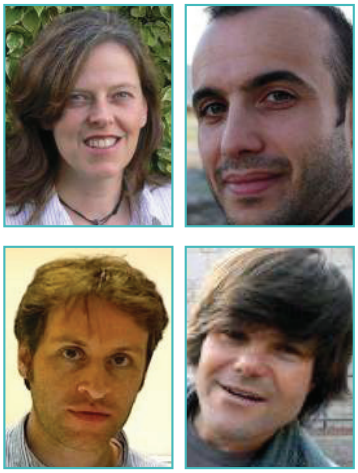
ICT needs and challenges for Big Data in the Life Sciences. A workshop report - SeqAhead/ISBE Workshop in Pula, Sardinia, Italy, 6 June 2013



Babette Regierer[1,2], Luca Pireddu[3], Martijn Moné[4], Andreas Gisel[5]

[1]Wageningen University, Wageningen, The Netherlands
[2]LifeGlimmer GmbH, Berlin, Germany
[3] Centro di ricerca, sviluppo e studi superiori in Sardegna (CRS4), Pula, Italy
[4]VU University Amsterdam, Amsterdam, Netherlands
[5]CNR - Institute for Biomedical Technologies, Bari, Italy

Received 15 July 2013; **Published** 19 August 2013

Sequencing has seen major breakthroughs in recent years and has paved the way for developing novel lifescience applications. Consequently, the life sciences are facing a rapidly increasing demand for data-handling capacity; in particular, when going from systems biology approaches to applications in, for example, systems medicine, the amounts of data to store, transfer and process overwhelm present-day capacities. New solutions must be developed by the Information and Communication Technologies (ICT) not only to adequately address the current challenges in the life sciences, but also to prepare for a future with an exponential growth of data during the next 10-20 years.

Now is the right time to address this important issue and bring experts from both fields – life sciences and ICT – together to formulate the necessary goals and prepare for this future. To stimulate this dialogue, two initiatives jointly organised a dedicated workshop with the aim of delineating life sciences' ICT-related requirements for the next ten years. Specifically, Babette Regierer, Luca Pireddu, Martijn Moné and Andreas Gisel joined efforts from the COST Action 'SeqAhead' (www.seqahead.eu) and the European ESFRI initiative 'ISBE – Infrastructure for Systems Biology in Europe[1] to analyse the ICT challenges and discuss the needs along the analytical pipeline, from data generation to data integration and modelling. In total, 20 experts from 12 European member states joined the workshop in Pula (Sardinia, Italy) on 6 June 2013 (Figure 1) to share their experience and perspectives on the Big Data challenge and discuss relevant topics that need new solutions. Emphasis was on: i) data processing – *i.e.*, tools and technologies essential to effectively process the avalanche of data; ii) data integration – *i.e.*, ICT and bioinformatics requirements for intelligently modelling and mining the vast bodies of data generated by NGS and other analytical technologies.

The morning session provided keynote presentations on relevant subjects and on-going activities, in order to stimulate discussions in the afternoon breakouts. Dr. Luca Pireddu (CRS4, IT), the local host and co-organiser of the workshop, and Dr. Andreas Gisel (CNR-ITB, IT) as representative of the SeqAhead COST Action introduced the workshop's goal and structure. The first talk by Dr. Martijn Moné (VU University Amsterdam, NL) on "*An infrastructure for European (systems) biology – ISBE*", gave an introduction to the new ESFRI infrastructure for systems biology in Europe, ISBE. Systems biology attempts to understand the functioning of organisms and of life in general via a process that includes data acquisition, analysis, integration and modelling. Sequencing has become one of the fundamental data-acquisition technologies for systems biology studies, exposing ICT challenges that need to be addressed to enable modelling of complex biological systems.

Dr. Babette Regierer (LifeGlimmer GmbH/ SeqAhead, DE) and Dr. Daniel Jameson (University of Manchester, UK) introduced the

1 www.isbe.eu

Figure 1. The SeqAhead/ISBE workshop participants at CRS4 in Pula (Sardinia) on 6 June 2013. (Source: Valentine Svensson, SciLifeLab)

European initiative on '*IT Future of Medicine (ITFoM) – ICT challenges for a virtual patient*'. This initiative aims to create a virtual patient to help health-care professionals better define personalised therapy and prevention strategies. An important goal is to identify major ICT challenges faced by the project, and to develop a roadmap for their resolution. The ITFoM consortium[2] thus developed a concept and roadmap for ICT challenges for the generation and implementation of a virtual patient. Major issues are expected in the optimisation of hardware and software, and in the efficiency of machine-learning and statistical methods; these are not just important for the virtual patient, but are generally applicable across all of the life sciences.

