EMBO reports

meeting point

Biggest challenges in bioinformatics

Jonathan C. Fuller, Pierre Khoueiry, Holger Dinkel, Kristoffer Forslund, Alexandros Stamatakis, Joseph Barry, Aidan Budd, Theodoros G. Soldatos, Katja Linssen, Abdul Mateen Rajput & HUB Participants

The third Heidelberg Unseminars in Bioinformatics (HUB) was held on 18th October 2012, at Heidelberg University, Germany. HUB brought together around 40 bioinformaticians from academia and industry to discuss the 'Biggest Challenges in Bioinformatics' in a 'World Café' style event.

The Heidelberg Unseminars in Bioinformatics (HUB) are participant-driven meetings. As Wikipedia notes (as of 18th January 2013), "the term 'unconference' [unseminar] has been applied, or self-applied, to a wide range of gatherings that try to avoid one or more aspects of a conventional conference, such as high fees, sponsored presentations, and top-down organization". At HUB, we have experimented with several formats to encourage participation in the meetings. For the third HUB, the organizers chose to discuss the 'Biggest Challenges in Bioinformatics'. We adopted a format called the 'World Café', with participants engaging in a series of short (approximately 20 min) conversations in groups of between four and five. After each round of conversation, the table host remained in place and the other participants visited another table with a specific topic of their choice. The table host then summarized the previous discussion to the new participants who added their ideas to the conversation. After a series of these conversations, the ideas were reported to the whole group in short form. We decided to take the idea one step further and share our deliberations with the wider scientific community through this article, which was written collaboratively at http://www.hub-hub.de.

This article therefore summarizes some of the main discussions around the biggest challenges in bioinformatics. The summaries are not intended to be comprehensive reviews of the state-of-the-art, but rather to reflect the discussions that took place at the meeting. As such, there are probably conflicting views on some areas, particularly relating to the question 'what is a species'?

Data deluge

The continuing development of high-throughput measurement techniques is leading to a constant increase in the volume of data available for analysis. For each piece of biological information in a measurement, any number of technical variables can be included, and it is not always clear which of these are relevant. Biological conclusions come only after multiple steps of quality control, filtering, normalization and processing have been undertaken, all of which might involve *ad hoc* cut-offs, settings and procedures. Unless all relevant information is retained, full reproducibility is not guaranteed. At the same time, the consideration of storage and processing, as well as the transfer of data between collaborating partners, is necessary and often limiting.

> **One main challenge […] is how to decide which data sets to archive and which to discard**

Even with sophisticated methods for information reduction, data-archiving costs can be considerable. One main challenge that the HUB meeting addressed is how to decide which data sets to archive and which to discard. The participants proposed that a benefit–cost ratio could be applied to each dataset to help to guide such decisions. Such a quantitative score would ideally take into account the estimated 'scientific value' to the community, the cost of archiving and the cost of recreating an equivalent data set. A formula to determine the cost–benefit ratio was even proposed; however, it became clear that an *a priori* measure of the scientific value of the data set was required to specify a truly useful cost–benefit ratio score. As the suggestion of the formula was only made at the end of a session, determining an absolute measure for 'scientific value' remained an open question, given that several participants considered this aspect hard to define, subjective and occasionally biased. A consensus from HUB participants was that quantitative scores should act only as an aid to those managing the data. Some data sets can never be recreated and so should arguably be archived even when their value to the community is low. Similar ideas to those discussed in this section have been explored in the context of DNA archiving [1].

Knowledge management

The fact that information is available is not sufficient; it also has to be made accessible and useable. Barriers to its use include a lack of standardized formats, a lack of common interfaces to data, inconsistency in identifiers for biological entities, insufficient support for data-exchange frameworks and insufficient visibility. Countless PhD projects involve attempts at solving this problem by introducing a new, common standard to eliminate inconsistencies. Unfortunately, unless these projects result in widely used, established standards, they just add additional layers of obfuscation.

A solution might be to accept the presence of parallel interfaces, whilst ensuring that new resources are available through as many formats as possible—for example, flatfile download, BioMart access and the Distributed Annotation System. Users can benefit from these resources according to their personal preferences. As with systems biology, redundancy of access to data can bring robustness to the tools using the data.

Predicting not just explaining

There has been much debate over the idea of data-driven hypotheses—the idea that the collection of data comes before the statement of a testable hypothesis. Indeed the data leads the researcher to the hypothesis itself. In this context, the HUB participants discussed hypothesis against data-driven science, tool integration and negative gold standards.

