JOURNAL OF PROTEOMICS 90 (2013) 77–84

Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot

Isobar^{PTM}: A software tool for the quantitative analysis of post-translationally modified proteins☆

*Florian P. Breitwieser, Jacques Colinge**

CeMM — Research Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 14, AKH-BT. 25.3, 1090 Vienna, Austria

ARTICLE INFO

Available online 5 March 2013

Keywords:
Bioinformatics
Computational proteomics
Quantitative proteomics
iTRAQ
TMT
Statistics

ABSTRACT

The establishment of extremely powerful proteomics platforms able to map thousands of modification sites, e.g. phosphorylations or acetylations, over entire proteomes calls for equally powerful software tools to effectively extract useful and reliable information from such complex datasets. We present a new quantitative PTM analysis platform aimed at processing iTRAQ or Tandem Mass Tags (TMT) labeled peptides. It covers a broad range of needs associated with proper PTM ratio analysis such as PTM localization validation, robust ratio computation and statistical assessment, and navigable user report generation. Isobar^{PTM} is made available as an R Bioconductor package and it can be run from the command line by non R specialists.

Biological significance
"IsobarPTM is a new software tool facilitating the quantitative analysis of protein modification regulation streamlining important issues related to PTM localization and statistical modeling. Users are provided with a navigable spreadsheet report, which also annotate already public modification sites."
This article is part of a Special Issue entitled: From Genome to Proteome: Open Innovations.

1. Introduction

The dynamic execution of the genetic program encoded in the genome is controlled by a multitude of regulatory mechanisms such as transcription factors, alternative splicing, silencing by non coding RNAs, and epigenetic marks. The large repertoire of gene products generated by the translation/transcription machinery is further submitted to another level of modulation provided by PTMs. These modifications increase the diversity of biomolecules available to cells to adapt to environmental changes or to assemble in specialized tissues.

A large number of PTMs have been described (591 entries in the RESID [1] database vers. 70.01) which modify the properties of proteins for diverse purposes and whose deregulated control can cause multiple disorders. A classical and very important example is the phosphorylation of threonine, tyrosine, or serine that is used to activate proteins upon specific stimuli and to realize signaling cascades [2]. Dysfunctions in such signaling can cause cell proliferation and cancer. More generally, PTMs participate in signal integration within the cell, protein degradation, binding, etc. Commonly studied PTMs are catalyzed by enzymes such as kinases, phosphatases, or acetyltransferases. It has been also shown that distinct PTMs can have a cross-talk, e.g. to establish substitution strategies when one is deficient [3].

Given the importance of PTM regulation in a broad range of biological processes, the analysis of their differences across biological samples is of prime interest in proteomics and is best achieved with quantitative techniques. The measure of PTMs by MS is generally challenging [4,5] since most modifications are lost upon ionization or fractionation resulting in low MS signals

☆ This article is part of a Special Issue entitled: From Genome to Proteome: Open Innovations.

* *Corresponding author*. Tel.: +43 14016070020; fax: +43 140160970000.
E-mail address: jcolinge@cemm.oeaw.ac.at (J. Colinge).

1874-3919/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jprot.2013.02.022

and it might be necessary to operate chromatography and MS equipments in particular conditions. A number of analytical protocols – often relying on chromatographic enrichment for the PTM of interest – have been established successfully, e.g. in the case of phosphorylation [6], ubiquitinylation [7], or acetylation [8].

In this work, we present isobar^{PTM} a new software tool aimed at analyzing the MS/MS spectra of modified peptides resulting from isobarically labeled samples using the Tandem Mass Tags [9] (TMT) or iTRAQ [10] reagents. Isobar^{PTM} is a peptide level extension of the isobar statistical and software framework which we introduced for the analysis of protein ratios [11]. The analysis of modified peptides does not only require determining peptide ratios instead of protein ratios but actually necessitates additional data processing steps. These include the validation of the modification sites on the peptides, the integration of publicly known PTMs, and the relation of modified peptide ratios with the corresponding protein ratios to eliminate apparent PTM regulation caused by the sole protein regulation. As it was the case previously, this new PTM extension is released as free open source software implemented in R and available as part of the isobar Bioconductor package. It provides a complete workflow for handling quantitative PTM data from their validation to user report generation. Currently, Mascot [12], Phenyx [13], Rockerbox [14], comma separated, and PSI mzIdentML identification formats are supported. Isobar is available from the Bioconductor web site (http://www.bioconductor.org).

2. Materials and methods

Programming was done in the R statistical programming language [15] and all the features described in this paper were implemented in the isobar package [11]. The novel PTM functionality is accessible via user report generation options and new specific functions of isobar.

The access to public PTMs from neXtProt [16] is performed via REST-compatible searches (URL http://www.nextprot.org/rest/). The results are retrieved in JSON format and parsed into the ptm.info data frame of the isobar package.

Integration of the PhosphoRS [17] phosphorylation localization tool was realized by using the free stand-alone command line version of PhosphoRS. PhosphoRS does not feature a graphical user interface but requires XML input instead. Isobar^{PTM} integrates generic readers and writers for such a situation and thus provides a seamless interface to PhoshpoRS and other similar external tools.

Validation of statistical models at the peptide level was achieved using data from isobar original publication [11] to assess true and false positive rates of peptide selection as well as the adequacy of the statistical distributions underlying isobar statistics. We further validated the ratio null distribution

2.1. *Application sample data*

We downloaded Phanstiel et al. raw MS data [18] from Tranche. Peak picking and processing was performed using ProteoWizard [19] and the resulting peak lists were searched with Mascot 2.3.0 against the UniProtKB/SwissProt human database [20] appended with sequences of common contaminants (sheep keratin and bovine serum albumin). Fixed modifications were set to cysteine Carbamidomethylation, iTRAQ 4-plex at the peptide N-terminus and lysine side chains. Methionine oxidation was set as variable modification. The phospho dataset was searched with phosphorylation on serine, threonine, and tyrosine residues as variable modifications and mass tolerance was set according to the original publication [18], i.e. precursors 4.5 Da and fragments 0.01 Da. In-house developed scripts were used to filter peptide-spectrum matches to a 1% false discovery rate (FDR) at the protein group and peptide level utilizing reversed database searches. Accordingly, proteins with 2 unique peptides above an ion score threshold of 16, or with a single peptide above a threshold of 40 were selected as unambiguous identifications. Additional peptides for these validated proteins with ion score >12 were also accepted. Only those peptides with a PhopshoRS [17] probability >0.9 were considered for quantitation. The quantification was performed with default isobar settings. From the peak lists, fragments with reporter tag mass +/– 0.005 m/z were extracted and corrected for isotopic impurities. iTRAQ channels were normalized to an equal median intensity. The higher-energy c-trap dissociation (HCD) noise model supplied with the isobar package was used.

3. Results and discussion

In our previous work [11] that established the isobar statistical framework we carefully integrated important elements for selecting significant ratios. Briefly, we eliminated outlier ratios from individual spectra obviously distorted by co-eluting peptides and modeled the technical as well as the biological variability. This allowed for a simple and safe selection of protein ratios that were reliably measured and with sufficient magnitude compared to the sample natural variability. This previous work also included generalized statistical models to take advantage of replicates with a single iTRAQ or TMT experiment, and, in general, put great emphasis on the value of statistically sound methods to obtain robust and competitive methods. Here, we describe isobar^{PTM}, the extension of isobar for the analysis of modified peptide ratios.

Clearly, to bring the whole analysis to the peptide level requires computing peptide ratios instead of protein ratios. That is, all the spectra assigned to a specific peptide/PTM combination (distinct copies of the same peptide can display different patterns of PTMs) are combined in a single weighted ratio calculation taking into account signal intensities and technical variability as previously described for the protein level [11]. Beyond the change in the analysis level, several additional issues that are specifically related to PTM analysis arise and must be properly addressed (Fig. 1). In this section, we present and discuss these various issues followed by two general improvements relevant to PTM quantitation and a comparison with other tools.

3.1. *Validation of PTM site localization*

The localization of PTM sites on modified peptides identified by MS can be ambiguous and, accordingly, only reliably localized PTMs should enter the quantitative analysis. This problem

