Bioinformatics training: a review of challenges, actions and support requirements

Maria Victoria Schneider, James Watson, Teresa Attwood, Kristian Rother, Aidan Budd, Jennifer McDowall, Allegra Via, Pedro Fernandes, Tommy Nyronen, Thomas Blicher, Phil Jones, Marie-Claude Blatter, Javier De Las Rivas, David Phillip Judge, Wouter van der Gool and Cath Brooksbank

Abstract

As bioinformatics becomes increasingly central to research in the molecular life sciences, the need to train non-bioinformaticians to make the most of bioinformatics resources is growing. Here, we review the key challenges and pitfalls to providing effective training for users of bioinformatics services, and discuss successful training strategies shared by a diverse set of bioinformatics trainers. We also identify steps that trainers in bioinformatics could take together to advance the state of the art in current training practices. The ideas presented in this article

Corresponding author. Maria Victoria Schneider, EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. Tel: +44(0)1223494448; Fax: +44(0)1223492468; E-mail: vicky@ebi.ac.uk

Maria Victoria Schneider is the Training Programme Project Leader in the Outreach and Training Team at EMBL-EBI where she is responsible for developing and coordinating the in-house and external training activities of EMBL-EBI.

James Watson is the Senior Scientific Training Officer in the Outreach and Training Team at EMBL-EBI, prior to this he was a postdoctoral researcher in the Thornton group looking at protein structure-function relationships.

Teresa Attwood is professor of Bioinformatics at the University of Manchester; she coordinates and teaches the introductory module of the Bioinformatics Masters course, and has written several bioinformatics reference and text-books.

Kristian Rother is teaching structural bioinformatics, programming, and software development to bioinformaticians at the UAM Poznan. He is also contributor to the BioPython and PyCogent software libraries

Aidan Budd is a computational biologist at the EMBL in Heidelberg, working in the Gibson Team. Working on a range of projects involving short protein linear motifs and phylogenies, he is also responsible for the training and outreach activities of the group.

Jennifer McDowall is a Scientific Training Officer in the External Services Team at EMBL-EBI where she is responsible for developing and teaching patent-related training activities.

Allegra Via is a researcher in Bioinformatics and Computational Biology at Sapienza University of Rome (Italy) and Adjunct Professor in Computer Programming for Bioinformatics at Tor Vergata University of Rome (Italy).

Pedro Fernandes is the creator and coordinator of the Gulbenkian Training Programme in Bioinformatics, with over 80 courses and 1600 attendees, held at the Instituto Gulbenkian de Ciência in Oeiras, Portugal on a regular basis, since 1999.

Tommi Nyrören is a development manager at the Finnish IT Center for Science—CSC where he is responsible for developing services and technology for biomedical sciences.

Thomas Blicher is associate professor at the NNF Center for Protein Research (University of Copenhagen) and Center for Biological Sequence Analysis (Technical University of Denmark). He is involved in teaching undergraduate and graduate students as well as developing methods for use of web-based tools in teaching.

Phil Jones is a senior software engineer on the InterPro project, a database of protein families, domains, regions, repeats and sites at the EBI. He has a background in both biological sciences and software engineering, and holds a PGCE in secondary level science education.

Marie-Claude Blatter is part of the communication team of the Swiss-Prot group (SIB Swiss Institute of Bioinformatics), where she coordinates teaching/training programs and also activities for the public at large.

Javier De Las Rivas is PI Scientist at the Cancer Research Center (CiC-IBMCC, CSIC/USAL) in Salamanca (Spain) leader of the Bioinformatics and Functional Genomics research group.

David Phillip Judge coordinates and runs wide range of undergraduate and graduate/staff courses at the University of Cambridge.

Wouter van der Gool organized bioinformatics training workshops and tutorial sessions for VIB's Bioinformatics Training and Services Facility (BITS).

Cath Brooksbank, Ph.D. is the Head of Outreach and Training at EMBL-EBI where she coordinates EMBL-EBI's public relations and user training programmes.

derive from the first Trainer Networking Session held under the auspices of the EU-funded SLING Integrating Activity, which took place in November 2009.

