 nanometres. This is not only used as a sort of molecular ruler along each microtubule but acts as an anchoring site to which components can bind. CCDC40 is indeed involved in the correct assembly of an axoneme's dynein components.

If a flagellum – or cilium – is defective, many things can go wrong. Typically, a flagellum beats so that a cell can move around; cilia, on the other hand, will beat either for a cell's mobility or to move fluids that surround a cell. Be it cilia or flagella, if something is wrong with CCDC40, the effects can be devastating. Besides male sterility, if cells are unable to move, an organism's development can be impaired. As an example, the position of internal organs in an organism can be mirrored, a condition known as *situs inversus* in humans. Defective cilia are also the cause of primary ciliary dyskinesia, a condition which affects the lungs that are unable to brush away the mucus and clear the airways. Clearly, CCDC40 seems to be an ideal target for designing drugs that would be able to re-establish impairments its mutation causes.

Cross-references to UniProt

Coiled-coil domain-containing protein 40 homolog, *Chlamydomonas reinhardtii*: A8IQT2
Coiled-coil domain-containing protein 40 homolog, *Homo sapiens* (Human) : CCD40

References

1. Oda T., Yanagisawa H., Kamiya R., Kikkawa M.
A molecular ruler determines the repeat length in eukaryotic cilia and flagella
Science 346:857-860(2014)
PMID: 25395538

2. Pazour G.J., Agrin N., Leszyk J., Witman G.B.
Proteomic analysis of a eukaryotic cilium
The Journal of Cell Biology 170:103-113(2005)
PMID: 15998802

3. Werner-Peterson R., Sloboda R.D.
Methylation of structural components of the axoneme occurs during flagellar disassembly
Biochemistry 52:8501-8509(2013)
PMID: 24152136

proteinspotlight

Swiss Institute of Bioinformatics

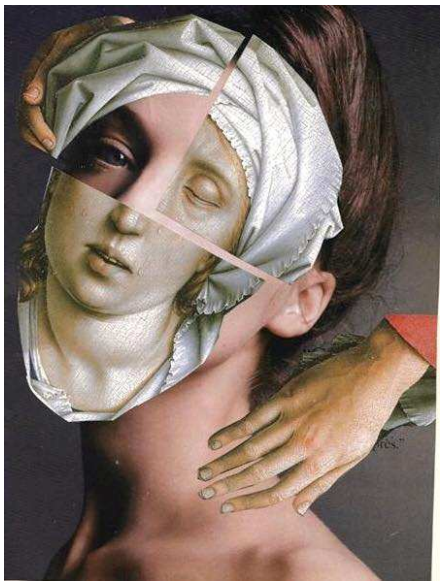
Protein Spotlight (ISSN 1424-4721) is a montly review written by the **Swiss-Prot** team of the **SIB Swiss Institute of Bioinformatics**. Spotlight articles describe a specific protein or family of proteins on an informal tone.
http://web.expasy.org/spotlight/

approaching happiness

Vivienne Baillie Gerritsen

We all know what happiness is. At least we know what it feels like to be happy. But the moment you begin to define it, things become complex. And trying to measure a feeling as ungraspable as happiness seems as far-fetched as weighing a poem. Yet understanding what sculpts high spirits is essential; as essential as understanding feelings at the other end of the scale – such as depression for instance. Over the years, scientists have demonstrated that there is undoubtedly a genetic component to happiness, as there is to depression. A gene discovered in the 1960s and known to be involved in antisocial behaviour has actually turned out to have its say in human happiness as well. Depending, that is, on how strongly the gene is expressed and an array of sociocultural, physiological and anatomical parameters. The gene is known as MAOA – for monoamine oxidase A – an enzyme that metabolizes neurotransmitters which each have their say in modulating our mood. And since MAOA is located on the X chromosome, it is argued that it influences happiness in women while, surprisingly, it has little incidence on men.

.



Approaching happiness, by Carol White

Courtesy of the artist

From a historical point of view, happiness was first defined as a condition that depended on events that were exterior to an individual, such as good luck or the prospect of a long-expected journey. As time went by and the notion of happiness began to revolve around something more self-centred, happiness became synonymous with a person's well-being. Cheerfulness and its making – or indeed its undoing – is currently understood as the intricate result of a sum of situations in which an individual is immersed. Namely: a person's age, gender, education, household income, marital status, employment status, mental disorder, physical health, relationship quality, religiosity, abuse history, recent negative life events and self-esteem. With the belief that happiness is something every human will naturally lean towards if doused in the "right" conditions.

