JOURNAL OF PROTEOMICS 87 (2013) 134–138

Available online at www.sciencedirect.com

SciVerse ScienceDirect

www.elsevier.com/locate/jprot

Letter to the Editor

# Computational proteomics pitfalls and challenges: HavanaBioinfo 2012 Workshop report

Yasset Perez-Riverol[a,b,*], Henning Hermjakob[b], Oliver Kohlbacher[c], Lennart Martens[d,o], David Creasy[e], Jürgen Cox[f], Felipe Leprevost[g], Baozhen Paul Shan[h], Violeta I. Pérez-Nueno[i,p], Michal Blazejczyk[j], Marco Punta[k], Klemens Vierlinger[l], Pedro A. Valiente[m], Kalet Leon[n], Glay Chinea[a], Osmany Guirola[a], Ricardo Bringas[a], Gleysin Cabrera[a], Gerardo Guillen[a], Gabriel Padron[a], Luis Javier Gonzalez[a], Vladimir Besada[a]

[a]Center for Genetic Engineering and Biotechnology, Ave 31 e/158 y 190, Cubanacán, Playa, Havana, Cuba
[b]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
[c]Center for Bioinformatics, Quantitative Biology Center, and Department of Computer Science, University of Tübingen, Germany
[d]Department of Biochemistry, Ghent University, Ghent, Belgium
[e]Matrix Science Ltd, London, UK
[f]Department for Proteomics and Signal Transduction, Max-Planck Institute for Biochemistry, Germany
[g]Laboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, Fiocruz, Paraná, Brazil
[h]Bioinformatics Solutions Inc., Waterloo, Ontario, Canada N2L 6J2
[i]INRIA Nancy-Grand Est, 615 Rue du Jardin Botanique, 54506 Vandoeuvre-lès-Nancy, France
[j]Beaulieu-Saucier Pharmacogenomics Centre, Montreal Heart Institute, Canada
[k]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK
[l]Molecular Medicine, AIT — Austrian Institute of Technology, A-1190 Vienna, Austria
[m]Center of Protein Studies, Faculty of Biology, University of Havana, 25 No 411, 10400 Havana, Cuba
[n]Systems Biology Department, Center of Molecular Immunology, Calle 216 esq 15, Atabey, Playa, Havana, Cuba
[o]Department of Medical Protein Research, VIB, Ghent, Belgium
[p]Harmonic Pharma, Espace Transfert, 615 rue du Jardin Botanique, 54600 Villers lès Nancy, France

## ARTICLE INFO

*Article history:*
Received 14 January 2013
Accepted 22 January 2013
Available online 29 January 2013

*Keywords:*
Bioinformatics workshop
Mass spectrometry
Course
Protein identification
Database searching
Proteomic repositories

## ABSTRACT

The workshop "Bioinformatics for Biotechnology Applications (HavanaBioinfo 2012)", held December 8–11, 2012 in Havana, aimed at exploring new bioinformatics tools and approaches for large-scale proteomics, genomics and chemoinformatics. Major conclusions of the workshop include the following: (i) development of new applications and bioinformatics tools for proteomic repository analysis is crucial; current proteomic repositories contain enough data (spectra/identifications) that can be used to increase the annotations in protein databases and to generate new tools for protein identification; (ii) spectral libraries, *de novo* sequencing and database search tools should be combined to increase the number of protein identifications; (iii) protein probabilities and FDR are not yet sufficiently mature; (iv) computational proteomics software needs to become more intuitive; and at the same time appropriate education and training should be provided to help in the efficient exchange of knowledge between mass

* *Corresponding author at:* Center for Genetic Engineering and Biotechnology, Ave 31 e/158 y 190, Cubanacán, Playa, Havana, Cuba. Tel.: +53 7 2718008; fax: +53 7 2736008.
E-mail address: yasset.perez@biocomp.cigb.edu.cu (Y. Perez-Riverol).


http://dx.doi.org/10.1016/j.jprot.2013.01.019

spectrometrists and experimental biologists and bioinformaticians in order to increase their bioinformatics background, especially statistics knowledge.



