Genomics Data 2 (2014) 37–39

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Report from the 2nd Summer School in Computational Biology organized by the Queen's University of Belfast

Frank Emmert-Streib [a,*], Shu-Dong Zhang [b], Peter Hamilton [b]

[a] Computational Biology and Machine Learning Laboratory, Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, United Kingdom
[b] Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, United Kingdom

ARTICLE INFO

Article history:
Received 9 November 2013
Received in revised form 17 December 2013
Accepted 20 December 2013
Available online 13 February 2014

Keywords:
Computational biology
Genomics data
High-throughput data

ABSTRACT

In this paper, we present a meeting report for the 2nd Summer School in Computational Biology organized by the Queen's University of Belfast. We describe the organization of the summer school, its underlying concept and student feedback we received after the completion of the summer school.

Introduction

This year the 2nd Summer School in Computational Biology organized by the Queen's University of Belfast took place 16–18 September 2013 in Belfast (UK). The event gathered a total of 25 students from 5 different countries, making it so far the largest summer school since its establishment in 2012. The purpose of the summer school was to provide the participating students with a systematic introduction to quantitative analysis methods for high-throughput data that are needed to analyze genomics data from biological or biomedical experiments [1,2]. Due to the fast paced progress of the field *genomics*, many university curricula have difficulties in catching-up with these developments, leaving room for a summer school like ours to provide the needed understanding for basic analysis principles and methods to the students to equip a new generation of scientists with the necessary skill set. In order to emphasize our goal for the summer school clearly we chose the subtitle: ‘Statistical and computational methods to analyze high-dimensional data in biology’.

The summer school consisted of 18 lectures that were provided by 10 instructors, covering topics from introductory lectures for the statistical programming language R, microarray and next-generation sequencing (NGS) data to advanced analysis methods, including clustering, classification and survival analysis. In this paper, we provide a report about this summer school, including a discussion of student feedback we received after the completion of the summer school.

Meeting report

The underlying goal of our summer school was to teach students the necessary skill set for computational biology so these can be applied to, e.g., biological, biomedical or clinical data; see Fig. 1. Due to the interdisciplinary character of computational biology, which consists of a mixture of skills and knowledge from statistics, machine learning, computer science, biology and medicine, this is usually a long-lasting and challenging endeavor, because no student can have a background in all subjects. For this reason, we provided an introduction to all three basic subjects (see Fig. 1). In Fig. 2 we show the schedule of the summer school, comprising 18 lectures distributed over 3 days. We emphasize the main purpose of each lecture by the same color code as used in Fig. 1. The lecture notes for all lectures can be downloaded from http://www.bio-complexity.com/QUBsscb13/QUBsscb.html.

The covered topics range from a basic introduction to the statistical programming language R [3] (lectures 2, 4 & 5) to advanced statistical analysis methods, including classification, clustering, pathway analysis, survival analysis and connectivity mapping (lectures 9, 10, 11, 15, 16 & 17) [4–9]. For each lecture, we aimed for a clear and hands-on explanation of a topic. For this reason we sacrificed a more comprehensive coverage, which would have forced us to extend the duration of the summer school considerably.

The actual need for such a wide range of topics and difficulty levels becomes clear from the educational background of the students. In

[*] Corresponding author.
E-mail address: v@bio-complexity.com (F. Emmert-Streib).

http://dx.doi.org/10.1016/j.gdata.2013.12.001

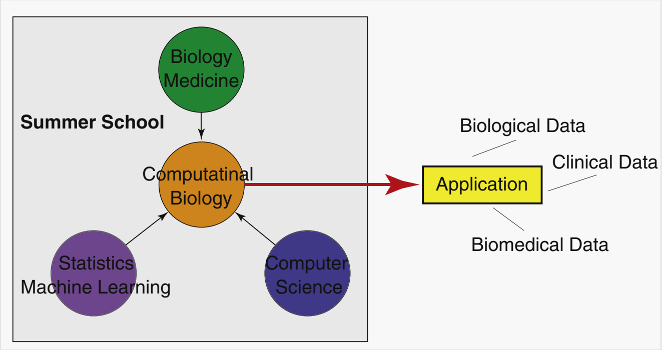

Fig. 1. A schematic overview of the constituting subjects of computational biology and its application domains. This visualization represents also the conceptual organization of the summer school.