Moods, however, are not solely due to circumstances outside an individual. Over the years it has become obvious that there is a genetic component to our traits of character too; something a mood can be built upon. So scientists turned to the genome to pin down genes that meddle with our humour. Antisocial personality traits are so diffuse and have such harmful effects on society in general that they have been intensively studied. And this is how, back in the 1960s, MAOA was identified as a gene that could having something to do with a person's antisocial behaviour. But these are notions to be handled with great care. Considerations such as these can be – and have been – used in legal procedures to lighten a sentence for instance. Genetic predisposition to antisocial behaviour does seem to exist, i.e. it can trigger off antisocial behaviour depending on a given life event. An example would be

child abuse. However, a person who carries the predisposition and has suffered child abuse does not necessarily develop aggressive behaviour. The difference is subtle but important to grasp.

So how can MAOA be involved in two states of mind that are situated on the opposite ends of the mood scale? First: a brief introduction to the enzyme behind MAOA. MAOA is a flavoenzyme that catalyses the oxidation of three neurotransmitters – dopamine, noradrenaline and serotonin – with the help of its cofactor flavine adenine dinucleotide (FAD). These three neurotransmitters are part of many different pathways amongst which those that tamper with mood regulation. Dopamine, as an illustration, is linked to our reward-motivated behaviour and has its say in depression and mania but also cognitive alertness. Noradrenaline is responsible for vigilant concentration and may well be involved in decision-making. And serotonin seems to play an important part in our feelings of well-being and happiness. To cut a long story short, when MAOA is expressed at low levels, it seems to mark a predisposition to antisocial behaviour. When it is expressed at higher levels, it could predispose us to happiness. And the fact that it is X-linked has made scientists wonder whether women are then more prone to happiness than men. But things have proved to be far more complex: an increased amount of MAOA does not make men happier.

MAOA could be compared to a corkscrew. The main body is made of two globular parts: one holds the cavity into which slips one of the three neurotransmitters (or indeed inhibitor); the second is the co-factor FAD binding domain. The corkscrew *per se* protrudes from the main body of the enzyme as a helix which is inserted into the mitochondrial membrane thus anchoring the enzyme on the mitochondrion's surface. To date it is not known why the mitochondrion was chosen as a place of mooring but anchoring is necessary for enzyme activity. The active site of MAOA is almost a sealed cavity whose opening is very narrow thus making it difficult for the substrate to squeeze through. However, when MAOA is inserted into the mitochondrial membrane, the whole structure becomes more supple and the substrate is able to slip into the active site.

Monoamine oxidases are important flavoenzymes that, over the years, have been linked to many psychiatric disorders. Happiness is far from a psychiatric disorder but it is a state whose detailed molecular description can help understand disabling mood disorders. MAOA is involved in antisocial behaviour but also in happiness, thus making it an ideal target for drug design. This said, happiness is dependent on many external parameters, and so will the efficiency of an MAOA-targeted drug. And to make things more complicated, studies have shown that the epigenetic methylation of MAOA could have a role in alcoholism and nicotine addiction in women. It is all so very complex and the more you read about happiness, what it means and how to measure it, the less you seem to know. And yet, the laugh of a child or a glint in an old woman's eye is able to express it in a fraction of a second.

Cross-references to UniProt

Amine oxidase (flavin-containing) A, *Homo sapiens* (Human): P21397
Amine oxidase (flavin-containing) A, *Rattus norvegicus* (Rat): P21396

References

1. Chen H., Pine D.S., Ernst M., Gorodetsky E., Kasen S., Gordon K., Goldman D., Cohen P.
The MAOA gene predicts happiness in women
Progress in Neuro-Psyhopharmacology & Biological Psychiatry 40:122-125(2013)
PMID: 22885141