The workshop "Bioinformatics for Biotechnology Applications" (HavanaBioinfo 2012) of the Heberprot-P Havana 2012 International Congress was held on December 8th to 11th, 2012 in the hotel "Occidental Miramar" located within an elegant area in Havana, Cuba. The workshop was fully subscribed, with more than 45 attendees and sixteen speakers, participating in two poster sessions and a panel discussion. The bioinformatics workshop was characterized by an informal atmosphere; questions were welcome at any time. The attendees responded enthusiastically and the room was always full of participants. The workshop was organized in three sessions dedicated to "System biology resources", "Protein identification and quantitation" and "Molecular drug design".

On the first day, after a brief introduction and the welcome words to the invited speakers and the students by Yasset Perez-Riverol (Center for Genetic Engineering and Biotechnology — CIGB); Dr. Gerardo Guillen (research director at CIGB) described the history and current developments of Cuban biotechnology [1], with particular emphasis on CIGB results. Cuba's first success in producing a modern biotechnological product was achieved in 1981, when, with the support of the US cancer specialist Randolph Lee Clark and Prof. Kary Kantell from Finland, the production of interferon was realized. Also in 1981, the Cuban government began to step up a "closed-cycle biotechnology initiative" encompassing everything from conceptualization and in vitro and animal studies up to clinical trials, commercialization, and post marketing follow-up. All invited speakers were surprised about the impact of the Cuban products on the health care system [1] and the current product pipeline of CIGB, particularly the Heberprot-P results [2].

Current developments of system biology tools at CIGB were the topic covered by Ricardo Bringas (CIGB). BisoGenet [3] is a Cytoscape plugin that constructs gene/protein networks containing information on molecular interactions and functional relations. It has a database (SysBiomics) to integrate information from multiple sources in a PostgreSQL warehouse. Henning Hermjakob (Proteomics Services, European Bioinformatics Institute — EBI) explained the EBI resources from Molecular Interactions (IntAct) [4] via curated human pathways (Reactome) [5] to Systems Biology Models (BioModels) [6]. Particularly, the Proteomics Standard Initiative (PSI) Common Query Interface (PSICQUIC) motivated an interesting discussion about remote access resources. PSICQUIC is a common computational interface for querying molecular interaction databases distributed worldwide [7]. It supports protein–protein and drug–target interactions and simplified pathway data. These widely used services distribute the efforts of curation, storage and support of molecular interaction databases, though central resources are still valuable and popular. Finalizing the system biology topic, Henning Hermjakob and Marco Punta described in detail the UniProt [8] and PFam resources [9].

Baozhen Paul Shan (Bioinformatics Solutions Inc.) described the history, theory, and practice of *de novo* identification strategy. The speaker demonstrated the actual scoring algorithm in PEAKS, and explained the fundamentals without losing the non-mathematicians in the audience. The integration of database search strategy and *de novo* sequencing in PEAKS increases the number of identifications compared to current database search tools of better accuracy and sensitivity [10]. Paul described that multiple-round search strategies decrease the efficiency of the target-decoy FDR estimation. The reason is that after the first round, there will be more target proteins than the decoys in the short list, and then if the second round search makes a mistake, it causes an FDR underestimation [10]. In PEAKS, a new approach, called decoy fusion, is used. Instead of mixing the target and decoy databases, the approach appends a decoy sequence to each target protein. After the fast search round, the protein shortlist will still contain the same number of target and decoy sequences. Spectra with highly confident *de novo* sequence tags but no significant database matches are extensively analyzed and the posttranslational modification module identifies peptides with unexpected modifications.

Continuing the *de novo* theme, Felipe Leprevost (Fiocruz, Brazil) explained the PepExplorer application, an integrated system to organize and statistically filter *de novo* sequencing results. The integration in one workflow, using the database search strategy and the *de novo* algorithm pepNovo, increases the number of peptides and proteins identified.