Fig. 3A, we show an overview of the subjects studied by the participating students of the summer school. As one can see, the majority of the students have an educational background in biology and biochemistry, however, there are also 29.5% of the students studying quantitive subjects (computer science, mathematics, statistics or physics). Furthermore, about half of the students are at the beginning of their university education (BSc student or MSc student) whereas the other half is on an advanced stage (PhD student or Postdocs), see Fig. 3B. This education diversity makes it challenging to provide an *optimal* level of the summer school lectures that fits all students equally. For this reason, we decided to provide basic lectures to give the students with a biology background the chance to catch-up basic skills that every student of a quantitative subject already has, and also to provide basic biology knowledge (lecture 3), e.g., computer science students to counterbalance their deficits.

In order to obtain feedback from the students regarding their experience from the summer school, we asked the students to fill in a questionnaire via 'survey monkey' after the completion of the summer school. In Fig. 3C we show the results for the question 'Please describe the difficulty level of the summer school for you'. It is reassuring to see that the majority of students (70.6%) felt comfortable with the difficulty level of the provided lectures and none (0%) considered the summer school as 'very difficult'.

Day 1.
[1.] Introduction to computational biology and high-throughput data.
[2.] Introduction to R: basics.
[3.] Basic Cancer Biology.
[4.] Introduction to R: applications I.
[5.] Introduction to R: applications II.
[6.] Biological databases and enrichment analysis.

Day 2.
[7.] Introduction to microarray data.
[8.] Introduction to microarray data analysis I: individual genes.
[9.] Introduction to microarray data analysis II: hypotheses testing for gene sets and pathways.
[10.] Supervised learning: classification.
[11.] Unsupervised learning: clustering
[T.] Invited talk

Day 3.
[12.] Introduction to next-generation sequencing data.
[13.] Introduction to microarray data analysis III: hypotheses testing for gene sets and pathways.
[14.] Analysis of next-generation sequencing data.
[15.] Survival analysis: Kaplan Meier curves.
[16.] Survival analysis: Cox proportional hazard model.
[17.] Connectivity map.
[18.] Summary and outlook to genome medicine.

Fig. 2. Schedule of the summer school. The colors correspond to the subjects in Fig. 1, indicating the emphasis of the corresponding lectures.

The next three questions we asked the students addressed their gained understanding from the summer school on different levels:

- Have you gained a basic understanding of the programming language R? See Fig. 3D.
- Have you gained a basic understanding of the statistical analysis of data? See Fig. 3E.
- Have you gained a basic understanding of high-throughput technologies? See Fig. 3F.

Overall, the majority of the students are of the opinion that they gained a basic understanding in these 3 categories. Given the educational diversity of the students, see Fig. 3A and B, it is probably hard to improve these numbers substantially without increasing the duration of the summer school that would allow one to cover more basic and advanced grounds in considerably more depth.

We would like to emphasize that the above discussion is not intended as a strict data analysis, but as a *quantitative* discussion of the feedback we received from the students.

Reproducible research

Aside from the actual topic of the summer school, teaching students the basics of computational biology, we intended to generate an awareness of the students for *reproducible research* [10–12]. In this respect, the key role of the statistical programming language R was discussed as a natural and efficient mediator between data, analysis methods and the conservation of the whole analysis process in a way that allows an error-free exchange/communication of a conducted analysis. Given the fact that it is anticipated to see in the next years an even increasing amount of data from more than one high-throughput technology, the reproducibility of a genomic data analysis, and means to realize it, is certainly becoming even more important in the near future.

Invited talk

Another feature of the summer school was to give the students the opportunity to see the application of the taught methods to real problems. For this reason, we invited Benjamin Haibe-Kains, from the Institut de Recherches Cliniques de Montreal (Canada), to provide a talk at the summer school. This talk was given at the end of the second day (see Fig. 2). Benjamin Haibe-Kains gave an interesting and exciting talk with the title *Significance Analysis of Prognostic Signatures*, which was based on recent results of his research [13,14]. This allowed the students to experience the translation of analysis methods into clinical practice and to engage into a lively discussion, which helped them to realize the importance of computational biology.

In this role to provide a lecture for the student, Benjamin Haibe-Kains is the successor of Simon Tavare, from the Statistics and Computational Biology Laboratory, Cancer Research UK Cambridge Research Institute, who provided last year's talk for our first summer school.

Conclusions

Overall, the experience we made during the last two years from organizing the Summer School in Computational Biology is very positive. We see not only a clear benefit for the students but also for ourselves, allowing us to provide more efficient lectures that can reach students with a heterogeneous educational background and professional level. For this reason the preparations for next year's summer school have already started and we anticipate the 3rd Summer School in Computational Biology to take place in September 2014.

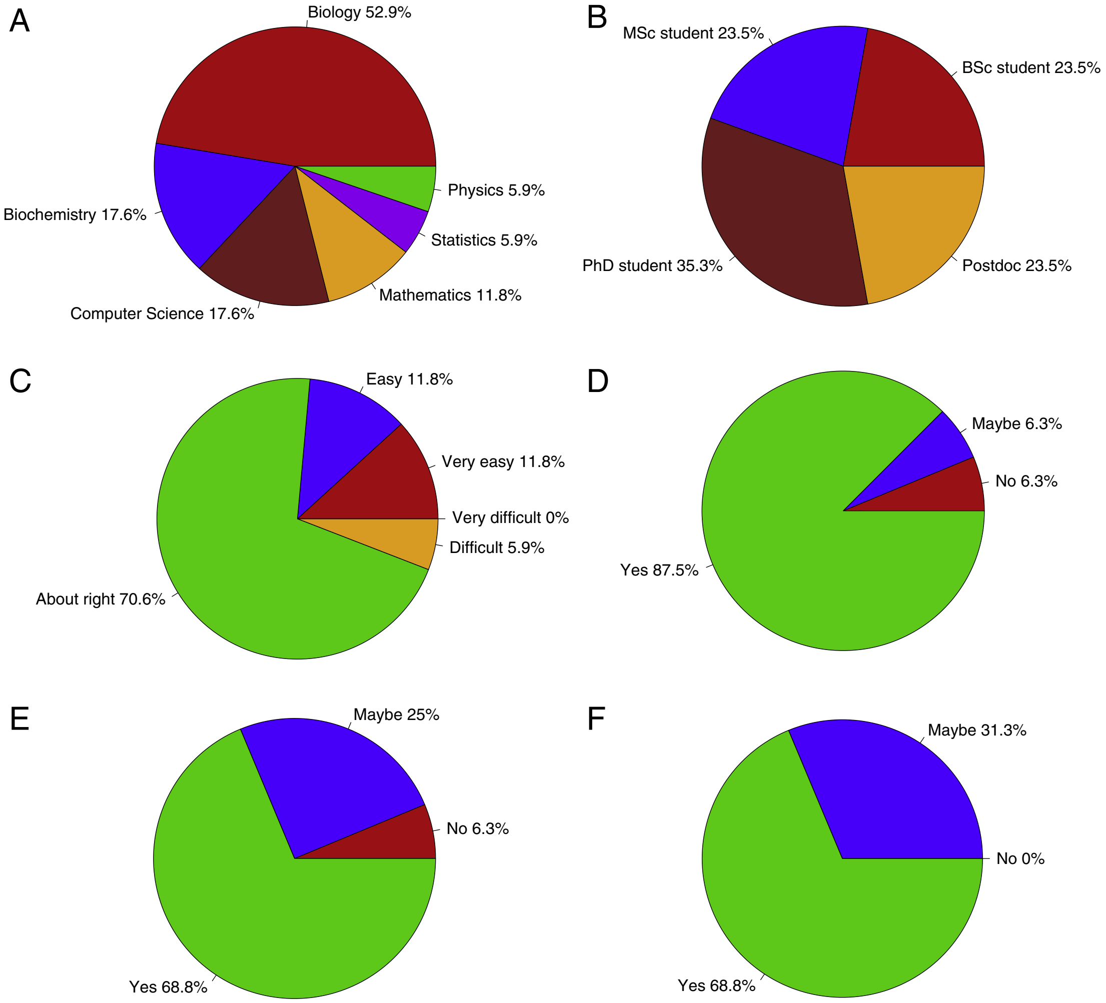

Fig. 3. Overview statistics of the participating students in the summer school. A and B provide information about the educational background of the students. C–F show feedback from the students to the questions: Please describe the difficulty level of the summer school for you (C). Have you gained a basic understanding of the programming language R? (D). Have you gained a basic understanding of the statistical analysis of data? (3E). Have you gained a basic understanding of high-throughput technologies? (F).

Acknowledgments

We would like to thank all the colleagues who contributed to the teaching of the summer school (in alphabetic order): Jaine Blayney, Ricardo de Matos Simoes, Kevin Keenan, Darragh McArt, James McCann, Ken Mills, and Salissou Moutari. For financial support we would like to thank the Center for Cancer Research and Cell Biology (CCRCB) and sAnalytiCo.

References

[1] G. Dutton, Computational genomics: the medicine of the future? Ann. Intern. Med. 131 (10) (1999) 801–804.
[2] E. Hernandez-Lemus, Further steps towards functional systems biology of cancer. Frontier in Physiology 4 (2013) 256, http://dx.doi.org/10.3389/fphys.2013.00256 URL http://www.frontiersin.org/systems_biology/10.3389/fphys.2013.00256/fulltext.
[3] R Development Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0.
[4] B. Clarke, E. Fokoue, H.H. Zhang, Principles and Theory for Data Mining and Machine Learning. Springer, Dordrecht, New York, 2009.
[5] F. Emmert-Streib, S. Tripathi, R. de Matos Simoes, Harnessing the complexity of gene expression data from cancer: from single gene to structural pathway methods. Biol. Direct 7 (2012) 44.
[6] A.J. Izenman, Modern Multivariate Statistical Techniques. Springer, New York, NY USA, 2008.
[7] D. Kleinbaum, M. Klein, Survival Analysis: A Self-Learning Text, Statistics for Biology and Health. Springer, 2005.
[8] D.G. McArt, P.D. Dunne, J.K. Blayney, M. Salto-Tellez, S. Van Schaeybroeck, P.W. Hamilton, S.-D. Zhang, Connectivity mapping for candidate therapeutics identification using next generation sequencing RNA-Seq data. PLoS ONE 8 (6) (2013) e66902.
[9] S. Tripathi, G.V. Glazko, F. Emmert-Streib, Ensuring the statistical soundness of competitive gene set approaches: gene filtering and genome-scale coverage are essential. Nucleic Acids Res. 6 (12) (2013) e53354.
[10] P.J. Diggle, S.L. Zeger, Embracing the concept of reproducible research. Biostatistics 11 (3) (2010) 375 (Oxford, England).
[11] R. Gentleman, Reproducible research: a bioinformatics case study. Stat. Appl. Genet. Mol. Biol. 4 (2005) (Article2).
[12] R.D. Peng, Reproducible research in computational science. Science 334 (6060) (2011) 1226–1227.
[13] B. Haibe-Kains, C. Desmedt, S. Loi, A.C. Culhane, G. Bontempi, J. Quackenbush, C. Sotiriou, A three-gene model to robustly identify breast cancer molecular subtypes. J. Natl. Cancer Inst. 104 (4) (2012) 311–325.
[14] S. Bentink, B. Haibe-Kains, T. Risch, J.-B. Fan, M.S. Hirsch, K. Holton, R. Rubio, C. April, J. Chen, E. Wickham-Garcia, J. Liu, A. Culhane, R. Drapkin, J. Quackenbush, U.A. Matulonis, Angiogenic mRNA and microRNA gene expression signature predicts a novel subtype of serous ovarian cancer. PLoS ONE 7 (2) (2012) e30269, http://dx.doi.org/10.1371/journal.pone.0030269.