Dr. Simon Heath (Centre Nacional d'Anàlisi Genòmica (CNAG), ES) presented the operations at CNAG '*Dealing with NGS data - the CNAG experience*', which are an excellent example of efficient NGS data processing. The close cooperation with a computer centre shows the advantages of coupling data-generation and computing facilities. The presentation included information about current practices and expected challenges in scaling up the processes. Dr. Luca Pireddu (CRS4, IT) suggested in his presentation on '*Data-intensive computing in NGS*' adoption of technologies developed in data-driven computing activities (Big Data) to help address the ICT challenges faced by modern bioinformatics. Distributed computing frameworks like Hadoop[3] could be a suitable solution to scale processing pipelines for sequencing data. Another principle expected to speed up the processing of high-volume data is the use of distributed, column-oriented databases; adopting these technologies, however, requires the creation of new software tools.

As the last speaker of the morning Dr. Heimo Müller (Medical University Graz, AT) presented the 'BiBBox – Biobanking in a box[4], a system that can store and grant access to patient data. The number of samples (*i.e.*, patients) involved in sequencing studies grows steadily; thus, it is im-

2 http://www.itform.eu

3 http://hadoop.apache.org

4 http://bibbox.org

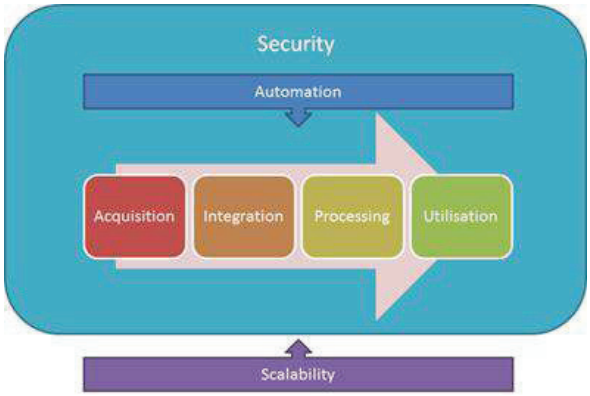


Figure 2. Major steps in the analysis of BIG DATA in the life sciences, including cross-cutting topics, such as scalability, automation and security (Source: D. Jameson).

portant to plan for scalable methods for tracking these samples. BiBBox looks to solve this problem for what could nowadays be considered large studies. On the other hand, the BBMRI initiative[5] looks further into the future, planning for biobanks at the national and European scale. The afternoon session included breakout groups to discuss and catalogue ICT needs and challenges from a range of different expert perspectives, encompassing sequencing technology, (bio) informatics, computer science, systems biology and user communities.

The presentations showed that the topic of the workshop spans a large spectrum of aspects that play key roles in the analysis, handling and management of lifescience data – these include not just steps along the analytical pipeline, like data integration, but also cross-cutting topics such as security (Figure 2).

Summary of the results

The overall process is structured in several different layers: i) the pillars of the principal pipeline, from data generation to modelling, comprise all aspects, from machine to model – and back to the application (*e.g.*, patient); all these areas have certain needs and challenges; ii) cross-cutting issues, like scalability, automation, timeliness, storage, privacy and security, are relevant for all areas; also, organisational aspects, communication, standardisation and metadata are of high relevance, and require development of strategies and agreements on a more general level; iii) time – *i.e.*, while we have to define needs and challenges for the next five years, we also need to prepare now for the next 10-20 years.

In both breakout sessions, the participants identified common themes, including the following:

- data accessibility must be easier and more efficient;
- as many processes as possible must be automated;
- standardisation is a prerequisite to generate automated processes;
- scaling up of data processing and storage needs new approaches (*e.g.*, Hadoop);

CRS4, located in Pula close to Cagliari, was the host institution of the workshop. (Source: CRS4)

5 http://www.bbmri.eu

- federated approaches might be more successful than a centralised solution;
- metadata and protocols must be standardised and implemented along the whole pipeline (this information must travel with the data);
- standardisation of data, processes, software, and so on, are required;
- intuitive user interfaces are key to allowing the data and tools to be used by non-experts;
- reproducibility and traceability of pipelines.

The participants will initiate a COST Action to address these ICT challenges, to organise the communities involved, to create a strategy for a better communication between the different communities, and to generate a collection of necessary ICT solutions to enable the future of data-intensive life sciences.

Supporting institutions: