Proceedings of the EuBIC Developer's Meeting 2018

Sander Willems[1], **David Bouyssié**[2], **Dieter Deforce**[1], **Viktoria Dorfer**[3], **Vladimir Gorshkov**[4], **Dominik Kopczynski**[5], **Kris Laukens**[6], **Marie Locard-Paulet**[2], **Veit Schwämmle**[4], **Julian Uszkoreit**[7], **Dirk Valkenborg**[8,9], **Marc Vaudel**[10,11], **Wout Bittremieux**[6,12]

[1]Laboratory of Pharmaceutical Biotechnology, Ghent University, Ghent, Belgium; [2]Institute of Pharmacology and Structural Biology, University of Toulouse, CNRS, UPS, Toulouse, France; [3]Bioinformatics Research Group, University of Applied Sciences Upper Austria, Hagenberg, Austria; [4]Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, Denmark; [5]Leibniz-Institut für Analytische Wissenschaften – ISAS – e.V., Dortmund, Germany; [6]Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium; [7]Medizinisches Proteom-Center, Ruhr University Bochum, Bochum, Germany; [8]Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Hasselt, Belgium; [9]Centre for Proteomics, University of Antwerp, Antwerp, Belgium; [10]KG Jebsen Center for Diabetes Research, Department of Clinical Science, University of Bergen, Bergen, Norway; [11]Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, Norway; [12]Department of Genome Sciences, University of Washington, Seattle WA, USA

Corresponding author: wout.bittremieux@uantwerpen.be

Abstract

The inaugural European Bioinformatics Community (EuBIC) developer's meeting was held from January 9th to January 12th 2018 in Ghent, Belgium. While the meeting kicked off with an interactive keynote session featuring four internationally renowned experts in the field of computational proteomics, its primary focus were the hands-on hackathon sessions which featured six community-proposed projects revolving around three major topics: (i) quality control; (ii) workflows, protocols, and guidelines; and (iii) quantification. Here, we present an overview of the scientific program of the EuBIC developer's meeting and provide a starting point for follow-up on the covered projects.

1 Introduction

The European Bioinformatics Community (EuBIC) is an initiative of the European Proteomics Association (EuPA) to promote the use of bioinformatics for computational mass spectrometry (MS) and MS-based proteomics. Our goal is to bring together the European MS bioinformatics community, including students and early-career researchers as well as long-standing experts from both academia and industry. Through the setup of community-driven dynamics, EuBIC mainly focuses on improving education in computational methods, job and funding opportunities, international collaborations, publication of specialized studies, and training of software tools. To this end, EuBIC maintains several web resources that include educational videos, grant overviews, a job fair, and tutorials (`https://www.proteomics-academy.org/`). Besides these online resources, EuBIC regularly organizes workshops and hubs at the major international conferences on computational MS and proteomics. Additionally, an annual conference on computational MS-based proteomics is organized by EuBIC itself, forming an important community outreach effort to bring together bioinformatics researchers from all over Europe.

The first EuBIC conference took place in January 2017 in Semmering, Austria [13]. As this turned out to be an overwhelming success, we envisioned to organize the EuBIC conference as an annual series. However, although this event brought together the European proteomics community, we observed that not all computational expertise was utilized to its full potential in the typical conference setup consisting of presentations and workshops. Therefore we decided to alternate the annual EuBIC conference between a Winter School targeting a broad end user-oriented audience and a developer's meeting for software developers.

The inaugural EuBIC developer's meeting was organized in Ghent, Belgium, from January 9th to January 12th 2018 (`http://uahost.uantwerpen.be/eubic18/`). A total of 43 participants (figure 1), including students, keynote speakers, and industry representatives from 14 different countries participated in the developer's meeting. To stimulate direct collaboration and the active development of bioinformatics applications, its main activity was a hackathon focusing on six important topics in computational proteomics which were crowd-sourced from the community. Additionally, prior to these hackathon sessions the meeting participants engaged in an interactive keynote session led by four internationally renowned scientists with experience in tool development for MS-based proteomics.

2 Keynote presentations

The EuBIC developer's meeting kicked off with four keynote presentations illustrating some important current drawbacks of MS-based data analysis and the crucial role of bioinformatics in solving these outstanding issues.

Prof. dr. Lennart Martens of Ghent University, Belgium, opened the meeting by describing his vision on the role of a bioinformatics scientist as a "researcher-developer". As life sciences research has accelerated enormously over the past two decades, nowadays it is heavily dominated by the huge amount of data that are generated and the

Figure 1: Participants of the EuBIC developer's meeting 2018.

advanced algorithmic techniques that are necessary to analyze these data. He outlined that the job of a researcher-developer is to use and develop sophisticated algorithms and powerful tools to increase our understanding of the sheer complexity of biological systems [5]. This was followed by an interactive discussion on career aspects and the growth path of bioinformatics researchers.

Next, dr. Frédérique Lisacek of the Swiss Institute of Bioinformatics (SIB), Switzerland, presented her work on bridging proteomics and glycomics. She described difficulties prohibiting the fully automated identification of glycoproteomics data and explained how her group has tackled some of these issues. By making use of open modification searching peptides with previously unconsidered post-translational modifications (PTMs) could be successfully identified [3]. Next, she explained how new computational tools can be used for the analysis of glycoproteomics data [4].

The third keynote speaker was dr. Laurent Gatto of the University of Cambridge, England, who gave a presentation on the ecosystem of open-source tools in the R programming language for the analysis of MS data [2]. Dr. Gatto showed a historical perspective on how increasingly powerful and popular R packages for the analysis of proteomics data have been developed. Based on a few use cases he demonstrated how several popular packages are related to each other and reinforce each other, thereby illustrating the effectiveness of open-source.

The final keynote speaker was prof. dr. Lukas Käll of the KTH Royal Institute of Technology, Sweden. Prof. Käll explained that although the characterized analytes in an MS proteomics experiment are peptides, researchers are typically interested in their parent proteins instead. As a result, protein inference has to be performed to reassemble protein sequences from the measured peptide sequence data. Based on simulated data and a sample of known content, prof. Käll demonstrated the effect of different design choices of protein inference algorithms [10]. Furthermore, he discussed the protein summarization problem, which aims to recreate proteins' relative concentration from peptides' abundances, and his Diffacto algorithm [14].

In addition to these invited scientific keynotes two sponsored presentations were given by company representatives. First, Adam Tenderholt from Veritomyx presented the PeakInvestigator^{TM} software, which helps with deconvoluting and centroiding mass spectra. Second, Lyle Burton from SCIEX explained which application programming interfaces (APIs) they provide and how to use them. He also showed some examples of how these APIs are already used in open source and proprietary projects.

3 Hackathon

During the subsequent days of the EuBIC developer's meeting the participants split up into small groups to actively develop bioinformatics applications. Project proposals for the hackathon sessions were crowd-sourced in a transparent and open process. Prior to the developer's meeting community members could submit project proposals for inclusion in the hackathon, which were subsequently evaluated on scientific merit and community interest. This resulted in a hackathon program consisting of six different projects in three main tracks: (i) quality control; (ii) workflows, protocols, and guidelines; and (iii) quantification.

3.1 Quality control

Dashboard for longitudinal QC monitoring During this hackathon session the participants developed a web tool for the visualization and analysis of quality control (QC) metrics. Based on data in the qcML format [11] an interactive R/Shiny dashboard was developed using a microservice architecture. The dashboard includes functionality to visualize specific QC metrics longitudinally and perform a robust principal component analysis to detect low-performing experiments [12].

Data management and instrument performance monitoring During this hackathon session the participants added novel functionality to assess the quality of an MS experiment to the Proline–MS-Angel proteomics management software system. First, an execution environment to run external scripts was added to MS-Angel to extract QC metrics from experimental raw files. Second, a semi-supervised approach to discriminate between high-quality and low-quality experiments was implemented in MS-Angel [8]. Third, the session participants established a roadmap to implement further QC features to Proline and MS-Angel.

3.2 Workflows, protocols, and guidelines

Implementation of software protocols in computational proteomics During this hackathon session the participants created a framework to implement fully documented and interactive protocols describing how to successfully carry out popular workflows to analyze MS data. Controlled environments in which to perform specific tasks were created using Docker containers and Jupyter notebooks to allow the full reproducibility of analysis pipelines and workflows.

Third-party tool integration and method development in OpenMS The participants of this session first got an introduction to the OpenMS software platform [7]. Afterwards they developed their own plugins under the guidance of experienced OpenMS maintainers. Examples of new OpenMS plugins that were developed include the MaRaCluster algorithm for spectral clustering [9].

3.3 Quantification

Statistical modelling to improve the quantitative analysis of post-translationally modified peptides Using a recent phosphoproteomics dataset [6], the participants of this session evaluated three strategies for the differential analysis of PTMs: (i) based on modified peptides only, (ii) based on modified peptides and any unmodified peptides from the corresponding protein, and (iii) based on modified peptides, their unmodified counterparts, and any other unmodified peptides from the corresponding protein. For each of these three cases linear models were developed to describe the quantification of modified peptides under different conditions.

Novel algorithms for DIA-based label-free quantification During this hackathon session the participants created new algorithms for label-free quantification of data-independent acquisition datasets to be included in IsoQuant [1]. A density-based clustering approach was developed to group corresponding features across the retention time, mass, and drift time dimensions.

4 Conclusion and outlook

The inaugural edition of the EuBIC developer's meeting was a resounding success. In a follow-up survey all participants expressed their overall satisfaction with the meeting, with two thirds of the survey respondents giving it a perfect score. Participants especially indicated that they enjoyed the unique interactive nature of the hackathon sessions. As envisioned, the restricted number of attendees allowed many interactions and facilitated effective communication and collaboration.

Even though the EuBIC developer's meeting only ran for a few days significant progress was made during the hackathon sessions on all projects. We are encouraged by the productivity of the participants to start solving important problems in only a limited time. The hackathon groups have committed to continue their collaboration and complete their projects, which will hopefully lead to scientific publications and ultimately better software solutions for MS-based proteomics end users.

Encouraged by the enthusiastic support of the community we are already planning the next EuBIC Winter School, which will take place in January 2019 in Zakopane, Poland.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

Funding for the EuBIC developer's meeting 2018 was provided by: the Research Foundation – Flanders (FWO), Ghent University's Doctoral Schools, the University of Antwerp's Framework for Young Researchers (OJO), the Flemish government, and EuPA. Additional sponsoring was provided by Thermo Fisher Scientific and the Belgian Proteomics Association (BePA).

Special thanks goes out to prof. dr. Lennart Martens for supporting the EuBIC developer's meeting through his FWO Scientific Research Network grant "Novel Knowledge from Public Life Sciences Data". Additionally, we would like to thank Inge Huyghe from the Laboratory of Pharmaceutical Biotechnology, Ghent University for her guidance through all the administration involved. Finally, we would like to thank all EuBIC members who volunteered to help on all fronts, as well as all keynote speakers and participants who contributed to the success of the developer's meeting.

References

[1] Distler, U., Kuharev, J., Navarro, P., Levin, Y., et al. "Drift Time-Specific Collision Energies Enable Deep-Coverage Data-Independent Acquisition Proteomics." In: *Nat. Methods* 11.2 (Feb. 2014), pp. 167–170. DOI: 10.1038/nmeth.2767.

[2] Gatto, L. and Christoforou, A. "Using R and Bioconductor for Proteomics Data Analysis." In: *Biochim. Biophys. Acta BBA - Proteins Proteomics* 1844.1 (Jan. 2014), pp. 42–51. DOI: 10.1016/j.bbapap.2013.04.032.

[3] Horlacher, O., Lisacek, F., and Müller, M. "Mining Large Scale Tandem Mass Spectrometry Data for Protein Modifications Using Spectral Libraries." In: *J. Proteome Res.* 15.3 (Mar. 4, 2016), pp. 721–731. DOI: 10.1021/acs.jproteome.5b00877.

[4] Horlacher, O., Jin, C., Alocci, D., Mariethoz, J., et al. "Glycoforest 1.0." In: *Anal. Chem.* 89.20 (Oct. 17, 2017), pp. 10932–10940. DOI: 10.1021/acs.analchem.7b02754.

[5] Martens, L., Kohlbacher, O., and Weintraub, S. T. "Managing Expectations When Publishing Tools and Methods for Computational Proteomics." In: *J. Proteome Res.* 14.5 (May 1, 2015), pp. 2002–2004. DOI: 10.1021/pr501318d.

[6] Rabiee, A., Schwämmle, V., Sidoli, S., Dai, J., et al. "Nuclear Phosphoproteome Analysis of 3T3-L1 Preadipocyte Differentiation Reveals System-Wide Phosphorylation of Transcriptional Regulators." In: *PROTEOMICS* 17.6 (Mar. 2017), p. 1600248. DOI: 10.1002/pmic.201600248.

[7] Röst, H. L., Sachsenberg, T., Aiche, S., Bielow, C., et al. "OpenMS: A Flexible Open-Source Software Platform for Mass Spectrometry Data Analysis." In: *Nat. Methods* 13.9 (Aug. 30, 2016), pp. 741–748. DOI: 10.1038/nmeth.3959.

[8] Solovyeva, E. M., Lobas, A. A., Kopylov, A. T., and Gorshkov, M. V. "Semi-Supervised Quality Control Method for Proteome Analyses Based on Tandem Mass Spectrometry." In: *Int. J. Mass Spectrom.* 427 (Apr. 2018), pp. 59–64. DOI: 10.1016/j.ijms.2017.09.008.

[9] The, M. and Käll, L. "MaRaCluster: A Fragment Rarity Metric for Clustering Fragment Spectra in Shotgun Proteomics." In: *J. Proteome Res.* 15.3 (Mar. 4, 2016), pp. 713–720. DOI: 10.1021/acs.jproteome.5b00749.

[10] The, M., Edfors, F., Perez-Riverol, Y., Payne, S. H., et al. "A Protein Standard That Emulates Homology for the Characterization of Protein Inference Algorithms." In: *J. Proteome Res.* Article ASAP (Apr. 16, 2018). DOI: 10.1021/acs.jproteome.7b00899.

[11] Walzer, M., Pernas, L. E., Nasso, S., Bittremieux, W., et al. "qcML: An Exchange Format for Quality Control Metrics from Mass Spectrometry Experiments." In: *Mol. Cell. Proteomics* 13.8 (Aug. 1, 2014), pp. 1905–1913. DOI: 10.1074/mcp.M113.035907.

[12] Wang, X., Chambers, M. C., Vega-Montoto, L. J., Bunk, D. M., et al. "QC Metrics from CPTAC Raw LC-MS/MS Data Interpreted through Multivariate Statistics." In: *Anal. Chem.* 86.5 (Mar. 4, 2014), pp. 2497–2509. DOI: 10.1021/ac4034455.

[13] Willems, S., Bouyssié, D., David, M., Locard-Paulet, M., et al. "Proceedings of the EuBIC Winter School 2017." In: *J. Proteomics* 161 (May 24, 2017), pp. 78–80. DOI: 10.1016/j.jprot.2017.04.001.

[14] Zhang, B., Pirmoradian, M., Zubarev, R., and Käll, L. "Covariation of Peptide Abundances Accurately Reflects Protein Concentration Differences." In: *Mol. Cell. Proteomics* 16.5 (May 2017), pp. 936–948. DOI: 10.1074/mcp.O117.067728.