EMBnet.journal
Bioinformatics in Action

Pfeil et al. (2018) EMBnet.journal 23, e905
http://dx.doi.org/10.14806/ej.23.0.905

Reports

Report on the "Big Data Training School for Life Sciences", 18-22 September 2017, Uppsala, Sweden

Juliane Pfeil[1]✉, **Sabrina K. Schulze**[2], **Eftim Zdravevski**[3], **Yen Hoang**[4]

[1] Division Molecular Biotechnology and Functional Genomics, Technical University of Applied Sciences, Wildau, Germany

[2] Cell2Fab (Synthetic Biology, Faculty of Biochemistry and Biology), University of Potsdam, Potsdam, Germany

[3] Department of Information Systems, Faculty of Computer Science and Engineering, Sts. Cyril and Methodius University in Skopje, Skopje, Macedonia

[4] Department of Signal Transduction, German Rheumatism Research Center Berlin, A Leibniz Institute, Berlin, Germany

Competing interests: JP none; SKS none; EZ none; YH none

Abstract

In September 2017 a "Big Data Training School for Life Sciences" took place in Uppsala, Sweden, jointly organised by EMBnet and the COST Action CHARME (Harmonising standardisation strategies to increase efficiency and competitiveness of European life-science research - CA15100). The week programme was divided into hands-on sessions and lectures. In both cases, insights into dealing with big amounts of data were given. This paper describes our personal experience as students' by providing also some suggestions that we hope can help the organisers as well as other trainers to further increase the efficiency of such intensive courses for students with diverse backgrounds.

Course of the training school

The "Big Data Training School for Life Sciences"[1] was a joint initiative of the EU COST Action CHARME[2] and EMBnet[3]. The main objective of these organisations is to network scientists of different countries (within Europe and neighbouring areas) to serve, support and sustain the biological and biomedical research. The aim of the training school was to increase the efficiency of life-science research and interdisciplinary collaboration by training students and researchers who have to cope with the need to manage big data for their research activity. The school took place from 18th to 22nd September 2017 at the Campus Ultuna of the Swedish University of Agriculture[4] (SLU) in Uppsala, Sweden.

The programme included lectures accompanied by hands-on sessions in the first three days (September 18-20), and lectures (Lecture days, 21-22 September) open to a wider audience in the last two days.

The hands-on sessions, as well as the lectures, dealt with different topics in the field of Bioinformatics for big data management and analysis. In total 25 students (the max. number allowed) took part in the first three days.

Figure 1 shows a nice group picture of some trainers and of trainees inside the building of the Ultuna Campus.

The participants for the hands-on sessions applied for a stipend (CHARME grants) several months in advance. The scientific committee[5] selected them based on diverse criteria, including research background and motivation, while also attempting to ensure an equal distribution of CHARME grants on the basis of national, gender and ethical diversity.

Trainees selected had different scientific backgrounds: bioinformaticians with limited knowledge in Big Data technologies, computer scientists that moved to bioinformatics, and scientists with biological/medical backgrounds with some knowledge in bioinformatics.

The first day was opened by Erik Bongcam-Rudloff, Professor of Bioinformatics at SLU and vice-Chair of CHARME. He gave an overview and showed the impact of Big Data in different areas of life and its potential for the future. The day was closed by Dr Jim Dowling, a distributed systems researcher at RISE SICS[6] and Associate Professor at KTH ICT School[7]. He presented a lecture and a hands-on session about Hops Hadoop, an open-source platform for analysing Big Data in an uncomplicated way. The second day was opened with an introduction to machine learning by Gioele La

[1] http://astrocyte.com/COST-CHARME/COST-CHARME/Home.html

[2] http://www.cost-charme.eu

[3] https://www.embnet.org

[4] https://www.slu.se/en/

[5] http://astrocyte.com/COST-CHARME/COST-CHARME/About.html

[6] https://www.sics.se/

[7] https://www.kth.se/en/ict

Article history
Received: 19 December 2017
Published: 05 February 2018

Pfeil et al. (2018) EMBnet.journal 23, e905
http://dx.doi.org/10.14806/ej.23.0.905

Figure 1. Group picture of the participants and some of the lecturers of the first three days. @Photo by Erik Bongcam-Rudloff.