INTRODUCTION

Bioinformatics is a rapidly evolving, interdisciplinary field. Advances in molecular technologies such as genome sequencing, microarray platforms and high-throughput technologies have resulted in a surfeit of data, in response to which the development of bioinformatics software tools and repositories has advanced rapidly, consuming a huge amount of time (and hence money) for both their creation and maintenance. As a consequence, bioinformatics training has had to evolve and grow in parallel with the technology, in order to keep users up-to-date. In addition to these technological changes, bioinformatics is an interdisciplinary subject, drawing expertise from a range of fields, from biology to chemistry, mathematics, physics and computer science. The evolution of the field is seen in terms both of the end-users who require training (who vary in their knowledge of bioinformatics and in their computer literacy), and of the trainers themselves. This leads to a training paradox: end users need training in bioinformatics tools and databases, and yet many trainers lack formal bioinformatics training and come from diverse scientific backgrounds; this makes creating universal, comprehensive training programs difficult. This group of co-authors reflects this situation, ranging in background from pure physics and chemistry, to evolutionary biology, biophysics and pure bioinformatics (i.e. those with a strong background in programming but not necessarily formal training in biology).

Bioinformatics training is challenging not only because of its interdisciplinary nature, but also because of the extensive technological changes witnessed in the field during the last two decades. As the types and quantity of data have changed, new databases and software have been developed to manage and analyse them, while many older resources have either disappeared or become redundant. Initially, sequence databases were of a manageable size, were linked to few external repositories, and were associated with a limited number of analysis tools. However, with the growing availability of digital data, the number of databases has increased exponentially, as have the tools developed to search and analyse the data they contain. Consequently, a basic introduction to bioinformatics now requires knowledge of nucleotide and protein sequences, protein structures, transcriptomic and proteomic data, chemical and systems biology data, data repositories, search algorithms, ontologies, and often a rudimentary knowledge of programming and statistics. The difficulty in tailoring training courses to accommodate such a diversity of knowledge domains is compounded by the diversity in the scientific backgrounds both of end-users and of trainers themselves. Consequently, narrowing the focus of training sessions to satisfy the range of possible student interests and learning objectives, while taking account of diverse trainer abilities, can be especially challenging.

The ideas presented in this article derive from a SLING (Table 1) meeting held in November 2009 concerning Bioinformatics Training. The meeting aimed to identify common key challenges and pitfalls, in addition to successful training strategies shared by a diverse set of bioinformatics trainers; moreover, we aimed to identify steps we could take together to advance the state of the art in current training practices. In particular, we sought to identify potential pilot projects we could initiate and develop ourselves. With this article, we aim to share the results of this meeting with a wider audience, and to stimulate discussion of everyday issues related to bioinformatics training. The body of the article therefore summarizes the perspectives that arose during the meeting regarding the current challenges and support requirements that face bioinformatics trainers, and offers a number of practical suggestions relating to the actions needed to address them.

In this article, we specifically concentrate on training bioscientists to use bioinformatics repositories and tools optimally—we do not discuss training in the context of computer programming, nor do we discuss bioinformatics education for undergraduate or master's students in general. Of course, training is required here too; however, the subject of graduate education is much wider, and our intention here is to focus on the more pragmatic activities involved in the delivery of the kind of short, sharp, effective professional development courses

Table I: Five examples (three from current and two from past) of pan-European initiatives that involve bioinformatics user training