In the afternoon, Lennart Martens (Ghent University and VIB) talked about the "CompOmics toolsuite". During the last twelve years Lennart's group has developed a broad set of Java tools for proteomics data analysis. The source code, documentation and a complete set of examples for the main code library are freely available at http://compomics-utilities.googlecode.com [11]. The group also produces a set of parsers for popular search engine output files (Mascot [12], X!Tandem [13], OMSSA [14] and Proteome Discoverer [15]). It also includes a collection of user-friendly tools, including: (i) ms_lims [16] and DBToolkit [17] for storing and performing different *in silico* analysis of proteomics data; (ii) Peptizer [18] for manual validation of MS/MS search results; (iii) Rover [19], for visualizing and validating quantitative proteomics data; (iv) FragmentationAnalyzer [20] for analyzing MS/MS fragmentation data; and (v) the new SearchGUI [21] and PeptideShaker (http://code.google.com/p/peptide-shaker/) combination, for comprehensive MS data analysis and visualization. Finally, ProteoCloud (http://proteocloud.googlecode.com) [41] is a tool for easily running exhaustive searches on large datasets in the cloud. A short discussion ensued about the future infrastructure for identification tools, especially using graphical processing units (GPU), dedicated servers and cloud technologies. In the near future dedicated servers will continue as the main computer technology for protein identification, while at the same time more computational proteomic tools will be available in the cloud, especially those that allow meta-proteome analysis. With the whole audience thinking about how we can start to use Lennart's tools, his second talk called "mining the public proteome" gave excellent examples on how to combine the various proteomics informatics tools in large-scale data analysis.

The first example is the study of the limitations in resolution of the current search engines combined with the common decoy database design [22]. Lennart described the problem to detect small permutations or variations in the sequence with current search engines and decoy designs. A solution he proposed, currently in the final stages of development, consists of a rescoring system that considers the intensity of the spectrum signals as well as their m/z. Structural mapping of protein modifications was the second example of his latest research [23].

Closing the first day, Klemens Vierlinger (Health & Environment Department/AIT/Vienna, Austria) described the current challenges in meta-analysis and data integration in biomarker discovery, especially in human fibrotic disease. After the afternoon coffee break, the poster session included the discussion of eleven posters by students from Cuba, Mexico and Colombia.

The second day was dedicated entirely to protein identification strategies and tools. The possibility to interact and discuss with David Creasy (Matrix Science, Mascot search engine [24]) and Jürgen Cox (Max Planck Institute, Martinsried, MaxQuant-Andromeda software [25]) about the scoring systems and platform fundamentals ensured a productive session. The Sunday opening lecture by Dr. Vladimir Besada described the computational proteomics tools developed at CIGB. The implementation of a new function to estimate the isoelectric point based on support vector machines for peptides and proteins [26], and the development of two algorithms for protein identification [27] and quantitation [28] are some of the most relevant results of the CIGB bioinformatics group.

David Creasy (Matrix Science) described the history, theory, and practice of Mascot search engine and tools. David pointed out some of the parameters in Mascot that may cause problems if not properly employed. For example, doing a non-enzyme search in Mascot is not a good idea unless there is a very high level of non-specific peptides expected in the sample. Semi-trypsin is almost always a better choice if the peptides came from a tryptic digest. Users interested in post-translational modifications, can use an error tolerant search [29] by checking the error tolerance box on the search form, rather than selecting more and more modification in the search. This is a much more efficient way to discover unusual modifications, as well as non-specific peptides and sequence variants. David also explained that one of the future very promising fields is the inclusion of spectral library search in the current proteomic workflows, as is already available through SpectraST [30] or X!Hunter [31].

The ensuing coffee break was particularly motivated by Mascot discussions, some of the non-answered questions were: Why is Mascot successful and extensively used even with the existence of different freely available tools such as X! Tandem, OMSSA and Andromeda?; How can the Mascot scoring system be at the same time powerful yet simple?; Why don't popular search engines consider the intensity of the signals in the scoring systems? The organizers decided to give an additional 10 min of coffee break time just to boost this dynamic and enthusiastic discussion environment.

Oliver Kohlbacher (University of Tübingen, Germany) introduced OpenMS [32] (http://OpenMS.de), a software framework for enabling rapid application development in mass spectrometry proteomics. It has been designed to be portable and robust, mainly developed in C++, while offering rich functionalities, ranging from the availability of basic data structures to sophisticated algorithms for data analysis. The framework architecture consists of several layers, a core application programming interface (API), which captures the MS data and complementary metadata, and a higher-level functionality API that contains database I/O, file I/O and other analysis algorithms. The framework contains several examples for extension and use of the libraries.

Based on OpenMS are the tools of the OpenMS Proteomics Pipeline (TOPP) [33], which is a suite of stand-alone tools suitable for the construction of high-throughput data analysis pipelines. TOPP tools interoperate easily with other software packages. They provide functionality for protein identification, protein/peptide quantification, and statistical analysis of these results. The quantitation tools allow the analysis of different samples using SILAC, iTRAQ, TMT and label-free experiments. OpenMS fully supports standard formats such as: mzIdentML, mzML and mzTab.

