Computational protein profile similarity screening for quantitative mass spectrometry experiments

Marc Kirchner1,2,†, Bernhard Y. Renard1,3,†, Ullrich Körte3, Darryl J. Pappin4, Fred A. Hamprecht1,3, Hanno Steen1,2,*, and Judith A. J. Steen3,‡

1Department of Pathology, Proteomics Center, Children’s Hospital Boston, 2Department of Pathology, Harvard Medical School, Boston, MA, USA, 3Interdisciplinary Center for Scientific Computing, University of Heidelberg, Heidelberg, Germany, 4Cold Spring Harbor Laboratory, Cold Spring Harbor, NY and 5Department of Neurobiology, Harvard Medical School and T. M. Kirby Neurobiology Center, Children’s Hospital, Boston, MA, USA

Received on May 22, 2009; revised on September 14, 2009; accepted on October 7, 2009

Advance Access publication October 27, 2009

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: The qualitative and quantitative characterization of protein abundance profiles over a series of time points or a set of environmental conditions is becoming increasingly important. Using isobaric mass tagging experiments, mass spectrometry-based quantitative proteomics deliver accurate peptide abundance profiles for relative quantitation. Associated data analysis workflows need to provide tailored statistical treatment that (i) takes the correlation structure of the normalized peptide abundance profiles into account and (ii) allows inference of protein-level similarity. We introduce a suitable distance measure for relative abundance profiles, derive a statistical test for equality and propose a protein-level representation of peptide-level measurements. This yields a workflow that delivers a similarity ranking of protein abundance profiles with respect to a defined reference. All procedures have in common that they operate based on the true correlation structure that underlies the measurements. This optimizes power and delivers more intuitive and efficient results than existing methods that do not take these circumstances into account.

Results: We use protein profile similarity screening to identify candidate proteins whose abundances are post-transcriptionally controlled by the Anaphase Promoting Complex/Cyclosome (APC/C), a specific E3 ubiquitin ligase that is a master regulator of the cell cycle. Results are compared with an established protein correlation profiling method. The proposed procedure yields a 50.9-fold enrichment of co-regulated protein candidates and a 2.5-fold improvement over the previous method.

Availability: A MATLAB toolbox is available from http://hci.iwr.uni-heidelberg.de/mip/proteomics.

Contact: hanno.steen@childrens.harvard.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

To whom correspondence should be addressed:
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.
‡The authors wish it to be known that, in their opinion, the last two authors should be regarded as joint Last authors.

1 INTRODUCTION

Current global quantitative proteomics experiments provide time-resolved insight into the dynamic behavior of cellular processes at the protein level and are more reflective of the immediate status of a cell compared with, e.g. transcriptional studies which completely ignore post-transcriptional regulation. In this context, the quantitative and qualitative characterization of protein expression-level profiles over a series of time points or a set of environmental conditions is becoming increasingly important. Quantitative mass spectrometry (MS) is the method of choice to directly identify, quantitate and characterize hundreds or thousands of proteins simultaneously, delivering accurate peptide abundance profiles that yield relative quantitative information (Bantscheff et al., 2007; Ong and Mann, 2005).

However, given the large numbers of proteins in these studies, the biochemical validation of the information gathered in such experiments is not feasible. It is hence desirable to develop computational screening procedures that can rank proteins based on their similarity to the abundance profile of a reference protein over a time course or a set of conditions. Although observing similar protein abundance profiles cannot prove specific biochemical properties, the associated ranking can yield a valuable enrichment of protein groups associated with the same or similar cellular processes and provide a criterion for the prioritization of biological validation experiments, i.e. a testable shortlist of candidate proteins (Andersen et al., 2003; Foster et al., 2006).