Name	Description
SLING	The Integrating Action 'Serving Life-science Information for the Next Generation' aims to support Europe's exploitation of bio-molecular information, in part through extensive pan-European user training to facilitate exploitation of the information. SLING organizes training roadshows with an emphasis on training bioscience researchers in the new EU member states (http://www.ebi.ac.uk/training/roadshow/); it also organizes networking sessions for trainers, and ultimately aims to build a collection of shared training materials to support these trainers http://sling-fp7.org/.
ELIXIR	The purpose of ELIXIR, the European Life-Science Infrastructure for Biological Information', is to construct a sustainable infrastructure for biomolecular data and related information in Europe, to support life science research and its translation to medicine and the environment, the bio-industries and society. Whilst ELIXIR's broad aim is to improve access to biological data for Europe's life science community, its training strategy is *to improve accessibility*, by empowering European researchers to make effective use of the data. Further information on ELIXIR can be found at http://www.elixir-europe .org/page.php. The Training Strategy Report is at: http://www.elixir-europe.org/bcms/elixir/Documents/reports/ WPII-Training.Strategy.Committee.Report.pdf
EMTRAIN	EMTRAIN aims to establish a sustainable, pan-European platform for education and training (E&T) covering the whole life-cycle of medicines research, from basic science through clinical development to pharmaco-vigilance. Its scope extends far beyond bioinformatics training, but it aims to overcome many of the challenges outlined in this review, including the harmonization of professional development programmes and the optimal use of teaching and learning methodologies: http://www.emtrain.eu/index.php/home
BioSapiens	The BioSapiens Network of Excellence, funded by the European Union's 6th Framework Programme and completed in 2009, created and organized the European School of Bioinformatics, which was held each time in a different European location and directed mostly at training newcomers to bioinformatics: http://www.biosapiens.info/page .php?page=esb
EMBER	Created by an EC-funded consortium tasked with developing a suite of multimedia bioinformatics educational tools, EMBER comprises a self-contained, interactive, web tutorial in bioinformatics and the equivalent stand-alone course on CD-ROM. A text book is in preparation. Using conventional text, coupled with Web- and CD-based media, EMBER aims to ensure that students for whom Internet access is not optimal also have access to the same fundamental level of bioinformatics educational materials. http://www.bioinf.manchester.ac.uk/dbbrowser/ember/ memberspage.html

in which the majority of us are already heavily involved. Last but not least, improvement in this area is most likely to be welcomed by the large group of scientists who seek to benefit from being able to use bioinformatics tools and services more efficiently.

USEFUL DEFINITIONS

First, it is important to define what we mean by 'bioinformatics training'—in particular, we seek to distinguish between the definitions of 'teaching' and 'training'. Throughout this article, we refer to 'training' where the main aim is to deliver skills to an audience in such a way that will allow them to optimally use bioinformatics tools and databases delivered in relatively short, practical, focused courses. By 'teaching', we mean longer, theoretical courses, with broader coverage, and where the main deliverable is knowledge and understanding of fundamental concepts.

It is also important to identify different groups of trainees we need to target which fall into three main categories:

(1) End users: these include users of the web interfaces for bioinformatics resources who need to access bioinformatics resources for their research.
(2) Developers/programmers: these include bioinformaticians who are developing databases and software tools; computational scientists; and other non-biologists (theoreticians, physicists, engineers, statisticians).
(3) Trainers: individuals from a range of disparate backgrounds who are involved with the production and delivery of training in bioinformatics.

DEFINING THE CHALLENGES AND PRIORITIZING

So, what are the challenges in delivering effective bioinformatics training? We identified several aspects

that need to be considered, but some specific points emerged as priorities:

- differences in trainee backgrounds;
- the preparation/availability of appropriate materials, keeping them up-to-date, and providing appropriate examples and/or use cases;
- identifying and understanding key concepts—what is it we should be training in?
- the need to provide support to trainers, in order to build and maintain a critical mass of training expertise that can meet the demands of a growing and diversifying user base;
- the availability of funding for training, and the recognition that bioinformatics training is not only a worthwhile endeavour, but also an essential one.

Funding is a complex issue, which we do not address in this article; nevertheless, it should be recognized that delivering good bioinformatics training requires adequate funds (such as those provided to the SLING Integrating Action by the European Commission), and sustainable funding for bioinformatics training may be a topic for future SLING meetings. Examples of existing and recently completed pan-European initiatives that involve bioinformatics user training are mentioned in Table 1.

DIFFERENCES IN TRAINEE BACKGROUNDS

Heterogeneity of trainee backgrounds is common, owing to the diversity of research areas in which bioinformatics plays a part. Trainees often differ substantially in both biological and computer knowledge, and their levels of familiarity with bioinformatics tools can vary widely (e.g. from 'basic' users to 'advanced' developers).

Striking the right balance between providing sufficient information to train individuals how to optimally use bioinformatics databases and tools, and ensuring that trainees are not overloaded with too many concepts, represents a considerable challenge.

In practice, bioinformatics trainers need to be able to offer training options that are tailored to the abilities of individual trainees (or risk leaving parts of their class overburdened or bored). In trying to accommodate such diverse abilities, a variety of different exercises could be prepared, but this would increase the necessary effort involved in material

preparation, and it is questionable whether a simplistic easy/hard scheme would adequately address such a diversity of training needs anyway. A possible solution is to use multi-stage learning aids that allow trainees to carry out the same exercise and adjust their difficulty to their individual learning speed [1, 2].

Another option is to pre-screen the audience prior to a particular training event, ideally with sufficient time to enable the trainer to tailor the materials to match trainee expectations. Collecting information about trainees upon registration can offer one way to enhance the effectiveness of training events. Typical questions could be:—What do you expect to gain from this course?; What is the relevance of this course to your research project and how will your research benefit from your attendance on this course?; Have you ever used bioinformatics resources before? If so, which ones?

MATERIALS: CHALLENGES

The bioinformatics training landscape is constantly changing: new tools become available, existing tools are changed, new user groups (with new sets of previous experience and knowledge) become candidate trainees. Therefore, bioinformatics trainers need to constantly adapt to this backdrop. A particular challenge relates to 'use cases', which can either be demonstrated by the trainer or form the basis of practical 'hands on' exercises. Stable databases to underpin specific exercises could overcome several of the hurdles mentioned here, and could prove especially useful, for example, for training courses that involve gene prediction and/or sequence searching, where one is teaching the principles and is not concerned about the exact data set retrieved.

Preparation of training materials is also often made more challenging by a lack of suitable documentation relating to the available tools and databases themselves. This makes it difficult for trainers to incorporate new information and new analyses into their presentations; consequently, this limits their ability to expand their materials to explore additional resources (especially when already faced with the difficulty of keeping their current materials up-to-date).

Bioinformatics training materials are often narrowly focused on a specific database or tool(s), rather than providing an overview of the field

being covered. As a result, trainers often fail to expand their materials to include and explore other repositories and tools that could be of value to end-users, hence limiting the knowledge and capabilities of their trainees.

One solution might be to produce up-to-date Web-based training materials that would allow end-users both to explore a range of current bioinformatics platforms, and to reinforce what they have learnt on short face-to-face courses, by providing the opportunity for additional learning, for exploring concepts and/or for trying hands-on exercises. The availability of such resources for self-education would be highly valuable and could provide long-term benefits, as the availability of Web-based courses introduced by a course tutor could be quickly spread to the research community both via typical Web-based social networking technologies and through word of mouth from course attendees.

Ultimately, the question we need to address is: why are so many trainers creating bespoke materials for bioinformatics training, rather than openly sharing them? Many bioinformatics trainers experience the same frustrations; we therefore believe that it would be beneficial to establish a Trainers' Network by means of which trainers could liaise with, and get help from, other trainers, and via which materials, ideas and best practices could be readily and rapidly communicated.

MATERIALS: ACTIONS

What types of training materials might such a network be able to make available in a centralized repository of training materials? These might include the following:

(1) Downloadable PowerPoint presentations could be useful both as teaching aids for trainers and as review material for trainees after they've attended a course. Existing examples include:
   (a) Examples from the EMBL Gibson group: http://tinyurl.com/gibsonTraining
   (b) Example from SIB: http://education.expasy.org/cours/Basel09/
   (c) Example from Hands on Courses at the EBI: http://www.ebi.ac.uk/training/handson/details/xcourse_091019_dip.html

   (d) Example from CSS courses: http://www.csc.fi/english/research/sciences/chemistry/courses/ gmx2007.

(2) Lesson plans would improve an individual trainer's ability to re-use another trainer's PowerPoint slides. Such plans would include details of 'how' one teaches the materials, adding clear, sound notes on the main teaching points for each slide. These could be included as notes in the PowerPoint slides themselves.

(3) Well-documented/annotated images and figures, in an editable format, could help to reduce material preparation times.

(4) A centralized list of Web-locations of useful training materials, with a **star-rating** system, would help trainers to choose relevant materials. Lists could also be provided for relevant books, references and reviews.

(5) Working tutorials would provide trainers with access to a pool of up-to-date and tested hands-on materials. The availability of such a resource could save trainers a considerable amount of time creating and updating tutorials and exercises (e.g. http://www.ember.man.ac.uk). This is not a trivial endeavour, owing to the fast pace of change in bioinformatics discussed above.