2. Son S.-Y., Ma J., Kondou Y., Yoshimura M., Yamashita E., Tsukihara T.
Structure of human monoamine oxidase A at 2.2A resolution: the control of opening the entry for substrates/inhibitors
PNAS 105:5739-5744(2008)
PMID: 18391214

proteinspotlight

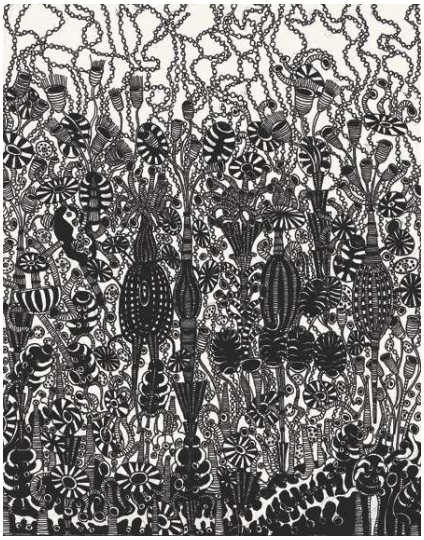
Swiss Institute of Bioinformatics

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

the smell of the sea

Vivienne Baillie Gerritsen

The chances a cloud will remind you of the ocean are slim. Yet without oceans, there would be fewer clouds drifting above our heads. This is because these huge stretches of sea are full of tiny creatures that produce a gas known as dimethyl sulphide, or DMS, which under certain circumstances has the faculty of initiating the beginnings of a cloud. These creatures – algae or bacteria – do not synthesize DMS for sculpting clouds; DMS happens to be the side product of a metabolite known as dimethylsulfoniopropionate, or DMSP, that phytoplankton probably require for buoyancy or protection, or perhaps even both. The air-borne DMS is responsible for the sometimes unpleasant pungent smell that is characteristic of the sea. The enzyme that is at the heart of this distinct sea scent was recently discovered in the marine phytoplankton *Emiliania huxleyi* and baptised DMSP lyase 1 – an algal enzyme that cleaves DMSP and in so doing releases the perfumed DMS into the earth's atmosphere.



Pen and ink drawing by Sue Bartfield

Courtesy of the artist

Emiliania huxleyi is the most abundant marine phytoplankton to float on the earth's oceans, and forms the massive green patches – or blooms – that can be seen from space. It is happy both in tropical and subarctic waters and is an important part of marine food webs. *E.huxleyi* is a single-celled eukaryotic phytoplankton surrounded by what is known as a coccolith, a term coined by the British comparative anatomist Thomas Huxley (1825-1895). Coccoliths are made of calcite. Their architecture is extraordinary and intricate, more often than not transparent and colourless. *E.huxleyi* coccoliths are made of calcite disks that were first described by Huxley and an Italian microbiologist Cesare Emiliani (1922-1995) – hence the name, *Emiliania huxleyi*. These calcite shells suggest some sort of protective role, yet no one knows it for sure. Coccoliths are perhaps a way of preventing zooplankton from grazing, or may just act as a physical barrier against viral or bacterial invasion and even harmful UV light. They could also simply be part of the physics that keeps the phytoplankton afloat or, in deep-dwelling species, a means of gathering and concentrating light for photosynthesis. These calcite shells are shed continuously and sink to the bottom of the sea where they form an important part of the deep-sea sediment. The white cliffs of Dover are an example of such sediment, deposited about 65 million years ago.

Besides its unique calcite shell, *E.huxleyi* produces – along with its fellow phytoplankton species and other marine microalgae – one of the most important and most abundant organic molecules in the world: DMSP. One billion metric tons of this organic compound are not only poured into the oceans every year but also turned over! DMSP is in fact so plentiful that it has become a signature molecule for life at sea. If so much is produced, surely it must have an important role in the life-cycle of phytoplankton and marine algae. No doubt. But to date nobody knows which. It could serve to protect organisms against osmotic stress. It could have a role in buoyancy. It could also be a means of communication in predator-prey interactions. First identified in 1948 in the red alga *Polysiphonia*, scientists recognised it to be the progenitor of DMS

which was already known to waft from seaweed. In 1956, a first enzyme able to cleave DMSP was discovered – one of many that were to follow, the most recent of which is DMSP lyase 1.

The ocean is awash with many different beasts able to split DMSP into DMS and acrylate by way of hordes of different enzymes. In the same way DMSP is a major compound of the ocean's sulphur sink, DMS is a major component of atmospheric sulphur – and together they represent key components of the ocean's sulphur cycle. DMSP lyases exist both in eukaryotes and bacteria, however *E.huxleyi* DMSP lyase 1 bears no resemblance to any other. It is tetrameric and belongs to the aspartate racemase superfamily while in marine bacteria, for instance, DMSP lyase belongs to the M24 peptidase family. DMSP lyase 1 cleaves DMSP, releasing in the process DMS – a gas – and acrylate. Crystal structures of various DMSP lyases suggest two active-site cysteines. The enzyme's activity is equally thought to be dependent on algal symbiosis or associated bacteria.