Quantitative MS methods provide direct information about abundance levels of endogenous proteins (Bantscheff et al., 2007; Ong and Mann, 2005). Quantitative MS is thus a method of choice for the comprehensive differential analysis of protein abundance profiles, which vary with time and/or experimental conditions (Bürckstümmer et al., 2006; Fields and Song, 1989; Paig et al., 2001; Rigaut et al., 1999; Ross et al., 2004; Selbach and Mann, 2006; Tedford et al., 2008; Thompson et al., 2003; White, 2008). Multiplexed isobaric mass tagging (IMT) approaches or multiplexed metabolic labeling allow for time-resolved protein abundance measurements for thousands of proteins simultaneously, overcoming the need for tedious individual protein testing. Recent computational analyses (Hill et al., 2008; Oberg et al., 2008) have...
provided means of statistical evaluation of differential abundances between IMT labels but have not focused on the statistical concepts necessary to compare peptide and protein profiles. Protein correlation profiling (PCP) is a heuristic protein profile screening approach that has been developed in the context of tracking interacting proteins over fractions of a sucrose gradient. It has successfully been used for large-scale proteomic characterization of the various human organelles (Andersen et al., 2003; Foster et al., 2006). Here, we use PCP as a de facto standard for performance comparison.

Our study introduces a protein profile similarity screening (PSS) procedure that utilizes abundance profiles from quantitative proteomics experiments using the IMT strategy. We investigate the statistical consequences of data normalization, which, if not accounted for, can jeopardize standard testing procedures. We establish the connection between IMT series and the statistically evaluated to yield a shortlist of coregulation candidates.

Section 2 of the article provides all methodological details and the proposed screening procedure is applied to real-world experimental data in Section 3. In Sections 4 and 5, we report and discuss results, suggesting that the proposed approach is indeed powerful: with only few protein IMT abundance measurements, the identification of a set of well-known co-regulated proteins is possible. Conclusions and perspectives are offered in Section 6.

2 METHODS

2.1 Workflow overview

We propose a novel procedure for the inference of protein abundance profile similarity from IMT analyses of proteomic time series experiments. Given a set of normalized IMT peptide reporter ion profiles (Fig. 1A), we apply a hierarchical clustering (Fig. 1B) method tailored to the statistical dependence structure that results from the normalization. The Dirichlet likelihood ratio test (DLRT) delivers a suitable cluster tree cutoff strategy and yields a data grouping on the peptide level. From there we construct protein signatures, representing the protein-wise peptide distribution over the clusters (Fig. 1C). The Mallows distance then provides a suitable measure for the inference of protein similarity (Fig. 1D). In the final step, proteins are ranked according to their profile similarity to one or more predefined marker proteins (Fig. 1E).

2.2 Statistical properties of IMT time-series measurements

2.2.1 Isobaric mass tagging. IMT labels such as TMT and iTRAQ generally consist of three parts: a reactive group which binds to the peptide, a reporter group and a balancer group. Varying combinations of light and heavy isotopes in the reporter and balancer groups yield four unique reporter ion masses while keeping the overall mass constant (Ross et al., 2004; Thompson et al., 2003). For quantitation experiments, K labels are attached to N peptide species from K experimental conditions. In LCMS analysis, the differentially tagged species have the same retention time and consequently form a single peptide isotope distribution in the MS parent spectrum. During fragmentation, the reporter/balancer/peptide compound breaks in three and yields K absolute reporter ion abundance measurements \( x = (x_1, x_2, \ldots, x_K)^T \), for each of the N peptide species. Given a protein, the vector \( x \) holds the respective reporter ion profile of observed abundances.
Hierarchical clustering iteratively interprets a normalized peptide reporter ion profile. It is no statistical evidence that the observations in the two branches stem to identify clusters within the tree, it is necessary to determine in which of the merges (groups of) observations and eventually yields a merge tree. In order to perform a meaningful analysis, it is required to identify a suitable clustering criterion. The method requires a suitable dissimilarity measure between the observed data points. In our case, as a direct consequence of sum normalization, the coefficients of any peptide reporter ion profile $x^*$ add to 1, i.e. $\sum_i x_i^* = 1$. This defines a hyperplane in $K$ dimensions and every vector $x^*$ lies on a $K$-dimensional simplex. Standard distance measures like the Euclidean distance cannot account for such dependency structures and thus we resort to the natural measure of distance on the simplex (Atchison, 1983) given by

$$\Delta(x^*, y^*) = \sum_i \left( \frac{x_i^*}{\sqrt{\sum_i x_i^*}} - \frac{y_i^*}{\sqrt{\sum_i y_i^*}} \right)^2$$

where $x^*$ and $y^*$ are $K \times 1$ vectors of sum normalized reporter ion profiles and $g(x^*) = (\prod_i x_i^*)^{1/K}$ denotes the geometric mean of $x^*$. For the calculation of agglomerative distances during the clustering procedure, we use average linkage (Cortis et al., 2007).