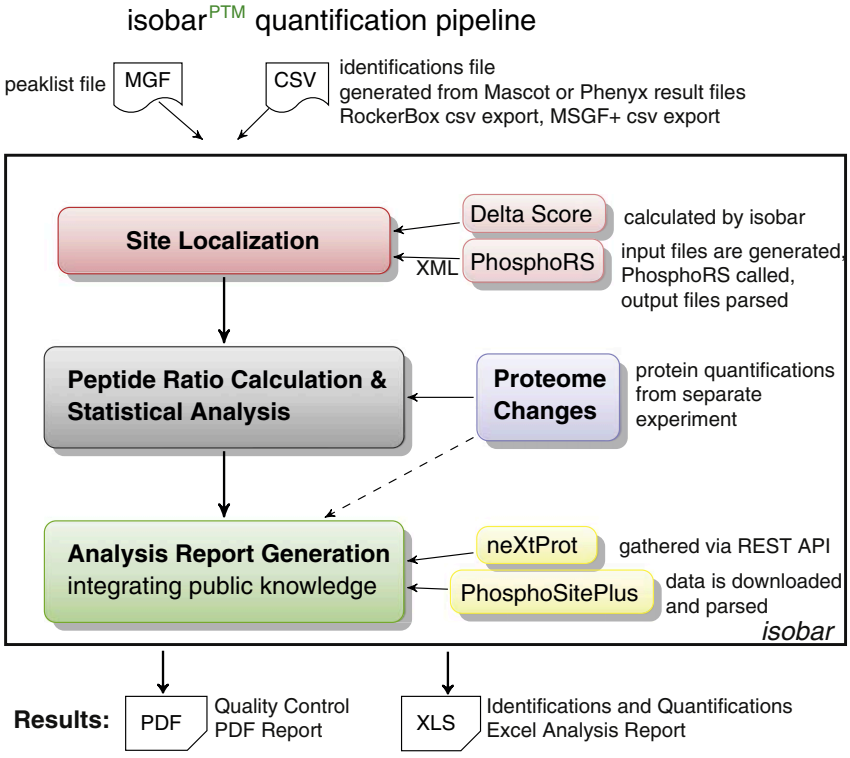


Fig. 1 – Workflow for generating quantitative PTM analysis reports. Peptide–spectrum matches with uncertain localizations of the modifications are removed using a difference or probability score (red box). Reliable matches are used to calculate ratios of modified peptides. Protein ratios from a separate experiment (blue box) can be used to correct modified peptide ratios (solid line) or integrated in the analysis report, and are displayed next to the modified peptide ratios (dotted line). The analysis report (green box) in Excel format integrates previously published knowledge on identified sites, harvesting neXtProt and PhosphoSitePlus. A PDF report containing quality control figures is generated.

mostly occurs when several amino acids of a peptide can carry a certain PTM. For instance the peptide AAGSWHSILSK can be phosphorylated at 3 positions (serines) and if it is singly phosphorylated there are 3 possible localizations. Protein identification search engines provide scores for peptide–spectrum matches that can identify the correct localization provided the peptide fragment coverage is sufficient. In practice, nonetheless, the score alone is not reliable enough [21]. To generally address this issue we integrated a universal method of validating PTM localizations, i.e. the Mascot Delta Score [22]. Although this technique was introduced for phosphorylations and is based on Mascot peptide ion scores, it is in reality of general applicability. It compares the difference between the best- and second best-scoring peptide–spectrum matches for a given peptide and PTM, with distinct modification sites, e.g. AAGS(phos)WHSILSK versus AAGSWHS(phos)ILSK to refer to the above example. The peptide identification score difference informs on the amount of information in the fragmentation spectrum to support one localization versus another one. It provides a measure of confidence in the localization and its analysis was performed by its authors. Since it only relies on score differences it is applicable to any PTM under the condition that the search engine provides multiple peptide/PTM matches for each spectrum and not only the best-scoring one. This is the case of Mascot and many other programs such as Phenyx.

Given the importance of identifying phosphorylated peptides, more advanced procedures of reliable localization have been proposed for this specific case [17,23–27]. To offer the possibility to implement or use external specialized and different PTM localization functions we introduced a generic mechanism of spectrum annotation in isobar^{PTM}, which we exploited to integrate PhosphoRS [17] for phosphorylation localization as an alternative to the Mascot Delta Score approach.

3.2. *Summarizing and quantifying at the level of the modified peptides*

As explained above the computation of modified peptide ratios necessitates introducing another level of organization in the data such that all the spectra – with safe PTM localizations – can be combined for one specific peptide sequence and PTM pattern. We validated that the statistical models introduced for the protein level [11] are still valid at the peptide level by repeating the analysis we conducted for protein ratios [11]. In particular, we assessed that (1) a heavy tailed distribution is appropriate to model peptide ratio null distributions (Supplementary Figs. S1–S3); (2) regulated peptide selection false positive rates are accurately estimated by the statistical models (Supplementary Table S1). We further estimated the true positive rate for different peptide ratios and underlying protein abundance

(Supplementary Table S2). These results, which resemble protein ratio results strongly, are not surprising since isobar protein and peptide ratios are computed identically. As a matter of fact, we do not distinguish between different peptides when we compute protein ratios [11] meaning that a ratio is always a weighted sum in our calculations (sum because we work in the log-scale and weighted by a variance estimate of each spectrum ratio [11]). We concluded this validation by showing that modified peptide ratios also follow a heavy tailed distribution (Supplementary Fig. S4).

The accurate modeling of modified peptide ratios is not necessarily sufficient to obtain biologically relevant results. The observed ratio of a modified peptide is the integrated change of the modification state and the underlying protein abundances and, when quantifying modification state changes, the change in protein abundance – if measured – should not be ignored. Wu et al., comparing the phosphoproteomes of FUS3 or STE7 yeast knockout strains against wild type [28], discussed this problem in great detail and found that 25% of the apparently regulated phosphopeptides disappeared after protein ratio correction. Having access to a high coverage of the proteome in yeast, they were able to calibrate over 96% of the phosphopeptide ratios. In our experience, working with human samples, the overlap between the proteins detected with both unmodified peptides, to estimate protein abundance change, and modified peptides simultaneously resides in the 60–90% range depending on the sample. Note that a PTM enrichment procedure preceding MS, as it is commonly done for phosphopeptide mapping, might require measuring the protein ratios from a separate set of samples. In isobar^{PTM}, we enabled the optional correction of modified peptide ratios when the protein ratio is available, in which case the peptide ratio is divided by the protein ratio. Namely, if Rn is the observed modified peptide ratio and Rp the observed protein ratio, then Rm, the corrected peptide ratio (i.e. its modification state change), is Rm = Rn − Rp (ratios in the log-scale). An adjustment to the estimated variance of Rm is also determined to comply with our general procedure of selecting significantly regulated peptides; the formulas are provided as Supplementary Information.

