JOURNAL OF PROTEOMICS 88 (2013) 41–46

Available online at www.sciencedirect.com

SciVerse ScienceDirect

www.elsevier.com/locate/jprot

Discussion

PROTEINCHALLENGE: Crowd sourcing in proteomics analysis and software development☆

Sarah F. Martin[a], *Heiner Falkenberg*[b], *Thomas F. Dyrlund*[c], *Guennadi A. Khoudoli*[d], *Craig J. Mageean*[e], *Rune Linding*[f,*]

[a]*Kinetic Parameter Facility, Centre for Synthetic and Systems Biology (SynthSys), University of Edinburgh, UK*
[b]*Molecular Proteomics Laboratory, Biological-Medical-Research Centre, Heinrich-Heine-University Düsseldorf, Germany*
[c]*Interdisciplinary Nanoscience Centre (iNANO) and Department of Molecular Biology, Aarhus University, Denmark*
[d]*Wellcome Trust Centre for Gene Regulation and Expression, College of Life Sciences, University of Dundee, UK*
[e]*Physiological laboratory, Institute for Translational Research, University of Liverpool, UK*
[f]*Cellular Signal Integration Group (C-SIG), Centre for Biological Sequence Analysis (CBS), Department of Systems Biology, Technical University of Denmark (DTU), Building 301, DK-2800 Lyngby, Denmark*

ARTICLE INFO

Available online 7 December 2012

Keywords:
Crowd sourcing
Community challenge
Data analysis
Software
Benchmarking
Open source

ABSTRACT

In large-scale proteomics studies there is a temptation, after months of experimental work, to plug resulting data into a convenient—if poorly implemented—set of tools, which may neither do the data justice nor help answer the scientific question. In this paper we have captured key concerns, including arguments for community-wide open source software development and "big data" compatible solutions for the future. For the meantime, we have laid out ten top tips for data processing. With these at hand, a first large-scale proteomics analysis hopefully becomes less daunting to navigate.

However there is clearly a real need for robust tools, standard operating procedures and general acceptance of best practises. Thus we submit to the proteomics community a call for a community-wide open set of proteomics analysis challenges—PROTEINCHALLENGE—that directly target and compare data analysis workflows, with the aim of setting a community-driven gold standard for data handling, reporting and sharing.

This article is part of a Special Issue entitled: New Horizons and Applications for Proteomics [EuPA 2012].

1. Introduction

Compared to other fields such as transcriptomics, computational biology, genomics and imaging, bioinformatics activities in mass spectrometry-based proteomics have been rather vendor specific, closed sourced and not involving large parts of the community [1–3]. This is in stark contrast to highly successful academic open source software solutions in other fields, such as Cytoscape.org, Openmicroscopy.org, Ensembl.org and BLAST. Academic software development efforts in proteomics have tended to be based within large individual labs, and only a relatively few have been shared openly and more widely. Despite

☆ This article is part of a Special Issue entitled: New Horizons and Applications for Proteomics [EuPA 2012].

* *Corresponding author*. Tel.: +45 4525 8611, +45 2365 1941(mobile); fax: +45 4593 1585.
E-mail addresses: linding@cbs.dtu.dk, proteinchallenge@gmail.com (R. Linding).

1874-3919/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jprot.2012.11.014

the lack of open source development, there have been successes, such as the broad community utilisation of academic—if closed-source—tools such as MaxQuant [4].

In this paper we share the fruit of the round-table discussion group at EUPA2012 on the topic of "Dealing with the Data Mountain". Participants included end-users, statisticians, programmers, software developers, biologists and medical and drug discovery researchers from both industry and academia. Firstly we will discuss the current problems with data analysis in proteomics and justify the case for community-based open source software development; secondly we propose a set of 10 tips to enable the robust analysis of data with current tools and lastly we propose a community-based challenge solution to enable the advancement of proteomics tools.