2.3.2 Dirichlet likelihood ratio test Hierarchical clustering iteratively merges (groups of) observations and eventually yields a merge tree. In order to identify clusters within the tree, it is necessary to determine in which of the tree nodes the merge operations are supported by the data and in which they are not. We approach this problem with a statistical hypothesis test for differences between groups of observations: the merge is accepted if there is no statistical evidence that the observations in the two branches stem from different distributions. Since the normalized underlying data violate the independence assumptions necessary for standard statistical tests, we interpret a normalized peptide reporter ion profile $x^*$ as a realization drawn from a Dirichlet distribution

$$p(x^*|\alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i (x_i^*)^{\alpha_i - 1}$$

where $\Gamma$ is the Gamma function, $x_i^* > 0$, $\sum_i x_i^* = 1$ and Dirichlet parameters given by $\alpha = (\alpha_1, \ldots, \alpha_K)$. When the number of parameters $K$ is large, these parameters can be estimated from the data using the method of moments. Alternatively, to obtain a fixed number of degrees of freedom, we can use a prior with a fixed number of Dirichlet parameters. We test the null hypothesis $H_0$ against the alternative hypothesis $H_1$. The likelihood ratio test statistic is given by

$$\Lambda(x^*, \alpha^*; x^*, \alpha^*) = \frac{\prod_i (x_i^*)^{\alpha_i - 1}}{\prod_i (x_i^*)^{\alpha_i - 1}}$$

where $\alpha^*$ is a set of parameters for which the alternative hypothesis is true. The rationale behind this approach is that IMT peptide reporter ion profiles are susceptible to post-translational modification effects: in the presence of PTMs, peptides of a protein may exhibit very diverse reporter ion profiles. Different types of reporter ion profiles aggregate in different clusters and determining the distribution of peptides over these clusters yields a robust and versatile protein representation. Subsequent comparison of protein signatures then allows for the calculation of protein-level abundance profile similarity.
We evaluated our method on an iTRAQ (a specific IMT strategy) (Levina and Bickel, 2001; Rubner et al., 1998; we define a discrete joint distribution $F(x_i, x_j) = \{ f_i(x_i, x_j) \}$ of flows between the signature entries $x_i$ and $x_j$ of proteins $k$ and $l$. We then identify the distribution $F^*$ that minimizes the expected cost $d_{ij}$:

$$F^*(x_i, x_j) = \arg\min_{F} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{ij} f_i(x_i, x_j).$$

(6)

Admissible solutions $F^*$ must fulfill the properties of a distribution function, i.e.

$$f^*_i(x_i, x_j) \geq 0, \quad \text{and} \quad \sum_{j=1}^{N} f^*_i(x_i, x_j) = 1,$$

and their marginals must correspond to the signature vectors,

$$\sum_{i=1}^{N} f^*_i(x_i, x_j) = s_i \quad \text{and} \quad \sum_{j=1}^{N} f^*_i(x_i, x_j) = s_j.$$

The costs of changes $d_{ij}$ are defined as the average squared distance between the peptide clusters $s_i$ and $s_j$ as given by

$$d_{ij} = \frac{1}{N} \sum_{k=1}^{N} \sum_{l=1}^{N} (x^{m_k} - y^{n_l})^2,$$

(9)

where $x^{m_k}$ with $k \in \{1, \ldots, N_i\}$ represents all normalized reporter ion profiles of peptides in the $i$th cluster and $y^{n_l}$ with $l \in \{1, \ldots, N_j\}$ represents all normalized reporter ion profiles of peptides in the $j$th cluster. This definition of $d_{ij}$ is consistent with the average linkage clustering scheme. The Mallows distance between two protein signatures $s_i$ and $s_j$ is then given by

$$m_{ij} = m(s_i, s_j) = \sqrt{\sum_i \sum_j d_{ij}^2}.$$

(10)

For the complete set of protein signatures, this yields a $P \times P$ protein distance matrix $M = [m_{ij}]$.