Jürgen Cox (Max-Planck Institute for Biochemistry, Munich, Germany) introduced the MaxQuant platform for high-resolution mass spectrometry experiments. Recent revolutionary advances in high accuracy mass spectrometry-based proteomics are providing a new basis for data-driven systems biology. Jürgen described the algorithms and whole workflows encompassing the mass spectrometry data analysis from intelligent data-driven acquisition, via algorithms for identification and quantification of proteins, to the statistical analysis of the final expression data for proteins and posttranslational modifications in the context of other omics and pathway data [34–36].

Before lunchtime, Henning Hermjakob described the current status of the proteomics repository services in the European Bioinformatics Institute. The PRotein IDEntification Database [37] started in 2005 and in the last update contains 11,629,064 identifications and 338,501,793 spectra, supporting the most common spectrum and identification file formats. PRIDE offers comprehensive support for the preparation of data deposition (Pride Converter 2 [38]) and data access (Pride Inspector [39]). PRIDE is a founding member of the ProteomeXchange consortium (http://www.proteomexchange.org) [40]. The members of the consortium, led by PRIDE and PeptideAtlas, are implementing a system to enable the automated and standardized sharing of MS-based proteomics data between the main existing MS proteomics repositories. One of the first question was: How useful are the current proteomics repositories?; Should the submission of the MS/MS data for publications be mandatory? Henning explained some of the ongoing projects with the current public data in PRIDE in the near future: PRIDE-Q for the provision of high quality subsets of the diverse PRIDE data, aiming to develop PRIDE from a specialist mass spectrometry resource into a protein expression resource for systems biology, and PRIDE-C, the development of PRIDE spectral libraries. In the discussion, the main consensus was that nowadays the submission of proteomics data supporting publications to public repositories should be mandatory. Data availability in public repositories allows the control of the published results and the reproducibility of the experiments. After lunch, all the invited speakers and students made a quick visit to the Old Havana City.

The last day was entirely dedicated to molecular drug design and chemoinformatics. The opening lecture entitled "Rational design of peptide inhibitors against Dengue virus" was given by Glay Chinea (CIGB). An overview regarding the Dengue virus, its prevalence and typical clinical outcomes was first introduced. Then, he presented the results which lead to the design of virus entry inhibitors based on a comprehensive approach combining a number of methods from bioinformatics, structural biology, computer simulations of molecular interactions and rational peptide and protein design.

Violeta Perez-Nueno from Orpailleur Team (INRIA Nancy) presented several approaches that can be used to model molecular interactions and more deeply a new 3D shape-based approach for predicting and quantifying drug promiscuity by correlating Gaussian clusters of ligand spherical harmonic shapes. The presentation entitled "Epitope-based vaccines — From high-throughput data to individualized therapies" by Oliver Kohlbacher triggered an enthusiastic exchange of ideas. Epitope-based vaccines (EVs) have recently been attracting growing interest. The success of an EV is determined by the choice of epitopes used as a basis. Numerous *in silico* approaches exist that can guide the design of EVs. In particular, computational methods for MHC binding prediction have already become standard tools in immunology. Oliver discussed some of the obvious problems in the design of EVs. After lunch, a conference entitled "Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions" was given by Marco Punta. Sequence alignment programs may miss or misidentify homologous relationships between proteins based on different factors, including homologous overextension and convergent evolution (as observed in compositionally biased amino acid regions). He presented a study where the Pfam collection of manually curated profile-hidden Markov models is used to test the accuracy with which the alignment program HMMER3 assigns protein sequences to homologous families. The last talk of the session aimed at presenting Montreal Heart Institute's and Pharmacogenomics Centre's Cardiovascular Genetics Clinic (CGC). Michal Blazejczyk's project is to identify deleterious mutations in patients suffering from several cardiovascular conditions (cardiomyopathies and arrhythmias) by using targeted resequencing (Sanger). High standards of quality assurance, quality control, standard operating procedures, and report approval rules are used to ensure that reports delivered to clinicians contain accurate data.

Concluding the workshop a "panel discussion" about some of the challenges and pitfalls in computational proteomics was realized. One of the most provocative ideas was "experimentalists (biologist or spectrometrist) need to realize that they are doomed without bioinformatics and statistical knowledge in current science". The abundance of information presents many hurdles to the investigators who need to interpret the data and derive new biological insights. Therefore, efforts need to focus on two directions: (i) computational proteomics software needs to become more intuitive; and (ii) bioinformatics and statistics knowledge is mandatory to understand the overall behavior of biological results. New bioinformatics applications need to be more professional in terms of error-tolerance, usability, and data integration; and researchers in life sciences need to take bioinformatics and statistics courses.