2. The case for open source software tools

The first issue to address as a community is to make proteomics analysis tools open source. The reason this is essential is that publication of scientific data should be accompanied by full disclosure of how they were analysed and processed. This ultimately requires access and insight into the source code, as even with full disclosure of parameter settings and algorithms, results will not be reproducible on an alternative platform. A serious concern with closed-source platforms is the omnipresent nature of software bugs. Imagine a bug in Mascot which affects search results: such an error could impact on hundreds of publications in our field. Such a scenario is not unheard of: in 2008 a bug was identified in the BLAST matrices [5]. While some may argue that closed-source code is better protected from the introduction of bugs, the bug in BLAST was in fact only discovered due to the global availability of the source code. The fear that open source code may lead to uncontrolled development and different spin-off versions on the other hand, can be fully controlled in both the method by which the code is shared and by licensing. It may be useful if journals adopt open access to underlying analysis source code as a requirement for publication of data.

Second, instead of having the entire responsibility for the flawlessness of a tool such as MaxQuant within just one or a handful of labs, it is common knowledge that far cleaner, better performing, less-error prone, faster evolving and more popular software arises from development in communities and more diverse groups. For example, software solutions designed initially for a single vendor's output files could be made compatible with other vendor-specific formats much faster if implemented as a community-wide effort, following the full disclosure of the source code. In academia the optimal way forward is through open source development, like for example the LINUX kernel and Cytoscape.org.

Finally, an ever pending issue is that of maintenance and funding of software. It is a big burden for any one lab (no matter how large) and again is better handled by community efforts or dedicated consortia. Cytoscape.org for example enables plugins from community projects, a powerful and embracing strategy. This does however not replace the need for full open-source access as discussed above.

3. The "big data" future

There are several technical reasons why a more broad development scenario is attractive. There is a lack of cross-platform deployment options and of high-performance computing (HPC) capable tools. In our experience, while some developers (e.g. Mascot) claim they support HPC operating systems such as Linux, the performance on these platforms is poor and not up to par with larger instrument fleets. Furthermore, the dependency of some tools on the Windows operating system makes them incompatible with high-performance pipelines. Deployment of tools in a pipeline mode is critical for any large scale proteomics project.

Automation of workflows including searching, fraction and sample handling and tracking is important for system-scale biology projects, such as implemented in the PepTracker project (www.peptracker.com). This is also critical for environments using laboratory information management systems and bar-coding for processing of large sets of samples or where tracking is mission critical, e.g. in clinical analysis. Such developments would also benefit the adoption of Standard Operating Procedures (SOPs) outside sample preparation, in the actual analysis of data and in the representation and sharing of results.

The reliance on workstation-centric software is also in direct conflict with robust storage solutions capable of handling terabytes of data with automated backup, versioning control systems and low-energy consuming data migration. Biology and in particular genomics and proteomics are 'big data' sciences and will benefit from adopting technologies used in other 'big data' fields such as physics, finance and engineering where HPC and open source tools are the norm rather than exception.

4. Sharing and collaboration tools

What happens with proteomics data once they are published in a public database? From a statistical point of view, could these be re-used? How are data shared with collaborators? These are all important and somewhat open questions.

The easiest way to share data is still via old-school (S)FTP, shipping of hard-drives or HTTP. However, with the ever increase of cheap cloud computing (for example Amazon or Google) data sharing via online accounts is increasingly popular. In terms of long-term public storage of proteomics data, the PRIDE database is a primary repository, much like PDB is for structural data. The future of alternative repositories (e.g. Proteome Commons) is currently uncertain which leaves PRIDE as the key global database for mass spectrometry data [1]. The PRIDE database provides statistical tools for using and building upon past data. While some participants still found it hard to get data into PRIDE it seems that this is currently the only way to ensure long-term public storage and increased likelihood of the data being re-used by others.

It is critical to release the RAW files alongside publications, as these contain some of the important parameters of the experiment necessary for reproducibility, such as measuring strategies, run-times and ideally HPLC gradients. Releasing

this information should be compulsory for any publication. The potential for re-use of data is particularly important since in the main text of publications authors typically discuss a subset of proteins (e.g. concerning a particular pathway), and only significantly up/down-regulated findings. Placing the full list of identified proteins, whether significant under treatment or not, in the public domain provides a potential source of data for similar studies and avoids duplication of expensive and time-consuming experiments. Identifying proteins that were not significantly altered following a particular treatment could be of great interest to other labs. A further reason for the publication of all identifications is that EBI is assembling such published information as evidence of expression in protein databases.