### 2.5 Identifying similar proteins

It is now possible to derive a shortlist of proteins that exhibit similar abundance profiles from the distance matrix $M$. Given a known substrate protein $p$, the elements of the column vector $m_p = (m_{p1}, m_{p2}, \ldots, m_{pP})$ are constrained to the interval $[0,1]$ and approximately follow a beta distribution. The parameters $m_{p1}$ and $m_{p2}$ are estimated by maximum likelihood and subsequently allow the computation of a cutoff quantile $q$ (generally the 0.01 or 0.05 quantile). All proteins $r$ with a Mallows distance $m_r$ below the quantile $q$ are then included in the protein shortlist.

### 3 EXPERIMENTS

We evaluated our method on an iTRAQ (a specific IMT strategy) MS experiment of the APC/C. The APC/C is a highly specific ubiquitin ligase that marks its substrates for degradation by the 26S proteasome and thus controls entry into and exit from mitosis in the cell cycle.

The analysis attempts to elucidate APC/C substrate candidates from a full cell extract, based on the temporal protein abundance profile of the known APC/C substrate Cyclin-B1 (CCNB1) (King et al., 1995).

We compared the proposed workflow against PCP (Andersen et al., 2003), which calculates peptide-level $x^2$ distances based on a predefined set of marker proteins and takes peptide medians to infer protein-level dissimilarity. PCP has been used in a large-scale proteomic organelle mapping study (Foster et al., 2006).

3.1 Experimental background

The data stem from lysates of HeLa S3 cells arrested in four time points in the cell cycle: prometaphase, M/G1, G1 and G1/S (Fig. 2). Over the selected time course cells divide and the observed changes in protein abundance also reflect changes induced by APC/C activity, i.e. controlled protein degradation. The samples were digested with trypsin, iTRAQ-labeled, combined, fractionated first by SCX then by reversed phase liquid chromatography and analyzed by MALDI-TOF/TOF MS (Applied Biosystems/MDS Sciex 4800 TOF/TOF). The iTRAQ reagents (Ross et al., 2004) consist of three parts: a reporter group with mass 114–117, a balance group with mass 28–31 and the amine-specific peptide reactive group (N-hydroxy succinimide, NHS), targeting the peptide N-terminal and the ε-amino group of lysine. The overall mass of the reporter-balance combinations is kept constant (145 Da) using differential isotopic labeling of $^{13}$C, $^{15}$N and $^{18}$O. Peptide and protein identifications were performed using the Mascot search engine (Matrix Science, version 2.2.1) (Perkins et al., 1999) with a fully tryptic human database (IPI human, version 3.23) and a false positive rate of 4.1% at the peptide level. The iTRAQ reporter group abundances were extracted from the raw MALDI-TOF/TOF data, isotope-correlated and matched to identified peptides using DataExplorer (Applied Biosystems, Foster City, CA, USA). In addition, the quality of the spectra and/or identification matches was also assessed requiring a spectral quality score (SQS; Parker et al., 2004) above 1000.

3.2 Computational analysis

The MS analysis yielded 19 619 MS/MS spectra with complete quantitative information, and identified 2443 proteins based on two or more of the 16 785 unique peptides. All reporter ion profiles were
sum-normalized and subjected to two computational analyses: (i) PSS was carried out as described in the previous section, with a DLRT significance level of 0.01 and (ii) PCP (Andersen et al., 2003). The resulting distance measurements and χ² values were used to derive a ranked protein list for each method. In both cases, we selected CCNB1 as a reference, and derived the top 1% shortlist for the proteins in the sample whose protein-level abundance profiles are most similar to the ones of the reference.