To exemplify ratio corrections on a human sample, we decided to reanalyze the iTRAQ 4-plex dataset published by Phanstiel et al. [18], who compared embryonic stem cell (ESC) lines with induced pluripotent stem cell (iPSC) lines and a fibroblast cell line. Using the ESC H1 as a reference, in line with the authors, we found that the strongest difference in phosphorylation is observed when comparing with the fibroblast cell line NFF (Fig. 2A), whereas the differences comparing H1 with another ESC line H9 and an iPSC line DF19.7 were very modest (ESCs are similar to iPSC [18]). Turning to the question of correcting phosphorylation site ratios with protein ratios, we found protein ratios for 77% of the phosphopeptides we identified. Applying the same fold-change threshold of 2 as Phanstiel et al., 48% of corrected phosphopeptide ratios were no longer significant after correction with a matching protein ratio, a massive change in the overall sample picture (Fig. 2B & C). Specific examples of four phosphorylated peptides are shown in Supplementary Fig. S5, including cases where the corrected ratio is augmented, reduced, and reversed compared to the original ratio.

Analyzing the enrichment of specific GO terms in differentially expressed and phosphorylated proteins using DAVID (http://david.abcc.ncifcrf.gov), we could recapitulate the findings of Phanstiel et al. Proteins higher in ESCs compared to NFF were enriched in cell cycle-related processes (e. g. chromosomal organization), those higher in NFF were enriched in cytoskeletal processes.

3.3. *Generation of user reports and integration with published PTM data*

The isobar package creates reports for quality control (Fig. 3) and quantification analysis and this feature has been extended to cover modified peptides. Reporting results at the peptide level dramatically augments the size of the data to return to the user and the PDF report we generate for the protein level is no longer appropriate. We hence extended and made fully navigable the already existing spreadsheet user report to also accommodate the peptide level (Fig. 3). It now provides links from quantified peptides to identified spectrum matches, enabling checking of the raw data, etc. Identification information includes search engine scores, modification site localization scores, and extracted isobaric report masses and intensities.

Public databases collect thousands of protein modification sites reported in the literature. To present an overview of existing knowledge about experimentally identified modification sites, we query PTM information-containing databases during user report generation. The neXtProt database [16] is our main source, which we reach via their on-line API (Materials and methods). An alternative source we also support is PhosphoSitePlus [29] that provides a second comprehensive resource of experimentally observed PTMs, primarily phosphorylations although ubiquitinylations and acetylations are covered as well. Isobar integrates PhosphoSitePlus data, automatically downloading the most recent of their monthly updated datasets at the time of report generation, parsing and mapping the data to the experimentally identified proteins. The isobar^{PTM} PTM annotation framework allows users to include supplementary PTM annotation resources if needed.

3.4. *Further improvements*

Having described all the necessary new functionalities implemented to support the analysis of quantitative PTM data, we briefly mention two improvements of isobar that are of general interest and thus impact modified peptide data processing as well.

Firstly, combinations of CID with HCD or electron transfer dissociation (ETD) fragmentation methods are commonly used in iTRAQ or TMT protocols to achieve more identifications on the basis of a fast method (CID), while more accurate quantification is obtained on the basis of the slower but more precise method (HCD or ETD) limited to a narrow mass range covering the iTRAQ or TMT channels [30]. In such a case, isobar can merge identification runs (e.g. from CID) and quantification runs (e.g. from HCD spectra) while reading the MS data, and even combine identifications obtained from quantification runs when they include regular fragment information as well. For instance, CID and HCD can provide complementary peptide identifications [31], which in our laboratory equipped with an