5. Ten tips for data processing in proteomics

While we are waiting for a community-wide, open source and high-performance analysis package to become available for any type of infrastructure, what can be done now with existing tools, and what are the guidelines that result in an optimal analysis of data? Which of the many tools out there can we rely on?

Here we provide a set of tips in the order of a typical analysis workflow:

1. Use a platform-specific peak selection and inspect spectra manually

A first key point is that if a peptide is not detected in a proteomics dataset, this does not mean it is not present in the sample. Efficient peak selection is the first critical analysis step to maximise identifications. It can be hard to navigate the tool jungle, and the choice of several tools claiming to be the best for the job is not always helpful to the end-user. This may be the result of the never ending search for better performance and a reflection of different users' requirements. In our experience, instrument producers have optimised platform-specific solutions for peak selection, and recommend relying on instrument specific software for this step. We do however suggest to always check a statistical representative set of spectra manually. While this inspection step may sound very old-school and tedious, we would argue it is a sound way to spot any systematic errors. You may want to check all spectra in your final set of proteins/peptides, or pick a random representative subset. We suggest that it would be useful to have more versatile spectra browsers available, also on mobile platforms to enable on-the-move browsing while streaming data.

2. Combine search tools for peptide identification

In a typical shotgun proteomics experiment, MSMS data is searched against protein databases to gain peptide identifications, using dedicated software (e.g. MASCOT, Andromeda or SEQUEST). While there is no single top-performing, high-performance and open source search engine out there yet, there has been a recent trend to utilize multiple search engines and combine the results [6]. Crowd sourcing of search algorithms has been shown to outperform individual algorithms [7]. But how are multiple search results best combined? One successful approach is by probabilistic integration, as the tools Scaffold [8], TPP [9] and Peptide Shaker perform (peptide-shaker.googlecode.com). While this type of meta-searching can increase the coverage and decrease the errors in identified peptides, there is still a need for new search algorithms to be developed to go beyond the performance of current tools.

3. Beware of FDR-assigned modification sites

The false discovery rate (FDR) of peptide identifications in database searches is typically assessed in comparison to a search of the same data against a reversed, or otherwise artificially created, decoy database. While FDR calculations are a useful and legitimate tool for peptide identification purposes, a big issue with the use of decoy searches arises when assigning confidence scores and localization prediction to post-translational modification sites. The problem is that basing statistical significance on a null-model that utilizes peptide sequences that do not exist in reality is inherently flawed.

The cut off set for FDR is often 1% by default, however to avoid losing identifications, a cut-off of 2–3% is more advisable [10].

There are two ways to improve this issue:

1) using novel mass-spectrometers and technologies such as ion mobility may resolve some localization problems
2) searching for numerous PTMs for the same set of spectra and then estimate a p-value for the best matching modification/localisation. This approach is very heavy computationally and would require streamlined search engines. The error tolerant search from Mascot or the Preview software [11] can be used for this, but these are not an open-source.

4. Customise the most suitable peptide quantification method

Peptide quantification has been one of the most successful aspects of mass spectrometry-based proteomics in recent years. Software solutions for SILAC [12,9], stable isotope [13], TMT and iTRAQ [14] labelled samples now enable convenient and robust relative quantification of pooled samples. Great improvements in label-free and spectral counting methods [15] now enable the relative quantification in silico without the need for labelling or pre-processing. Several targeted proteomics (MRM/SRM/SWATH) tools such as SkyLine [16] have also been developed.

These tools are commonly complemented with in-house scripting based pipelines using Python, Perl, R and SQL for large projects where full control and automation of the quantification workflow are required. It is also important to consider proline and arginine conversion in SILAC quantification workflows, often this can be corrected for by custom developed R-scripts.