DMS let loose in the atmosphere plays a key role in the earth's sulphur cycle. Its tangy aroma serves as a chemical attractant which guides a variety of marine animals – such as sea birds, invertebrates and even mammals – to possible food supplies. Atmospheric DMS is a huge part of the global cycle of sulphur from the sea into the air, and then back into the sea or onto land when it rains. DMS even has a part to play in cloud formation, and hence in the world's weather. How? When the gas is flung into the atmosphere, it is *per forza* prone to oxidation. DMS oxidation products – DMSO – act as condensation nuclei causing water molecules to coalesce, leading to cloud formation or enhancing it. And where there are clouds, there are consequences on the local climate – or perhaps on an even more global scale. Clouds increase the reflection of solar radiation, sending it back where it came from, thereby influencing the earth's atmospheric temperature.

DMS has this particular tangy smell to it. Some say cabbage-like. Others, more poetically, refer to it as the smell of the sea. The characteristic scent is also released when cooking beetroot, asparagus and sea foods for example. An amusing anecdote: twenty years ago, a French chemist, Thierry Talou, was the first to identify DMS as the chemical responsible for the tangy scent, and used truffles and pigs to support his theory. Today, DMS is in fact used – in very small concentrations – as a food additive to impart a savoury flavour. Oxidised DMS is an important industrial solvent.

There is something hugely lyrical in a chemical component that is capable of creating a cloud. However, clouds spell climate change – a particularly sensitive subject today. Getting to know DMSP lyases better, the way they contribute to atmospheric DMS – which is dependent on the global distribution of bacteria and eukaryotes – and how they are affected by environmental parameters, or indeed how they affect the environment, are all important issues in our day and age. What is more, there are no doubt other forms of enzymes that act as DMSP lyases which are still unknown.

Understanding the physiological and signalling roles of DMS in the phytoplankton's resistance to viral infection, its fight against predators or its symbiotic interaction with other organisms will also help to give a fuller picture of a form of life that is so essential in the marine sulphur cycle besides having such a large impact on the earth's atmospheric conditions and hence the planet we are part of too. *E.huxleyi* was actually the creature that inspired Lovelock's famous Gaia hypothesis claiming that the earth's biochemistry and geochemistry are intimately intertwined.

Cross-references to UniProt

DMSP lyase 1, *Emiliania huxleyi* (Pontosphaera huxleyi) : P0DN21

References

1. Alcolombri U., Ben-Dor S., Feldmesser E., Levin Y., Tawfik D.S., Vardi A.
Identification of the algal dimethyl sulfide-releasing enzyme: A missing link in the marine sulfur cycle
Science 348:1466-1469(2015)
PMID: 26113722

2. Johnston A.W.B.
Who can cleave DMSP?
Science 348:1430-1431(2015)
PMID: 26113706

proteinspotlight

Swiss Institute of Bioinformatics

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

Organisational Members of EMBnet

Biocomputing Group
Belozersky Institute, Moscow, Russia

BMC
Uppsala Biomedical Centre, Computing Department, Uppsala, Sweden

Centre of Bioinformatics
Peking University, Beijing, China

CMBI
Radboud University, Nijmegen, The Netherlands

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

CSC
Espoo, Finland

EMBL-EBI
Hinxton, Cambridge, United Kingdom

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires, Buenos Aires, Argentina

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

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

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

ITICO
Information Technology Infrastructure for Collaborative Organizations, United Kingdom

KEMRI
Wellcome Trust Research Programme, Kilifi, Kenya

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

LCSB
University of Luxembourg, Luxembourg, Luxembourg

ReNaBi
French bioinformatics platforms network, France

SIB
Swiss Institute of Bioinformatics, Lausanne, Switzerland

TGAC
The Genome Analysis Centre, Norwich, United Kingdom

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

UMBER
Faculty of Life Sciences, The University of Manchester, Manchester, United Kingdom

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

EMBnet.journal
ISSN 1023-4144

Dear reader,

If you have any comments or suggestions regarding this journal we would be very glad to hear from you. If you have a tip you feel we can publish then please let us know. Before submitting your contribution read the "Instructions for authors" at http://journal.EMBnet.org/index.php/EMBnetnews/about and send your manuscript and supplementary files using our on-line submission system at http://journal.EMBnet.org/index.php/EMBnetnews/about/submissions#onlineSubmissions.

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

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

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