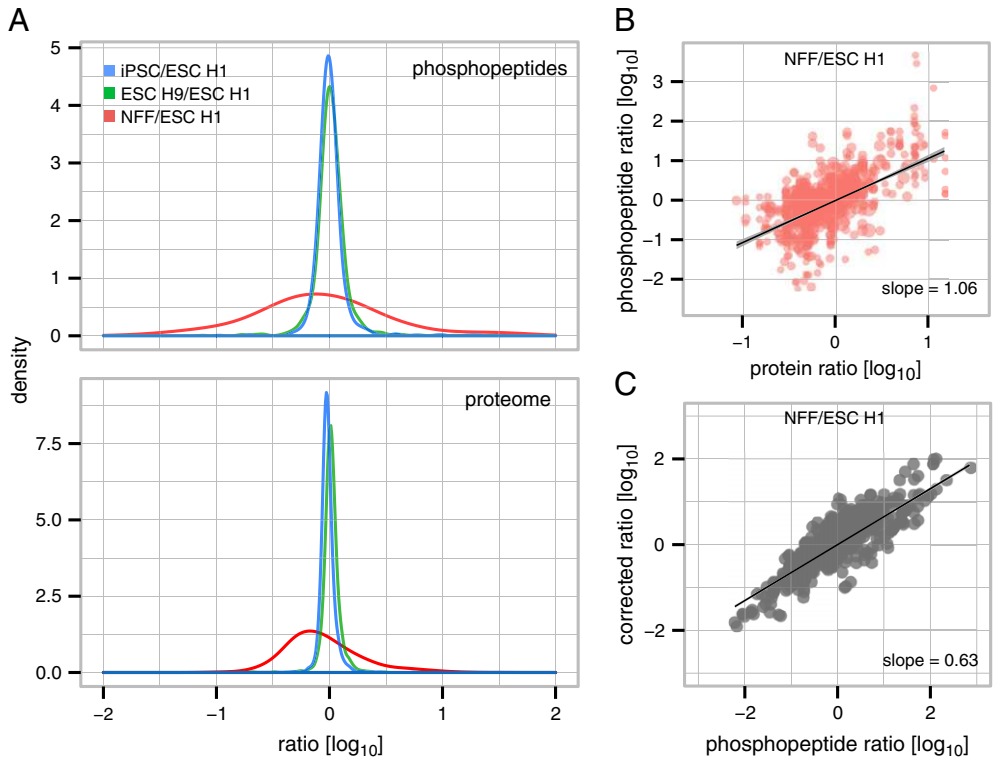


Fig. 2 – Analysis of Phanstiel et al. data. Ratios are relative to the 114 channel corresponding to H1 embryonic stem cells. (A) We observe the larger spread of ratios both in the phosphoproteome (top) and the proteome (bottom) when comparing to NFF fibroblast cells (red, channel 115) compared to H9 embryonic stem cells (green, channel 116) and DF19.7 induced pluripotent stem cells (blue, channel 117). (B) Protein ratios versus phosphopeptide ratios. We note the positive correlation indicating that a significant part of the phosphopeptide ratios originate from the protein regulation and not the phosphorylation site regulation. (C) Original versus corrected phosphopeptide ratios. The slope 0.63 < 1 confirms the general reduction of the ratios after correction.

LTQ-Orbitrap Velos (ThermoFisher Scientific, Waltham, MA) each account for 20–30% of the peptide–spectrum matches in the analysis of phosphopeptide enriched fractions.

Secondly, we could find a more accurate model of heavy tailed distribution than the Cauchy. We have observed that generalized Student's t distribution better models the tails and thus improve the sensitivity of isobar (Supplementary Figs. S1–S4, S6). This distribution belongs to the generalized logistic distribution family that is a very general model of heavy tailed distribution parameterized by five parameters, which is too much for practical applications where data can be sparse. The generalized Student's t distribution has three parameters as compared to Cauchy which has only two, and it is a widely used model for heavy tailed distributions. Cauchy remains isobar default to ensure maximum robustness with smaller datasets (less than 1000 ratios, Supplementary Table 3).

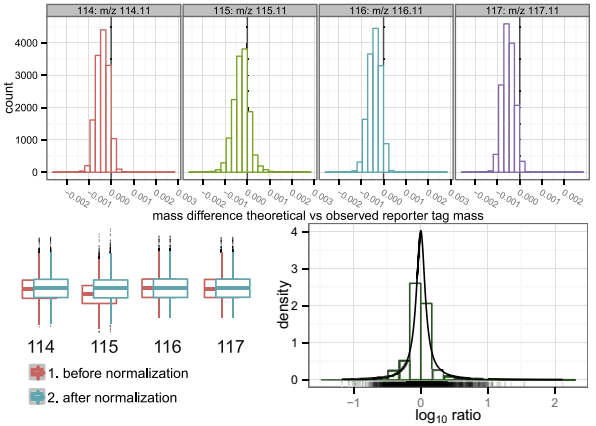
3.5. *Use without programming*

The presented tool can be used with minimal configuration and no direct interaction with R: a plain text property file specifies basic parameters such as the isobaric tagging kit used (iTRAQ/TMT, 2-, 4-, 6-, or 8-plex), peak list and identification file names, and how the report and quantification should be produced (see Fig. 3). An R script, which can be called from the command line, runs the analysis with the provided parameters and generates the results. Many further options can be specified to customize the analysis and report — examples are provided with the package to guide beginners.

3.6. *Comparison with existing tools*

In Table 1 we present a feature comparison of software used in recent publications for the quantitation of isobarically tagged PTM experiments. The Coon group has developed the COMPASS [32] proteomics analysis suite for OMSSA, used recently for the quantitation of stem cell proteomes and phosphoproteomes [18]. The Marto group introduced Multiplierz [33] that provides an excellent basis for extensible workflows and data access and has been used for example for the quantitation of the mTOR regulated phosphoproteome [34]. Thermo Scientific's commercial Proteome Discoverer enables to construct a workflow from identification to quantitation. As it can be appreciated from the table, isobar's distinguishing features are its statistical fundament for quantitation and significance analysis, the high level integration of public PTM data for report generation, and the configurability and extensibility with bioinformatics packages for R/Bioconductor.