The choice of quantification strategy, whether labelled or not, must be part of the earliest stages of experimental planning, with considerations including desired proteome coverage, expected fold-changes in signals and statistical power (replicates required to achieve significant results at the desired sensitivity), as well as expertise—quantification is a specialist topic and experience in applying a technique will influence reproducibility [17].

These applications are increasingly adapted to enable novel proteome-wide measurements beyond plain intensity

comparisons between samples, two examples being quantitative measurements of protein turnover [18,19,13] and sub-cellular localisation [20].

5. Use one sequence identifier type and stick to it

This is a never-ending, time consuming issue in almost any project involving bioinformatics. The short answer is 'pick one and stick to it across the project'. I.e. select a well annotated database and use it in all aspects of the project. It is a waste of time to try to translate across databases and will result in data loss and corruption, no matter how robust a scheme is used. We also note that it is important to use representative genome-wide databases which have developed highly robust mapping/linkage schemes, for example ENSEMBL at EBI, which ensure data linkage and integrity.

6. Choose network mining tools over pathway analysis

A frequently asked question is whether to perform network or so-called 'pathway' analysis and if it is worth investing in commercial pathway tools [21]. The recommendation is a clear yes on network analysis but no to existing 'pathway tools'. The main problem is that unlike metabolic pathways, protein interactions cannot be described as linear pathways [22–24]. In addition to this we know very little about signalling networks, and in particular their dynamics, which are essential to their impact on the cell phenotype. Thus relying on tools that claim they have a 'pathway knowledgebase' for proteins is fundamentally flawed.

That said, it does make sense to use probabilistic resources that integrate the 'entire scientific literature' such as STRING [25] and GeneMania [26]. These tools attempt to comprehensively index all reported interactions and functional associations between proteins across all sequenced species. As such these tools by far outperform 'pathway tools' in terms of sheer coverage of the interactome. They also have the advantage of assigning confidence scores to any reported interaction which gives the analysis a probabilistic basis which is important. In addition to this we recommend using NetPhorest [27] and NetworKIN [28] to model and reconstruct networks specifically for phospho-proteomics data.

7. Integrate your data with other sources

Once we have a network model of identified proteins, peptides and PTMs it is critical to visualize it and to perhaps overlay with other types of data (not necessarily from proteomics). This is currently best facilitated by Cytoscape.org. While the interface can be slow and complicated at times, it is the most advanced network tool available and once basic tasks such as file-based colouring and node-shaping are mastered, it is relatively easy to write own scripts to create input files for it. Cytoscape also has a plethora of plug-ins that allow sub-graph and functional enrichment analyses. More features are continuously added. Cytoscape integrates many different types of data and once a network model is set up it can be exported as a vector graphic and customised for publication in a drawing tool.

8. Use only experimentally evidenced GO annotation

GO annotation [29] has become cyclic, with electronically inferred terms propagating uncontrolled errors. GO-annotated results should always be filtered based on experimental data, using e.g. UniProt for filtering and PantherDB for functional searching.

At the end of the day we are more interested in mechanistic network dynamics rather than a pie diagram showing 10% transcription. GO may be useful for categorising data but not for drawing conclusions.

9. Make use of packages in R, Perl and Python

It is also a matter of respecting data. While people happily spend months preparing samples, often the data analysis can't go fast enough. Data which was so carefully gathered deserves a careful and customised analysis to help answer the biological questions at hand.

10. Test multiple modelling approaches

While a discussion of computational modelling of e.g. cell signalling networks is beyond the scope of this write-up, we did raise some general points which may help people starting out:

1) Modelling greatly depends on the question at hand, and which data are available to answer it. Thus while several modelling methods have been used to analyse mass spectrometry data in the past with some degree of success [30], it does not mean that they will be applicable more generally in other cases.
2) There is no 'best' or 'only approach', and often several options will work equally well. In fact tools like e.g. DataRail [31] enable trials of different approaches. Ideally one would sample across many strategies and pick the best performing one and—crucially—determine whether this choice is statistically robust.
3) In general probabilistic approaches are preferred as they give confidence values for decision making. Also these are much less dogmatic in nature compared to modelling approaches that assume we know the system. For example ordinary differential equation (ODE) modelling may perform poorly if not all rate constants are known. Furthermore, the concentration of proteins in cells can be highly localized and interactions resemble solid phase chemistry more than diffusion limited reactions, yielding measured rate constants irrelevant [32].
4) It is likely that some development in house will be required, but this may result in new features being discovered in data.