4 RESULTS

Table 1 lists the ranks of 10 known APC/C substrates and PRC1 that were observed in the acquired data as reported by PSS and PCP. See the Supplementary Material for detailed references concerning the chemical validation of the respective compounds. The CCNB1 reference profile is reported with rank zero and excluded from all following statistics.

Figure 3 displays the normalized peptide reporter ion profiles (gray lines) for the same set of proteins along with the geometric means over the profiles of all associated peptides. The geometric means serve as a measure of simplicial central tendency and are suitable for visual comparison and discussion of the results. High-ranking substrates (TK1, NUSAP, PLK1, TPX2) and PLK1 exhibit U-shaped tendencies similar to CCNB1, whereas the low-ranking AURKA, CDCA5, DNMT1 and GTSE1 show clearly different tendencies.

At a 1% confidence level, PSS reports five of the known APC substrates, PCP reports two. Both approaches report confident hits for PRC1, a mitotic spindle-associated microtubule bundling protein that is essential to cell cleavage. Its tight regulation is necessary to maintain the spindle midzone and to guarantee microtubule interdigitation. For PRC1, there is a body of evidence indicating that it tightly co-regulates with CCNB1 and that it indeed may be an APC/C substrate (Jiang et al., 1998; Mollinari et al., 2002), although biological validation is still pending. For all following statistics, we included PRC1 into the list of known coregulating proteins.

The PSS results on the APC/C iTRAQ dataset yield an 50.9-fold enrichment of CCNB1 co-regulated proteins as compared with the original raw data; the likelihood to observe an CCNB1-coregulating protein (i.e. an APC/C substrate candidate) in the set of significant proteins in the sample: peptide reporter ion profiles are shown in gray, protein-wise tendencies support the algorithmic findings that the data do not exhibit detectable correlation for AURKA, CDCA5, DNMT1 and GTSE1.

<table>
<thead>
<tr>
<th>Description</th>
<th>PSS</th>
<th>PCP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCNB1: G2/mitotic-specific cyclin-B1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TK1: Thymidine kinase cytosolic</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>PRC1: Protein regulator of cytokinesis</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>TPX2: Targeting protein for Xklp2</td>
<td>7</td>
<td>54</td>
</tr>
<tr>
<td>NUSAP: Nucleolar/spindle-assoc protein 1</td>
<td>12</td>
<td>623</td>
</tr>
<tr>
<td>PLK1: Serine/threonine-protein kinase</td>
<td>24</td>
<td>28</td>
</tr>
<tr>
<td>CKAP2: Cytoskeleton-associated protein 2</td>
<td>399</td>
<td>624</td>
</tr>
<tr>
<td>AURKA: Serine/threonine-protein kinase 6</td>
<td>548</td>
<td>186</td>
</tr>
<tr>
<td>CDCA5: Sororin</td>
<td>1565</td>
<td>1958</td>
</tr>
<tr>
<td>DNMT1: DNA methyltransferase 1</td>
<td>1598</td>
<td>876</td>
</tr>
<tr>
<td>GTSE1: G2 and S phase-expressed protein 1</td>
<td>1724</td>
<td>373</td>
</tr>
<tr>
<td>Confirmed proteins in top 1% ranks</td>
<td>5/10</td>
<td>2/10</td>
</tr>
<tr>
<td>Ratio of confirmed proteins (q = 1%)</td>
<td>20.8%</td>
<td>8.3%</td>
</tr>
<tr>
<td>Enrichment factor (q = 1%)</td>
<td>50.9</td>
<td>20.4</td>
</tr>
</tbody>
</table>

The table displays the list of known (i.e. biochemically validated) APC/C substrates present in the sample. The entries are ordered by the ranking derived from computational PSS and annotated with the ranking delivered by PCP (Andersen et al., 2003). PSS identifies 5 of the 10 known coregulating proteins among the top 1% ranks whereas PCP identifies only two. PSS thus yields a 50.9-fold enrichment of CCNB1-coregulation candidates among the top 1% proteins in the shortlist and a 2.5-fold increase compared with PCP.
5 DISCUSSION