Manno (Karolinska Institutet) followed by a practical example in Apache Spark presented by Marco Capuccini (Uppsala University). In another hands-on session in the afternoon, Apurva Nandan (CSC - IT Center for Science, Finland) introduced Apache Spark SQL, a toolkit for easy parallelisation of computations. The day was finished by Dr Vaughan Wittorf from PetaGene[8], explaining the idea of how storing and transferring next-generation sequencing (NGS) data can be done more efficiently. On Wednesday, Dr Witold Rudnicki (Associate Professor at University of Białystok) started with a session about gene expression data and a possible way to extract informative changes with R and the classification method Random Forests. Kim Kultima, Payam Emami, Stephanie Herman, and Ola Spjuth from Uppsala University were the last speakers on this day. They introduced their method of handling metabolomics data produced by mass spectrometry with OpenMS and Pachyderm. Their work is dedicated to the two major projects Caramba.clinic[9] and Phenomenal[10].

The last two lecture days were opened by Erik Bongcam-Rudloff and Domenica D'Elia (Chair of EMBnet and of the Dissemination Working Group in CHARME). After the introduction of the objectives of EMBnet and CHARME, there was a presentation from Roxana Merino Martinez (Karolinska Institutet) that presented the B3Africa project[11]. During the day, there were presentations by the two speakers Gaurav Kaul and Erik Gullbring from the technology leaders Intel and Microsoft, respectively, that showed powerful frameworks for cloud computing and machine learning. The importance of the programming language R was highlighted by Seija Sirkiä (CSC). During this day, Prof. Witold Rudnicki presented more information about the work he shortly introduced during the hands-on sessions. Ola Spjuth gave a deeper insight into the project Phenomenal. Bjorn Lindell from SLU focused on the impact of cloud solutions. The day was closed with an open and fruitful discussion about the use of Big Data technologies in bioinformatics. The last day was introduced by Anders Herlin (SLU), who showed possibilities to improve the conditions of livestock farming using Big Data. Kim Kultima also gave more insights into the projects Caramba.clinic and Phenomenal. After that, Nataša Sladoje (Centre for Image Analysis, Uppsala University) showed her work and some possible training opportunities of the COST

[8]https://www.petagene.com/

[9]http://www.caramba.clinic

[10]http://phenomenal-h2020.eu

[11]http://www.b3africa.org/

Pfeil et al. (2018) EMBnet.journal 23, e905
http://dx.doi.org/10.14806/ej.23.0.905

Action NEUBIAS[12] (Network of European Bioimage Analysis). In this context, one participant of the training school was asked for a spontaneous presentation of the mobile microscopic system from the company Oculyze[13]. Possible applications, especially in the B3Africa project, were discussed. The last day was closed by Professor Dimitrios P. Vlachakis from the Biotech Department of the Agricultural University of Athens, Greece, who presented a long-term scientific project about NGS data of inhabitants in Cyprus.

Pros and Cons from students' point of view

As mentioned above, this was the first edition of this training school. The transmitted contents of the lessons fit very well into the area of Big Data and highlighted the topic from different perspectives. Despite the heterogeneous education background of the participants, everyone had certainly obtained helpful information about the way to improve the quality of their work, and the exchange of information and comments among participants was very rewarding. All the lecturers were keen in passing along their knowledge and their experience to the audience. When questions came up, they were immediately answered or their assistants came to solve the problems directly with the person who had asked help. Unfortunately, the available time was too short to satisfy the needs of all attendants, especially during the hands-on sessions.

After the school conclusion on Friday afternoon, the organisers invited us to discuss and report our personal opinion on the school to improve possible future editions. The discussion was open to all attendants and after a constructive exchange of comments and ideas we were invited by Domenica D'Elia and Erik Bongcam-Rudloff to submit this article for its publication in the EMBnet.journal. The main objective was to allow a wider dissemination of the challenges encountered during hands-on training of aspects related to Big Data issues. We hope that it will allow trainers to better programme the learning objectives of this type of schools to the actual possibility to transmit huge amounts of knowledge in the space and time that a one-week training school can have.

Our first impression was that it would have been more advantageous if the interesting lectures, which had taken place on Thursday and Friday, would have occurred before the hands-on sessions. Much of the basics could have been covered here instead of during the sessions. By doing so, the time of the hands-on sessions would not have been as limited to introduce the methods by the lecturer. More than once it was noticed that the lecturer had more to tell, but could not do because the time slots ended. Still, this introduction was necessary due to the diverse background of the attendees.