Major conclusions of the workshop included: ((i) development of new applications and bioinformatics tools for proteomic repository analysis is crucial; current proteomic repositories contain enough data (spectra/identifications) that can be used to increase the annotations in protein databases and to generate new tools for protein identification; (ii) spectral libraries, *de novo* sequencing and database search tools should be combined to increase the number of protein identifications; (iii) protein probabilities and FDR are not yet sufficiently mature; (iv) computational proteomics software needs to become more intuitive; and at the same time appropriate education and training should be provided to help in the efficient exchange of knowledge between mass spectrometrists and experimental biologists and bioinformaticians in order to increase their bioinformatics background, especially statistics knowledge.

## Acknowledgments

The workshop was supported by The International Centre for Genetic Engineering and Biotechnology, Trieste, Italy and The Center for Genetic Engineering and Biotechnology, Havana, Cuba.

## REFERENCES

[1] Cuba's biotech boom. Nature 2009;457:130.
[2] Fernandez-Montequin JI, Betancourt BY, Leyva-Gonzalez G, Mola EL, Galan-Naranjo K, Ramirez-Navas M, et al. Intralesional administration of epidermal growth factor-based formulation (Heberprot-P) in chronic diabetic foot ulcer: treatment up to complete wound closure. Int Wound J 2009;6:67–72.
[3] Martin A, Ochagavia ME, Rabasa LC, Miranda J, Fernandez-de-Cossio J, Bringas R. BisoGenet: a new tool for gene network building, visualization and analysis. BMC Bioinformatics 2010;11:91.
[4] Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, et al. The IntAct molecular interaction database in 2012. Nucleic Acids Res 2012;40:D841–6.
[5] Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res 2011;39:D691–7.
[6] Li C, Donizelli M, Rodriguez N, Dharuri H, Endler L, Chelliah V, et al. BioModels database: an enhanced, curated and annotated resource for published quantitative kinetic models. BMC Syst Biol 2010;4:92.
[7] Aranda B, Blankenburg H, Kerrien S, Brinkman FS, Ceol A, Chautard E, et al. PSICQUIC and PSISCORE: accessing and scoring molecular interactions. Nat Methods 2011;8:528–9.
[8] Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 2012;40:D71–5.
[9] Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res 2012;40:D290–301.
[10] Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, et al. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. Mol Cell Proteomics 2012;11 [M111 010587].
[11] Barsnes H, Vaudel M, Colaert N, Helsens K, Sickmann A, Berven FS, et al. Compomics-utilities: an open-source Java