At the meeting, much discussion focused on whether bioinformaticians can effectively formulate new hypotheses before experimental work takes place. In many fields, experimentalists can generate new data faster than bioinformaticians can make informed predictions, which is especially true if bioinformaticians are approached as an afterthought once the experiment has been performed. Traditionally, experimental design and hypothesis formulation might have been done by experimentalists, with bioinformatics often seen as a *post hoc* analysis step. With the advent of high-throughput methodologies, the hypotheses of which are sometimes developed post-experiment, there can be bias in data interpretation. We felt that input from bioinformatics at early project stages could help to formulate hypotheses and enable experimental design with the appropriate statistical power to confirm or deny the hypotheses, and ensure that data interpretation bias is minimized. Further insights from data exploration can be used to generate new hypotheses for an iterative cycle between hypothesis generation, data collection and hypothesis testing. To some extent this is attempted in large systems biology projects such as the Virtual Liver Network (www.virtual-liver.de). Here models are built from existing biochemical knowledge, hypotheses are generated from the model and experiments are designed to test these hypotheses in an iterative cycle, the new results of which are fed back into the initial model. One method to improve this collaboration might be pressure from funding agencies encouraging cooperation between bench scientists and bioinformaticians at the grant-writing stage, or alongside investment in facilities that will produce high-throughput data, which probably requires bioinformatics analysis.

> **…much discussion focused on whether bioinformaticians can effectively formulate new hypotheses before experimental work takes place**

Further to the above improvements, the general integration of bioinformatics tools into standard tools such as molecular viewers, work flow managers and modular pipelines should allow bioinformaticians to work faster and more effectively. For example, in the field of structural bioinformatics and simulation, the integration and reuse of tools within molecular viewers, as well as the use of libraries such as Open Babel, in projects for which several file formats need to be read and written, is proving fruitful. To improve the adoption of these strategies, bioinformatics education needs to focus on the use of these types of helper libraries as a first choice, rather than as an afterthought when new code has been developed.

The meeting participants also noted that in some disciplines the existence of high-quality 'negative gold standards' is important—such as, decoys in protein or ligand docking studies, lists of non-drugable proteins or sets of negative protein–protein interactions. In this context, access to high-quality data sets for new training methods is of crucial importance to ensure that findings can be generalized sufficiently to predict new observations, for example, in drug-discovery and personalized medicine.

Personalized medicine

Advances in sequencing technologies mean that personalized medicine will almost certainly become a scientific reality, although the commercial and public health benefit remains to be defined and tested. However, before any new method of diagnosis can be implemented, important technical choices must be made, especially in regards to the decision to follow either a holistic approach—correlating different causative associations—or a reductionist one—targeting specific biomarkers. Initially, among HUB participants, some leaned more towards holistic approaches and others towards reductionist approaches. Whilst holistic approaches might be powerful for generating new knowledge and understanding the behaviour of an entire network or system, translation of this knowledge into the clinic might require the subsequent reductionist distillation of findings. We concluded that this choice depends on our understanding of the molecular and the genetic basis of each disease. In both cases, however, stringent quality controls and gold standards are needed to reduce bias, for example due to the use of different techniques or sequencing platforms.

In addition to the scientific questions, where some leaned more towards system biology approaches and others towards focusing on specific targets, the participants discussed the ethical and privacy issues relating to the confidentiality of the data generated and the impact it might have on the job market and insurance companies. This inevitably led participants to wonder about more practical concerns, such as whether the proper social structures are in place to provide fair access to this new medicine and to help people to cope with such definite diagnoses. Such requirements include providing appropriate education and promoting trust in scientific methods. Answers to these important questions will probably determine whether the financing of new research to overcome the technical challenges—including the secure storage of information—will be available. Fortunately, some governments and government agencies have started to implement regulations related to privacy and ethical issues in personalized medicine.

> **…in some disciplines the existence of high-quality 'negative gold standards' is important…**

In a best-case scenario, understanding the complexity of a disease can lead to its cure. This requires the systematic study of the complex molecular interactions that contribute to human health—drugs, proteins, pathways, mutations and others. The HUBs' point of view about these aspects tended to be optimistic in the main, especially given recent applications, technical improvements, bioinformatics advances and increased financial support for personalized health-research projects, such as the 1000 Genomes Project, large-scale open access databases of pharmaceutical side-effects, high-throughput experimental data repositories and increasingly predictive tools, including for treatment outcomes.

What is a species?

Deciding how to define 'species' is, in general, of fundamental importance to biology. More pragmatically, the HUB participants share the concerns of others about the massive global loss in biodiversity and its impact on environments and people across the world. This worldwide calamity highlights the importance of understanding the basic mechanisms and processes that influence biodiversity, including speciation.

Whilst defining genetic markers that delineate species is relatively easy among groups that sexually reproduce, such as mammals, it can be quite hard in others. Moreover, it is not clear whether it is possible or even desirable to define a universally accepted species concept and whether in the case of bacteria, for instance, a different concept for defining an 'evolutionary unit' might be more appropriate. It is apparent, however, that to identify a species or other evolutionary unit reliably, distinct types of data from different sources need to be integrated. These data include molecular, morphological, ecological and environmental information. Thus, a crucial part of the challenge is to develop methods and theories that allow the integration of large, ever-expanding, heterogeneous data sets to address questions of species delimitation. The term 'integrative taxonomy' was previously introduced to denote this emerging discipline (for a review see [2]). It was clear to participants that 'integrative taxonomy' is a huge challenge for the field, and that incorporating distinct data sources seems to be the only feasible solution. However, it is unclear how to incorporate and weigh all of the distinct data types into a unified integrative taxonomic model.

There was also an intense debate regarding approaches to delimiting species by using molecular data—for example, environmental sequencing—on the basis of mostly arbitrary or empirical sequence similarity cut-offs. One potential improvement might consist in inferring and deploying variable empirical thresholds for different parts of the underlying species tree. In this context, new statistical methods based on coalescent theory [3] were mentioned as potentially promising solutions.

It was also noted that data sets frequently contain a mix of sequences from within a species, and from across a group of species, such that species delimitation and population genetic methods can no longer be separated from phylogenetic inference. Instead, once again, integrative approaches and analysis methods need to be deployed.

Inferring the tree of life

The issue of orthology assignment, which represents the first step for assembling a multi-gene or whole-genome multiple sequence alignment (MSA), was mentioned as an unresolved problem, since objective criteria for performing this task do not exist, although reasonably engineered pipelines do. The number of universal (housekeeping/core) genes common to most organisms on earth was mainly perceived by participants as being too small to infer a robust tree of life reliably. Therefore, we coined the term 'gene sampling pyramid' that would rely on these few genes to 'get the big picture'. Additional genes can then potentially be used to resolve phylogenetic relationships on a per-family, per-rank basis.

> **…HUB participants share the concerns of others about the massive global loss in biodiversity and its impact on environments and peoples…**

Given the ability to determine accurately orthology and resolve the problem of gene sampling, several methodological challenges such as handling lateral gene transfer, the problem of gene tree—species tree discordance—and the reconstruction accuracy *per se* of methods for MSA and phylogenetic inference need to be critically assessed. Moreover, resource access (storage and computing capacity) is expected to become a bottleneck, because of the data deluge—the field is transitioning into a 'classic' computational science, similar to fluid dynamics or astrophysics. Another issue discussed in the context of 'inferring the tree of life', but which remains relevant in any computational discipline, is the potential danger generated by code incorrectness and lack of verification in large and complex data analysis pipelines.

Summary

Unseminars are a welcome alternative to traditional seminars, successfully bringing together researchers from different institutes and testing new ways to discuss science. The 'World Café' approach used in the HUB meeting is a great way to debate multiple topics with the engagement of every individual in a way that might not be practical in a big group discussion, as it encourages contributions to topics that one might not feel confident in contributing to in a big group discussion. The topics chosen at the HUB meeting allowed every participant to be informed on and to contribute to problems outside their individual field of research, and thus gain an understanding of the challenges in different fields. The concept of 'table hosts' ensured that as people moved from one table to the next, each group was brought up to speed on each discussion and valuable contributions were noted down.

We found that a weakness of this format was that topics were not always covered in great detail, due to the short time of each round of discussion. In one case this fixed time for discussion even resulted in discussion along one avenue—namely cost–benefit ratio analysis to determine whether scientific data should be stored—not being fully explored during the course of the event. Furthermore, it was difficult to provide traceability of statements, meaning that it can be difficult to follow up interesting or new ideas with the participant who contributed them. This could be easily addressed by encouraging participants to add their name to any ideas that were brought to the conversation. We would recommend trying this method, with the suggested modification, as it is a time-efficient way to discuss many topics even in a mixed group of researchers. Further discussions could be arranged to find solutions to the more challenging problems.

ACKNOWLEDGEMENTS

The authors acknowledge Jean-Karim Heriche and Rosario M. Piro for comments that improved the manuscript. Furthermore HUB acknowledges Rebecca C. Wade for encouragement in organizing the HUB meetings.

REFERENCES

1. Cochrane G, Cook CE, Birney E (2012) *GigaScience* **1:** 2
2. Fujita MK *et al* (2012) *Trends Ecol Evol* **9:** 480–488
3. Yang Z, Rannala B (2010) *Proc Natl Acad Sci USA* **20:** 9264–9269

Jonathan C. Fuller and Alexandros Stamatakis are at the Heidelberg Institute for Theoretical Studies, Heidelberg, Germany.
Pierre Khoueiry, Holger Dinkel, Kristoffer Forslund, Joseph Barry and Aidan Budd are at the European Molecular Biology Laboratory, Heidelberg, Germany.
Theodoros G. Soldatos is at Molecular Health GmbH, Heidelberg, Germany.
Abdul Mateen Rajput is at Life Science Informatics, Bonn, Germany.
Katja Linssen is affiliated with Heidelberg Unseminars in Bioinformatics (http://www.hub-hub.de).
The HUB Participants are listed at http://hub-hub.de/index.php/3rdUnseminarParticipants.
E-mail: jonathan.fuller@h-its.org

EMBO *reports* (2013) **14,** 302–304; published online 15 March 2013; doi:10.1038/embor.2013.34