As highlighted earlier, the participants had different scientific backgrounds. This diversity made it a challenging task to adequately set the learning goals. The goals of the school ranged from introductions to machine learning and Big Data technologies, to ways of applying such technologies for addressing computational challenges in bioinformatics. The selection of more uniform groups of participants could have allowed trainers to focus much on trainees' specific needs. Nevertheless, this approach has a major drawback that it hinders interdisciplinary collaboration, one of the most important goals of this training school. Indeed, the advantage of the education background diversity was that the students were forced to interact more with each other by asking questions and offering support in a complementary way. By doing so, personal interactions were spontaneously fostered.

A way to somewhat balance out the different trainees' backgrounds could be achieved by "assigning some homework" before the training school. In this case, the trainee could have gotten a list of required knowledge and maybe texts/links to study by themselves beforehand. This would improve the effectiveness of training by allowing trainees to concentrate much more on the hands-on work. Indeed, during the hands-on sessions, participants faced for the first time with a large amount of material to deal with. It was necessary to understand the lesson and to do some practical work in parallel. Although this helped enhance the effects of the learning, in some cases (*e.g.*, students with a poorer bioinformatics background) it lowered its efficiency.

Another aspect that should be improved is related to the preparation of the computer environment for the training school that includes software and other materials. The downloads of this material during the training sessions took away precious time, which could have been used to better address the issues in Big Data. Moreover, because of the different computers' efficiency in downloading, not all machines were ready to work in due time. Therefore, the hands-on sessions were resumed after the first few successful downloads, as the lecturers had to respect the timetable for the explanation of methods and tools foreseen to be completed in the session. Many participants had to follow the meaningful input and monitor the installation in parallel. This was complicated and forced people to focus on one thing or the other.

Another suggestion we would like to provide is about the number of topics that the school should include. It was good for us to know the huge variety of bioinformatic issues related to Big Data and the variety in finding new and more effective solutions, but it would have been even better if the focus was on fewer topics, but with more intensity.

If bioinformaticians want to use Big Data technologies, it is not always clear where to start, as the information provided in a school can be overwhelming. To improve this aspect, the application for the school can include a survey about what common algorithms prospective students use or have used in the past. Then,

[12]http://eubias.org/NEUBIAS/

[13]www.oculyze.de

Pfeil et al. (2018) EMBnet.journal 23, e905
http://dx.doi.org/10.14806/ej.23.0.905

the results of such a carefully crafted survey can help in identifying some algorithms that could be presented during the course for a Big-Data approach and a traditional method. Comparison in terms of performance for some use-cases could further complement this. This can engage bioinformaticians more and incline them to push for Big Data approaches in their departments. This can ultimately lead to initiatives for reusable open source implementation of popular algorithms in Big Data technologies, which could draw people from computer science into bioinformatics.

Overall, we had an interesting week where we learnt a lot about dealing with Big Data. The organisers, as well as the lecturers, did a good job communicating present Big Data challenges and solutions. Our above remarks should be considered as mere constructive feedback and suggestions for improvement of other training schools, not only on Big Data but in general.

Conclusions

The “Big Data Training School for Life Sciences” was an excellent idea for learning trends in this field for young scientists and perfect for networking purposes. Realising the existence of immense computational challenges in bioinformatics is of great importance for interdisciplinary collaboration, and the school was particularly successful in this regard. The networking part was done during the breaks and in the evenings. One of these opportunities was during the “Ethno-Party” on Thursday. Every participant had been asked to bring some food and drinks from their home country to share.

Some people stayed in touch even after the end of the school. A follow-up symposium with the title “Bioinformatics meets Synthetic Biology” was organised. It took place on 20th October 2017 in Potsdam-Golm, Germany. Five participants of the original school and five other scientists presented their projects and work.

Acknowledgement

We would like to thank SLU for organising the school. A huge thanks to Erik Bongcam-Rudloff and his organisation team for their great work. Also, thanks to EMBnet and to COST Action CHARME for funding this school and the stipends so that we had the opportunity to attend. Another thanks to all the lecturers and speakers during this week who gave us some insight in their work and personal experiences. The "Big Data Training School for Life Sciences" 18-22 September 2017 Uppsala, Sweden, was financed by the COST action CA15110 supported by the EU framework program H2020.