Approaches such as Principal Component Analysis (PCA) coupled with Partial Linear Regression (LDR) modelling have been used in different projects which have provided new biological insight [33]. However, there are also weaknesses with these approaches as they are linear and tend to scale-up noise in the data. Thus there is a need for new non-linear approaches and for hybrid approaches aiming to use probabilistic suppression of noise.

6. Crowd sourcing initiatives

Comparisons of performance in both computational and experimental approaches and tools are essential to technological and scientific progress. This has been shown repeatedly in structural biology, bioinformatics, interactome screening, RNAi screening and even in experimental proteomics. Some of these

efforts have been conducted as individual studies, while others have been formalized into global, open and highly productive challenges. In the latter category the CASP (structural biology) [34] and DREAM (network biology) [35] competitions are two bright success stories. The CASP competition has for years powered new innovation in algorithms for structural biology. DREAM, a younger initiative, has already led to impressive new tools and knowledge about the effectiveness of crowd sourcing algorithms (something also known from CASP in the so-called meta-servers). Proteomics needs new algorithms and not just meta-servers. New innovative solutions are essential to prevent stagnation. While crowd sourcing can outperform existing individual tools they can only keep doing so if new innovative algorithms are created.

As a spinoff from DREAM an initiative called IMPROVER [36] has been launched which aims at conducting biological challenges based on systems biology approaches. Thus there are now community efforts in trying to identify network based biomarkers and assessing the cross-species utility of data obtained in model organisms. In addition IMPROVER and DREAM are actively exploring different types of benchmarking approaches while annual competitions are held in modelling of networks and other types of systems biology analysis.

7. PROTEINCHALLENGE—call for community-wide open analysis challenges

Here we call for a global sister project to DREAM and IMPROVER in the proteomics community, PROTEINCHALLENGE. We suggest that this should be held as part of HUPO. It is critical that we, as a community, begin to systematically test datasets, search algorithms and platform robustness [37]. Just like DREAM we will use community shared 'golden data sets' for blinded benchmarking of search algorithms across platforms and data types (e.g. labelled versus non-labelled). Both academia and industry will be invited to participate as is the case in IMPROVER.

We suggest that PROTEINCHALLENGE be launched as soon as possible and contain sub-competitions in areas such as:

Experimental challenge:

- sample enrichment
- labelling technologies
- platform comparisons

Data analysis/pipeline challenge:

- search algorithms
- localization scores
- decoy methods
- network reconstruction algorithms
- modelling of proteomics data
- protein/peptide identification algorithms
- quantification algorithms

The winning prize will receive a vendor sponsored research prize, which will encourage both participation and further development. It will be important to collate results across different labs, using samples and data sets sent out specifically for this golden test [1]. Multi-lab studies are regularly performed with experimental samples [38] but have not yet contrasted or defined a best-practice workflow for proteomics analysis pipelines.

The foundation of PROTEINCHALLENGE will be robust benchmarking and performance evaluation criteria to ensure the fair, robust and accurate comparison of individual tools as well as declaration of winners. This will involve rigorous probabilistic assessments using redundancy reduced data and round-robin schemes for cross-validation. It is essential that tools participating in the competition release their source-code and algorithms and provide full documentation on their correct use. While this requirement would initially exclude a proportion of existing tools, it will in the long term encourage adoption of successful tools and spawn enhanced developments. It will also provide a much speedier innovation within our community. The PROTEINCHALLENGE will ensure that tools do not disappear off the radar after publication as they will be applied by end-users during the challenge. As part of the challenge a WIKI will be maintained (similar to DREAM) where an updated list of top-performing tools in the different categories will be listed. Researchers can then add their tools to the list and a status field will show if the tool has been evaluated in the PROTEINCHALLENGE. Tracking of parameters and new versions of software will also be important to report to the end-user. This will be a publically accessible wikipedia page rather than a restricted mailing list.