The biologically validated set of top-ranked APC/C substrates includes: CCNB1, TK1, NUSAP, PLK1, TPX2 and PRC1. The examination of the peptide reporter ion profiles of the known APC/C substrate (AURKA, CDC5, DNMNT1 and GTSE1), which were not reported as coregulation candidates at a 1% cutoff shows significant deviations from the CCNB1 reporter ion profiles (Fig. 3). The two observable peptide reporter ion profiles for CKAP2 exhibit a U-shape with higher starting and lower ending points compared with CCNB1. The cluster assignment of one of the peptide profiles is close to a CCNB1 cluster (data not shown). However, because only two reporter ion profiles are available, only half of the CKAP2 protein signature matched to CCNB1; we assume that if better sequence coverage were available, CKAP2 would be ranked closer to the top. In this context, limiting the approach to proteins with a minimum amount of sequence coverage might be a worthwhile step to increase the screening accuracy. In summary, the proteins that fall out of the top 1% ranks feature protein signatures very different from the reference which result in increased distance measures. This intuitive assessment of performance also underlines the different distance measures used by PSS and PCP: PCP orders PLK1 and AURKA further to the top. This is due to the definition of the median and in particular in the case of AURKA, the median-based PCP delivers less intuitive results than PSS. Based on the experiments conducted in this study, PSS provides promise for practical application: among the top 1% ranked proteins, the likelihood of finding a truly coregulating protein was 2.5 times higher with PSS than with PCP; given that screening experiments in general need to be followed up with labor-intensive biological validation, this is a significant difference.

6 CONCLUSIONS

The proposed data analysis procedure enables PSS from IMT experiments. The procedure introduces novel statistical methodology for the treatment of IMT abundance reporter ion profiles that takes into account the dependency structure inherently present in the measurements. It also introduces advances in exploratory data analysis that enable protein-level inference based on peptide-level measurements. The experimental results indicate that the methodology is sufficiently powerful to cope with practical requirements. In addition, the protein signatures $s_p$ hold the information across which reporter ion profile clusters the peptides of a particular protein are distributed. This information can be used to gain insight if different homologs of a protein are present in an experiment. PSS identifies proteins with similar abundance profiles without the need for tailored biochemistry or high-effort experimental protocols. In particular, the method is applicable to full cell lysate measurements at endogenous protein levels. As a consequence, the method is unbiased. In practical application, similarity screening is carried out in a fully automated manner, requiring only a single, well-interpretable user-parameter (the DLRT significance level). The underlying statistical methodology is thus applicable to a wide range of proteomic research questions.

Ultimate validation of substrate relationships has to be carried out in the biochemical domain. However, in the case of APC/C co-regulation, our findings indicate that high-confidence candidates reported by the proposed methodology are well-chosen candidates for biochemical validation. Of particular importance for the proposed approach is the fact that each analysis step makes use of the correct metrics with respect to the underlying statistical dependency structures. Thus, the overall approach maintains statistical power and is able to generate usable results even with comparatively small sample sizes. The underlying methods, including the DLRT, can be applied to a wide field of use cases and PSS can be used as a drop-in replacement for PCP.

Future developments in time-resolved IMT experiments will likely include the ability to measure the sample under investigation at much better temporal resolution, providing a much more complete description of quantitative protein behavior and a significant increase in the amount of available discriminative information.

ACKNOWLEDGEMENTS

The authors would like to thank Michael Hanselmann, Xinghua Lou (Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Germany), Flavio Monigatti and Wiebke Timm (Department of Pathology, Children’s Hospital, Boston, MA, USA) for comments, suggestions and fruitful discussions.

Funding: DFG (HA436/2-1 to B.Y.R. and F.A.H.; KI1498/1-1 to M.K.); the Alexander von Humboldt-Foundation (3.1 DEU/1134241 to M.K.); the Children’s Hospital Trust (to J.A.J.S. and H.S.i; Harvard Medical School (Junior Faculty Award to J.A.J.S.).

Conflict of Interest: none declared.

REFERENCES

Computational PSS for quantitative MS experiments


Rubin, Y. et al. (1998) A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision*.