A) Quality Control Report



B) Report Properties

general	type	"iTRAQ4plexSpectra", …	isobaric tagging kit
	isotope.impurities	matrix	isotope impurity matrix
	peaklist	file name(s)	MGF files
	identifications	file name(s)	identification CSV files
ptm	modif	"PHOS", "ACET", …	modification to track in the report
	ptm.info.f	function name	*e.g.* "getPtmInfoFromNextProt"
	correct.ratios.with	data.frame	proteome ratio table
quantification	noise.model	NoiseModel object	technical variability model
	class.labels	vector	classes of tag channels
	normalize	boolean	normalize intensities?
	summarize	boolean	summarize ratios of the same class?
	ratios.opts	list, see '?peptideRatios'	additional properties for quantitation
report	write.qc.report	boolean	generate quality control report?
	write.xls.report	boolean	generate spreadsheet analysis report?
	xls.report.format	"long" or "wide"	layout of spreadsheet
	xls.report.columns	"p.value.ratio", …	columns visible in spreadsheet

C) Analysis Report

	Sequence	Phosphorylation Position	ACs	ID	Description	Gene	Spectra	Channels	Ratio	Significa	Ratio Minus Sd	Ratio Plus Sd	P Value Rat	P Value Sample
1	SSPNPFVGS(p)PPK	S401*;S380*	P98082-[1,3]	DAB2_HUMAN	Disabled homolog 2	DAB2	1	116 / 114	6.48	1	4.13	10.17	0.0000	0.0492
5	KAEPS(p)EVDMNSPK	S65	Q9NR30-1	DDX21_HUMAN	Nucleolar RNA helicase 2	DDX21	5	115 / 114	0.06	1	0.03	0.15	0.0004	0.0338
6	NEEPS(p)EEELDAPKPK	S121*	Q9NR30-1	DDX21_HUMAN	Nucleolar RNA helicase 2	DDX21	6	115 / 114	0.09	1	0.03	0.29	0.0224	0.0377
2	EES(p)EEEEDEDDEEEEEEEK	S32*	P35659	DEK_HUMAN	Protein DEK	DEK	2	115 / 114	0.02	1	0.01	0.06	0.0003	0.0231
2	S(p)LLVEGK	S51*	P35659	DEK_HUMAN	Protein DEK	DEK	2	115 / 114	0.11	1	0.07	0.17	0.0000	0.0419
2	KPATPAEDDEDDDLDLFGS(p)D	S162*;S528*	P		on factor 1-delta	EEF1D	2	115 / 114	7.04	1	5.43	9.13	0.0000	0.0471
5	AS(p)STSTPEPTR	S485	Q		enabled homolog	ENAH	5	115 / 114	0.01	1	0.00	0.11	0.0348	0.0186
3	LWT(p)PLK	T143	Q9NYF3	FA53C_HUMAN	Protein FAM53C	FAM53C	3	115 / 114	0.06	1	0.05	0.08	0.0000	0.0331
3	ATEDGEEDEVS(p)AGEK	S1435*	Q9BXW9-2	FACD2_HUMAN	Fanconi anemia group D2 pro	FANCD2	3	115 / 114	0.07	1	0.03	0.17	0.0015	0.0349
28	RAPS(p)VANVGSHC(c)DLSLK	S2152*;S2144*	P21333-[1,2]	FLNA_HUMAN	Filamin-A	FLNA	28	115 / 114	7.01	1	3.50	14.05	0.0025	0.0472
4	AGGSAALSPS(p)K	S33*	Q92522	H1X_HUMAN	Histone H1x	H1FX	4	115 / 114	0.07	1	0.05	0.10	0.0000	0.0346
2	S(p)APAPK	S7*	O60814,P06899,	H2B1B_HUMAN,	Histone H2B type 1-B, Histon	H2BFS, HIS	2	115 / 114	0.10	1	0.06	0.14	0.0000	0.0394
2	LEDVGS(p)DEEDDS(p)GKDK	S255*&S261*	P08238	HS90B_HUMAN	Heat shock protein HSP 90-be	HSP90AB1	2	115 / 114	0.12	1	0.08	0.19	0.0000	0.0441
1	C(c)TPAC(c)LS(p)FGPK	S40	P34932	HSP74_HUMAN	Heat shock 70 kDa protein 4	HSPA4	1	115 / 114	42.38	1	28.64	62.72	0.0000	0.0247

P29692-1 pos 162: Phosphoserine; by CK2
P29692-2 pos 528: Phosphoserine; by CK2

Quantifications | Identifications | Analysis Properties | Log

links to spectrum level information

Fig. 3 – Isobar^{PTM} quantification reports. (A) Quality control report showing reporter tag mass precision, reporter tag intensities before and after normalization, and a histogram of peptide ratios along with the fit Cauchy biological variability ratio distribution [11]. (B) Report generation is controlled by a properties file. Columns: property name, possible values, and explanation. (C) Spreadsheet user report. It includes modified peptide sequence with the positions of the modifications in the protein sequence (separated by semicolons if in multiple identical peptides or by ampersands if multiple occurrences in the same peptide). A star identifies positions previously reported in the literature, tooltips display information on the latter PTMs (here from neXtProt). The report has multiple tabs for identifications and contains multiple links to navigate them, e.g. from a modified peptide as featured in the figure to all the spectra supporting its identification.

Table 1 – Comparison with similar software packages.