The definition of requirements for standard data sets, in writing, is an important first step for the challenge. We propose this be an aim of an inaugural workshop at HUPO.

In terms of vendor specific tools and platforms, we suggest the use of one data set and a top-five list of vendor platforms for the challenges. This could be based on raw unbiased market share according to official NYSE/DOW JONES numbers. Currently this list would be: Thermo Fisher Scientific, AB SCIEX, Agilent Technologies, Waters and Bruker.

More and more labs are using mass spectrometers, in particular with the advent of bench-top instruments and thus the need for robust, community accepted and state-of-the-art analytical tools is more pressing than ever. We propose that there is an urgent need for proteomics software development to mature and move into a challenge based, community-wide, open source and open competition powered reality. We argue that this is essential for the future success of the field and that a call is made for an open challenge at global and international community meetings moving forward.

Acknowledgements

R.L. is a Lundbeck Foundation Fellow and is also supported by a Sapere Aude Starting Grant from the Danish Council for Independent Research and a Career Development Award from Human Frontier Science Program. See www.networkbio.org and www.lindinglab.org for more information on cancer related network biology. S.F.M. is funded by SynthSys Edinburgh, which is a Centre for Integrative Systems Biology (CISB) funded by BBSRC and EPSRC. H.F. is funded by the BMBF (German

Federal Ministry of Education and Research). C.J.M. is the recipient of a Wellcome Trust Prize Studentship. T.F.D is funded by the Danish Agency for Science, Technology and Innovation, and the Faculty of Science at Aarhus University.