library for computational proteomics. BMC Bioinformatics 2011;12:70.
[12] Helsens K, Martens L, Vandekerckhove J, Gevaert K. MascotDatfile: an open-source library to fully parse and analyse MASCOT MS/MS search results. Proteomics 2007;7:364–6.
[13] Muth T, Vaudel M, Barsnes H, Martens L, Sickmann A. XTandem Parser: an open-source library to parse and analyse X!Tandem MS/MS search results. Proteomics 2010;10:1522–4.
[14] Barsnes H, Huber S, Sickmann A, Eidhammer I, Martens L. OMSSA Parser: an open-source library to parse and extract data from OMSSA MS/MS search results. Proteomics 2009;9:3772–4.
[15] Colaert N, Barsnes H, Vaudel M, Helsens K, Timmerman E, Sickmann A, et al. Thermo-msf-parser: an open source Java library to parse and visualize Thermo Proteome Discoverer msf files. J Proteome Res 2011;10:3840–3.
[16] Helsens K, Colaert N, Barsnes H, Muth T, Flikka K, Staes A, et al. ms_lims, a simple yet powerful open source laboratory information management system for MS-driven proteomics. Proteomics 2010;10:1261–4.
[17] Martens L, Vandekerckhove J, Gevaert K. DBToolkit: processing protein databases for peptide-centric proteomics. Bioinformatics 2005;21:3584–5.
[18] Helsens K, Timmerman E, Vandekerckhove J, Gevaert K, Martens L. Peptizer, a tool for assessing false positive peptide identifications and manually validating selected results. Mol Cell Proteomics 2008;7:2364–72.
[19] Colaert N, Helsens K, Impens F, Vandekerckhove J, Gevaert K. Rover: a tool to visualize and validate quantitative proteomics data from different sources. Proteomics 2010;10:1226–9.
[20] Barsnes H, Eidhammer I, Martens L. FragmentationAnalyzer: an open-source tool to analyze MS/MS fragmentation data. Proteomics 2010;10:1087–90.
[21] Vaudel M, Barsnes H, Berven FS, Sickmann A, Martens L. SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. Proteomics 2011;11:996–9.
[22] Colaert N, Degroeve S, Helsens K, Martens L. Analysis of the resolution limitations of peptide identification algorithms. J Proteome Res 2011;10:5555–61.
[23] Vandermarliere E, Martens L. Protein structure as a means to triage proposed post-translational modification sites. Proteomics submitted for publication. http://dx.doi.org/10.1002/pmic.201200232.
[24] Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 1999;20:3551–67.
[25] Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 2008;26:1367–72.
[26] Perez-Riverol Y, Audain E, Millan A, Ramos Y, Sanchez A, Vizcaino JA, et al. Isoelectric point optimization using peptide descriptors and support vector machines. J Proteomics 2012;75:2269–74.
[27] Sanchez A, Perez-Riverol Y, Gonzalez LJ, Noda J, Betancourt L, Ramos Y, et al. Evaluation of phenylthiocarbamoyl-derivatized peptides by electrospray ionization mass spectrometry: selective isolation and analysis of modified multiply charged peptides for liquid chromatography-tandem mass spectrometry experiments. Anal Chem 2010;82:8492–501.
[28] Fernandez-de-Cossio J, Gonzalez LJ, Satomi Y, Betancourt L, Ramos Y, Huerta V, et al. Isotopica: a tool for the calculation and viewing of complex isotopic envelopes. Nucleic Acids Res 2004;32:W674–8.
[29] Creasy DM, Cottrell JS. Error tolerant searching of uninterpreted tandem mass spectrometry data. Proteomics 2002;2:1426–34.
[30] Lam H, Deutsch EW, Eddes JS, Eng JK, King N, Stein SE, et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics 2007;7:655–67.
[31] Craig R, Cortens JC, Fenyo D, Beavis RC. Using annotated peptide mass spectrum libraries for protein identification. J Proteome Res 2006;5:1843–9.
[32] Sturm M, Bertsch A, Gropl C, Hildebrandt A, Hussong R, Lange E, et al. OpenMS — an open-source software framework for mass spectrometry. BMC Bioinformatics 2008;9:163.
[33] Bertsch A, Gropl C, Reinert K, Kohlbacher O. OpenMS and TOPP: open source software for LC–MS data analysis. Methods Mol Biol 2011;696:353–67.
[34] Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, et al. Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol 2011;7:548.
[35] Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res 2011;10:1794–805.
[36] Michalski A, Cox J, Mann M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC–MS/MS. J Proteome Res 2011;10:1785–93.
[37] Vizcaino JA, Cote RG, Csordas A, Dianes JA, Fabregat A, Foster JM, et al. The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res 2013;41. http://dx.doi.org/10.1093/nar/gks1262 [Database issue D1063 -D1069].
[38] Cote RG, Griss J, Dianes JA, Wang R, Wright JC, van den Toorn HW, et al. The PRIDE Converter 2 framework: an improved suite of tools to facilitate data submission to the PRIDE database and the ProteomeXchange consortium. Mol Cell Proteomics Dec 2012;11(12):1682–9. http://dx.doi.org/10.1074/mcp.O112.021543.
[39] Wang R, Fabregat A, Rios D, Ovelleiro D, Foster JM, Cote RG, et al. PRIDE Inspector: a tool to visualize and validate MS proteomics data. Nat Biotechnol 2012;30:135–7.
[40] Hermjakob H, Apweiler R. The Proteomics Identifications Database (PRIDE) and the ProteomExchange Consortium: making proteomics data accessible. Expert Rev Proteomics 2006;3:1–3.
[41] Muth T, Peters J, Blackburn J, Rapp E, Martens L. ProteoCloud: a full-featured open source proteomics cloud computing pipeline. J Proteomics in press. http://dx.doi.org/10.1016/j.jprot.2012.12.026 [pii: S1874-3919(13)00013-4].