(6) Podcasts and recordings: there are now various well-known online collections of videos (e.g. JOVE, SciVee, YouTube; iTunesU, LabTV), and various initiatives to promote podcasts (e.g. Cell Press' 'Article of the Future'). However, few, if any, of these are dedicated to bioinformatics training. It might therefore be worthwhile to explore, and perhaps to pilot, using these in a future training event.

(7) Training methods: a dedicated area with which to share training methodologies and experiences would be a valuable asset. It is hoped that this article will help to launch such a training repository, and we invite those who are interested in expressing their views to contact us via e-mail to: btn@ebi.ac.uk

One approach is the development of 'problem spaces'—units of didactic material prepared in such as way as to allow future change [3]. A typical problem space consists of (i) biological questions, (ii) a

data body (e.g. query sequences, a database to search), and (ii) tools and Web resources that are potentially useful to answer the questions. The essence of this approach is that, once created, the individual components of a problem space become interchangeable (e.g. replacing one tool by an improved one). Moreover, it is possible to create many possible exercises from a problem space differing in length, complexity, and target audience, because no such assumptions are made in the problem-space definition *per se*.

PROVIDING SUPPORT TO TRAINERS

Bioinformatics specialists are often asked to help their user communities, based at a variety of institutes and universities, to develop local bioinformatics training initiatives. Providing such support not only benefits trainers and those developing bioinformatics tools and databases, but also benefits the wider scientific community when trainees take what they have learned to teach and support others. This concept of *trainer support* is important to keep in mind when developing bioinformatics courses, as it recognizes two types of trainee: those who wish to learn how to use particular tools/databases in order to facilitate their day-to-day work, and those wishing to learn how to use them in order to train others. Thus, we want to start developing and delivering courses with the dual aim of training trainees how to optimally use bioinformatics resources, and of providing them with appropriate materials and information to allow them to pass on to their peers the knowledge and understanding that they acquired during the course. To ensure long-term benefits of such events, it is crucial both to be able to rely on support from others and to be able to find adequate, relevant and up-to-date training materials as necessary. Ideally, if we are to derive the maximum benefit from more strategic and organized, community-based approaches, trainers should be able to answer the following questions before embarking upon teaching a course: (i) what do the trainees expect from the course?; (ii) What are the learning objectives of the course?; (iii) How can the subject be subdivided into smaller, more easily consumed units?; (iv) How do I create a lesson plan?; (v) How do I encourage trainees to ask questions?; (vi) How do I prepare engaging, interactive presentations?; (vii) How do I prepare written exercises?; (viii) How can I save time preparing materials when time is limited?; (ix) How do I get constructive feedback from the trainees?; (x) How do I address that feedback next time I teach the course?

BEST WORKING PRACTICE: LEARNING FROM THE TRAINEES

Course feedback is a valuable way of gaining information on course contents, training methods and much more. In order to cross-compare teaching methods and programs at different institutes, we feel that it would be valuable to share this feedback. A common set of feedback questions could enable bioinformatics trainers to gather feedback in a consistent way from training events, which in turn could catalyse the spread of best working practices. The following minimal set of questions was selected as a draft recommendation for those delivering bioinformatics training: (i) would you recommend this course?; (ii) What would you like to see done differently?; (iii) What was the best thing about the course that should definitely be included in a repeat of the course?; (iv) How could the materials provided be made more useful?; (v) Would you prefer a different balance in the proportion of lectures to practical sessions/exercises?; (vi) Were there suitable opportunities to have discussions with trainers?; (vii) Rate (a) the quality of the trainer (b) the relevance of the course for your own research on a scale of 1–5 (1—poor; 5—excellent).

In general, when dealing with negative feedback, it is useful to ask responders to provide information on the reasons for their dissatisfaction, and what could be done to improve the course.

BIOINFORMATICS COURSES AND EVENTS: WHAT, WHERE AND WHEN

Getting a clear overview of what bioinformatics training is available and for whom is a difficult task. A central point that collects such data from event organizers would be a valuable resource

that could be used by course organizers and trainers. Although several efforts have been made in this direction (see examples below), ensuring that information is accurate and reliable, yet secure from malicious attack, is not a trivial task; as it requires a gatekeeper who can commit enough time to filter information. Nevertheless, we feel that this would be a beneficial development, perhaps more so for potential trainees, who could get an idea of what events are taking place near them and, in the event of missing a relevant training opportunity, be able to identify another one somewhere else.