REFERENCES

[1] Nature Methods Editorial: a home for raw proteomics data. Nat Methods 2012;9:419.
[2] Fenyö D, Beavis RC. Informatics and data management in proteomics. Trends Biotechnol 2002;20:S35–8.
[3] Jones AR, Lister AL. Managing experimental data using FuGE. Methods Mol Biol 2010;604:333–43.
[4] 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.
[5] Styczynski MP, Jensen KL, Rigoutsos I, Stephanopoulos G. BLOSUM62 miscalculations improve search performance. Nat Biotechnol 2008;26:274–5.
[6] 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.
[7] Prill RJ, Saez-Rodriguez J, Alexopoulos LG, Sorger PK, Stolovitzky G. Crowdsourcing network inference: the DREAM predictive signaling network challenge. Sci Signal 2011;4 [mr7].
[8] Searle BC. Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. Proteomics 2010;10:1265–9.
[9] Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, Tasman N, et al. A guided tour of the Trans-Proteomic Pipeline. Proteomics 2010;10:1150–9.
[10] Colaert N, Degroeve S, Helsens K, Martens L. Analysis of the resolution limitations of peptide identification algorithms. J Proteome Res 2011;10:5555–61.
[11] Kil YJ, Becker C, Sandoval W, Goldberg D, Bern M. Preview: a program for surveying shotgun proteomics tandem mass spectrometry data. Anal Chem 2011;83:5259–67.
[12] de Godoy LMF, Olsen JV, Cox J, Nielsen ML, Hubner NC, Fröhlich F, et al. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature 2008;355:1251–4.
[13] Martin SF, Munagapati VS, Salvo-Chirnside E, Kerr LE, Le Bihan T. Proteome turnover in the green alga *Ostreococcus tauri* by time course 15N metabolic labeling mass spectrometry. J Proteome Res 2012;11:476–86.
[14] Arntzen MØ, Koehler CJ, Barsnes H, Berven FS, Treumann A, Thiede B. IsobariQ: software for isobaric quantitative proteomics using IPTL, iTRAQ and TMT. J Proteome Res 2011;10:913–20.
[15] Neilson KA, Ali NA, Muralidharan S, Mirzaei M, Mariani M, Assadourian G, et al. Less label, more free: approaches in label-free quantitative mass spectrometry. Proteomics 2011;11:535–53.
[16] MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics 2010;26:966–8.
[17] Turck CW, Falick AM, Kowalak JA, Lane WS, Lilley KS, Phinney BS, et al. The Association of Biomolecular Resource Facilities Proteomics Research Group 2006 study: relative protein quantitation. Mol Cell Proteomics 2007;6:1291–8.
[18] Schwanhäusser B, Gossen M, Dittmar G, Selbach M. Global analysis of cellular protein translation by pulsed SILAC. Proteomics 2009;9:205–9.
[19] Ahmad Y, Boisvert FM, Lundberg E, Uhlen M, Lamond AI. Systematic analysis of protein pools, isoforms, and modifications affecting turnover and subcellular localization. Mol Cell Proteomics 2012;11 [M111.013680].
[20] Trotter MW, Sadowski PG, Dunkley TP, Groen AJ, Lilley KS. Improved sub-cellular resolution via simultaneous analysis of organelle proteomics data across varied experimental conditions. Proteomics 2010;10:4213–9.
[21] Müller T, Schrötter A, Loosse C, Helling S, Stephan C, Ahrens M, et al. Sense and nonsense of pathway analysis software in proteomics. J Proteome Res 2011;10:5398–408.
[22] Jørgensen C, Linding R. Simplistic pathways or complex networks? Curr Opin Genet Dev 2010;20:15–22.
[23] Bakal C, Linding R, Llense F, Heffern E, Martin-Blanco E, Pawson T, et al. Phosphorylation networks regulating JNK activity in diverse genetic backgrounds. Science 2008;322:453–6.
[24] Kholodenko B, Yaffe MB, Kolch W. Computational approaches for analyzing information flow in biological networks. Sci Signal 2012;5 [re1].
[25] Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, et al. STRING 8—a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 2009;37:D412–6.
[26] Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res 2010;38:W214–20.
[27] Miller ML, Jensen LJ, Diella F, Jørgensen C, Tinti M, Li L, et al. Linear motif atlas for phosphorylation-dependent signaling. Sci Signal 2008;1:ra2.
[28] Linding R, Jensen LJ, Ostheimer GJ, van Vugt MA, Jørgensen C, Miron IM, et al. Systematic discovery of in vivo phosphorylation networks. Cell 2007;129:1415–26.
[29] Camon E, Barrell D, Lee V, Dimmer E, Apweiler R. The Gene Ontology Annotation (GOA) Database—an integrated resource of GO annotations to the UniProt knowledgebase. In Silico Biol 2004;4:5–6.
[30] Lee MV, Topper SE, Hubler SL, Hose J, Wenger CD, Coon JJ, et al. A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. Mol Syst Biol 2011;7:514.
[31] Saez-Rodriguez J, Goldsipe A, Muhlich J, Alexopoulos LG, Millard B, Lauffenburger DA, et al. Flexible informatics for linking experimental data to mathematical models via DataRail. Bioinformatics 2008;24:840–7.
[32] Andersen JS, Lam YW, Leung AK, Ong SE, Lyon CE, Lamond AI, et al. Nucleolar proteome dynamics. Nature 2005;433:77–83.
[33] Janes KA, Albeck JG, Gaudet S, Sorger PK, Lauffenburger DA, Yaffe MB. A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. Science 2005;310:1646–53.
[34] Moult J, Fidelis K, Kryshtafovych A, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)—round IX. Proteins 2011;79:1–5.
[35] Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. Nat Methods 2012;9:796–804.
[36] Meyer P, Alexopoulos LG, Bonk T, Califano A, Cho CR, de La Fuente A, et al. Verification of systems biology research in the age of collaborative competition. Nat Biotechnol 2011;29:811–5.
[37] Yates III JR, Park SKR, Delahunty CM, Xu T, Savas JN, Cociorva D, et al. Toward objective evaluation of proteomic algorithms. Nat Methods 2012;9:455–6.
[38] Friedman DB, Andacht TM, Bunger MK, Chien AS, Hawke DH, Krijgsveld J, et al. The ABRF Proteomics Research Group Studies: educational exercises for qualitative and quantitative proteomic analyses. Proteomics 2011;11:1371–81.