	Isobar^{PTM}	Proteome Discoverer	COMPASS	multiplierz
Availability	Open source	Commercial	Open source	Open source
iTRAQ and TMT Quant	Yes	Yes	Yes	Yes
Statistical Framework	Yes, technical and biological variability	no	no	Technical variability modeled [a]
PTM Localization	Yes [b]	Yes [c]	No	Yes [a]
Annotation of PTM sites	Yes [d]	No	No	No
Correction with Protein Ratios	Yes	Yes	Yes	Yes
Restrictions	No graphical user interface	Closed source	For usage with OMSSA only	Scripting skills required

[a] Scripts for robust error model and Mascot Delta Score available on the multiplierz homepage http://blais.dfci.harvard.edu/index.php?id=106.
[b] PhosphoRS and Mascot Delta Score.
[c] PhosphoRS.
[d] NextProt and PhosphoSitePlus.

4. Conclusion

To measure and understand PTMs in disease and biological processes is an important objective of current research in proteomics. Such experiments remain challenging but the technology has made such tremendous progresses that in-depth and proteome-scale mappings of specific PTMs can be realized with unprecedented accuracy. As a consequence, data analysis faces difficulties that are common to most omics fields: the access to reliable and highly automated methods of processing and selecting relevant data conditions the extent to which discoveries can be accomplished. With this consideration in mind, we started to develop a combined statistical and software framework – isobar – that we originally targeted towards protein expression studies [11]. The work presented here implements a second step aimed at including the peptide PTM regulation level within the scope of the analyses supported by this platform. We named this specific branch of the project isobar^{PTM}.

The approach we have followed remains in line with the original concepts that guided isobar design: the establishment of robust and accurate statistical models provides the most appropriate basal layer to construct a successful software platform. In isobar^{PTM} we greatly benefited from the initial effort to the point where no real additional statistical modeling was necessary, just validations and small adaptations. The models developed for the proteins turned out to be adequate for the peptides as well and we could concentrate on establishing the new software functionalities. Doing so, we also benefited from the general improvements and bug-fixes we kept introducing in the isobar libraries that has been applied to a multitude of projects by ourselves [35] and others [36] meanwhile.

Practically, successful and high quality analysis of PTM data on a large-scale preventing the manual inspection of each and every interesting spectrum implies the execution of several tasks that are generally not all accessible to the average proteomics laboratory in the best conditions. With isobar^{PTM} we have streamlined the fundamental steps of extracting and combining identification and MS data, including when hybrid fragmentation strategies e.g. CID-HCD are adopted, performing an automatic validation of the localization of the modification sites and removing dubious cases, and applying state of the art statistical modeling to compute ratios and assess their significance (Fig. 1). Furthermore, convenient user reports are produced which include a navigable sophisticated spreadsheet that represents a convenient paradigm for reporting large sets of results as generated by peptide level studies.

Finally, we believe that bioinformatics tools should be as interoperable as possible and the development of open source R Bioconductor packages represents an effective way of implementing this goal. In particular, follow up functional analyses such as GO term or pathway enrichments are made straightforward thanks to many existing Bioconductor packages. Developing within the R platform allows other bioinformaticians to use isobar at all possible levels, from calling high-level functions down to completely redesigned analyses capitalizing on the low-level functions. For non-bioinformaticians and for usage within an automated pipeline, we make the complete analysis with report generation accessible on the command line requiring simple configuration via text files only. In the future of the isobar project, we will give significant attention to the development of a graphical user interface.

Isobar and isobar^{PTM} can be downloaded from http://www.ms-isobar.org or from the Bioconductor web site.

Acknowledgments

We thank all our CeMM colleagues and in particular André Müller, Uwe Rix, Alexey Stukalov, and Keiryn Bennett for useful feedback and advices. JC is supported by an Austrian Science Fund (FWF) grant No P 24321-B21.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jprot.2013.02.022.