Awareness of course availability would also help course organizers and trainers to more efficiently evaluate what training is required locally, thereby avoiding course redundancy. Some efforts along these lines have been made already, and below we list some examples of initiatives that aim to share information about bioinformatics courses where lists of available courses, their locations, websites, and so on, are presented. Nevertheless, a centralized place from which to find coordinated information regarding courses is still lacking.

- The International Society of Computational Biology's Student Council distributes training information via its Regional Student Groups: http://www.iscbsc.org/; the International Society of Computational Biology also maintains a list of online courses at http://www.iscb .org/online-courses:
- Shared Google calendars: some training programmes offer a variety of courses on a regular basis, often with similar versions of the same course at different locations and times (Cambridge UK, Oeiras PT). Although the sharing of Google Calendars between coordinators started as a means of avoiding course collisions, it soon became a useful tool for potential course candidates seeking training opportunities. A project is currently under way to build a web resource that accesses and merges these calendars, which can still be individually maintained by the course organizers, thereby offering additional useful information in several aspects that are important when selecting training events: contents, teaching staff, time, location, fees, subsistence cost, lodging options, availability of financial support, etc.
- Bioinformatics.org (http://www.bioinformatics .org/wiki/Educational_services) devotes part of its efforts towards the promotion of bioinformatics education.
- EMBnet, the European Molecular Biology Network (http://www.embnet.org/) [4], has an Education and Training Committee devoted to the collation and dissemination of training information and organization of bioinformatics courses.
- The EMBL-EBI has started a similar project under ELIXIR. Users can view a map of bioinformatics user-training courses in Europe: http://www .elixir-europe.org/page.php?page=user_training_ map or the list of courses: http://www.elixir-europe.org/page.php?page=user_training_modu les. For more information go to http://www .elixir-europe.org/page.php?page=wp11

A START: SUPPORTING EACH OTHER

A recurring theme of our first SLING trainer networking session was that we all share and need to address the same problems. We view the ideas presented in this article as the first steps needed to build a community of bioinformatics trainers with a shared set of goals and a defined action plan. We acknowledge that other such groups, such as the EMBNet community, have done much to support bioinformatics training initiatives; what is new is that we are now in a position to tackle these issues with a mature understanding of many of these issues, gained from a wide range of training projects. Several of these are discussed in the ELIXIR Training Strategy Report at http://www.elixir-europe.org/page.php?page=reports.

Many of the recommendations discussed in the ELIXIR report will require significant funding for new infrastructure, but there are some simple, community-based projects that could be started immediately. To this aim, the co-authors of this article have started a bioinformatics training network using Google groups. Long-term plans involve developing a more optimal solution that would enable fast, easy access to the materials, information and ideas discussed above. The aim is to provide a centralized space to share materials, to list training events (including course contents and trainers), to list suitable venues for hands-on training in bioinformatics, and to share and discuss training experiences. We encourage the bioinformatics community to join this initiative and to enrich it with their active participation.

Key Points

- Bioinformatics Training presents a series of challenges common to all those delivering training of databases and tools at different levels as well as in different areas.
- Four main aspects are particularly relevant:—differences in trainee backgrounds;—dynamic nature of resources and data which leads to dynamic nature of training materials and updates;—support for trainers and appropriate funding.
- A good start to encompass these challenges is for trainers to support each other, sharing experiences as well as materials; this article represents an initiative in this direction.
- Bioinformatics Training needs a long term solution providing a centralized space to share materials, to list training events, to list suitable venues for hands-on training in bioinformatics, and to share and discuss training experiences.

References

1. Christensen C, Hansen AJ. *Teaching and the Case Method.* Boston, MA: Harvard Business School Publishing Division, 1986.
2. Forschergruppe Kassel. Aufgaben mit gestuften Lernhilfen [Stepped supporting tools for problems in school science], Lernchancen, 2004;**7**(41):38–39.
3. Allen D, Tanner K. Approaches to cell biology teaching: learning content in context -problem-based learning. *Cell Biol Educ* 2003;**2**:73–81.
4. Attwood T, Selimas I, Buis R, *et al.* Report on the EMBER project – a European Multimedia Bioinformatics Educational Resource. *Biosci Educ* 2005;**6**:Article 4.