REFERENCES

[1] Garavelli JS. The RESID database of protein modifications as a resource and annotation tool. Proteomics 2004;4:1527–33.
[2] Bodenmiller B, Wanka S, Kraft C, Urban J, Campbell D, Pedrioli PG, et al. Phosphoproteomic analysis reveals interconnected system-wide responses to perturbations of kinases and phosphatases in yeast. Sci Signal 2010;3:rs4.
[3] van Noort V, Seebacher J, Bader S, Mohammed S, Vonkova I, Betts MJ, et al. Cross-talk between phosphorylation and lysine acetylation in a genome-reduced bacterium. Mol Syst Biol 2012;8:571.
[4] Mallick P, Kuster B. Proteomics: a pragmatic perspective. Nat Biotechnol 2010;28:695–709.
[5] Mann M, Jensen ON. Proteomic analysis of post-translational modifications. Nat Biotechnol 2003;21:255–61.
[6] Bodenmiller B, Aebersold R. Quantitative analysis of protein phosphorylation on a system-wide scale by mass spectrometry-based proteomics. Methods Enzymol 2010;470: 317–34.
[7] Kim W, Bennett EJ, Huttlin EL, Guo A, Li J, Possemato A, et al. Systematic and quantitative assessment of the ubiquitin-modified proteome. Mol Cell 2011;44:325–40.
[8] Henriksen P, Wagner SA, Weinert BT, Sharma S, Bacinskaja G, Rehman M, et al. Proteome-wide analysis of lysine acetylation suggests its broad regulatory scope in *Saccharomyces cerevisiae*. Mol Cell Proteomics 2012;11:1510–22.
[9] Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, et al. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem 2003;75:1895–904.
[10] Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, et al. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. Mol Cell Proteomics 2004;3:1154–69.
[11] Breitwieser FP, Muller A, Dayon L, Kocher T, Hainard A, Pichler P, et al. General statistical modeling of data from protein relative expression isobaric tags. J Proteome Res 2011;10:2758–66.
[12] 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.
[13] Colinge J, Masselot A, Giron M, Dessingy T, Magnin J. OLAV: towards high-throughput tandem mass spectrometry data identification. Proteomics 2003;3:1454–63.
[14] van den Toorn HW, Munoz J, Mohammed S, Raijmakers R, Heck AJ, van Breukelen B. RockerBox: analysis and filtering of massive proteomics search results. J Proteome Res 2011;10: 1420–4.
[15] R_Core_Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012.
[16] Lane L, Argoud-Puy G, Britan A, Cusin I, Duek PD, Evalet O, et al. neXtProt: a knowledge platform for human proteins. Nucleic Acids Res 2012;40:D76–83.
[17] Taus T, Kocher T, Pichler P, Paschke C, Schmidt A, Henrich C, et al. Universal and confident phosphorylation site localization using phosphoRS. J Proteome Res 2011;10:5354–62.
[18] Phanstiel DH, Brumbaugh J, Wenger CD, Tian S, Probasco MD, Bailey DJ, et al. Proteomic and phosphoproteomic comparison of human ES and iPS cells. Nat Methods 2011;8:821–7.
[19] Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics 2008;24:2534–6.
[20] Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 2006;34:D187–91.
[21] Chalkley RJ, Clauser KR. Modification site localization scoring: strategies and performance. Mol Cell Proteomics 2012;11: 3–14.
[22] Savitski MM, Lemeer S, Boesche M, Lang M, Mathieson T, Bantscheff M, et al. Confident phosphorylation site localization using the Mascot Delta Score.Mol Cell Proteomics 2011;10 [M110 003830].
[23] Bailey CM, Sweet SM, Cunningham DL, Zeller M, Heath JK, Cooper HJ. SLoMo: automated site localization of modifications from ETD/ECD mass spectra. J Proteome Res 2009;8:1965–71.
[24] Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol 2006;24:1285–92.
[25] Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, et al. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell 2006;127:635–48.
[26] Ruttenberg BE, Pisitkun T, Knepper MA, Hoffert JD. PhosphoScore: an open-source phosphorylation site assignment tool for MSn data. J Proteome Res 2008;7:3054–9.
[27] Swaney DL, Wenger CD, Thomson JA, Coon JJ. Human embryonic stem cell phosphoproteome revealed by electron transfer dissociation tandem mass spectrometry. Proc Natl Acad Sci U S A 2009;106:995–1000.
[28] Wu R, Dephoure N, Haas W, Huttlin EL, Zhai B, Sowa ME, et al. Correct interpretation of comprehensive phosphorylation dynamics requires normalization by protein expression changes. Mol Cell Proteomics 2011;10(8):M111 009654.
[29] Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, et al. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res 2012;40:D261–70.
[30] Kocher T, Pichler P, Schutzbier M, Stingl C, Kaul A, Teucher N, et al. High precision quantitative proteomics using iTRAQ on an LTQ Orbitrap: a new mass spectrometric method combining the benefits of all. J Proteome Res 2009;8:4743–52.
[31] Frese CK, Altelaar AF, Hennrich ML, Nolting D, Zeller M, Griep-Raming J, et al. Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-Orbitrap Velos. J Proteome Res 2011;10:2377–88.
[32] Wenger CD, Phanstiel DH, Lee MV, Bailey DJ, Coon JJ. COMPASS: a suite of pre- and post-search proteomics software tools for OMSSA. Proteomics 2011;11:1064–74.
[33] Parikh JR, Askenazi M, Ficarro SB, Cashorali T, Webber JT, Blank NC, et al. multiplierz: an extensible API based desktop environment for proteomics data analysis. BMC Bioinformatics 2009;10:364.
[34] Hsu PP, Kang SA, Rameseder J, Zhang Y, Ottina KA, Lim D, et al. The mTOR-regulated phosphoproteome reveals a mechanism of mTORC1-mediated inhibition of growth factor signaling. Science 2011;332:1317–22.
[35] Winter GE, Rix U, Carlson SM, Gleixner KV, Grebien F, Gridling M, et al. Systems-pharmacology dissection of a drug synergy in imatinib-resistant CML. Nat Chem Biol 2012;8:905–12.
[36] Gluck F, Hoogland C, Antinori P, Robin X, Nikitin F, Zufferey A, et al. EasyProt — an easy-to-use graphical platform for proteomics data analysis. J Proteomics 2012;79C:146–60.