BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 1 2010, pages 77–83
doi:10.1093/bioinformatics/btp607

Gene expression

Computational protein profile similarity screening for quantitative

Marc Kirchner[1,2,†], Bernhard Y. Renard[1,3,†], Ullrich Köthe[3], Darryl J. Pappin[4], Fred A. Hamprecht[1,3], Hanno Steen[1,2,‡,*] and Judith A. J. Steen[5,‡]

[1]Department of Pathology, Proteomics Center, Children's Hospital Boston, [2]Department of Pathology, Harvard Medical School, Boston, MA, USA, [3]Interdisciplinary Center for Scientific Computing, University of Heidelberg, Heidelberg, Germany, [4]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY and [5]Department 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.

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; Puig *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

[*]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.

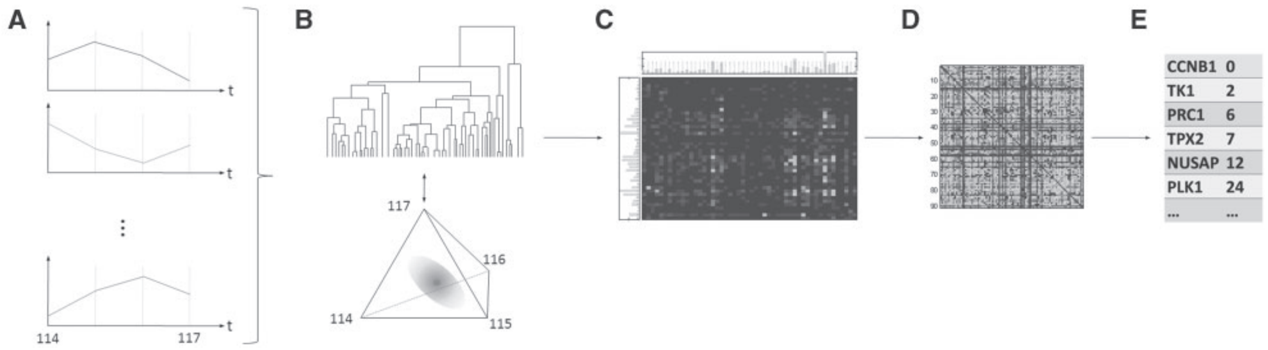

Fig. 1. Data analysis workflow for protein coregulation estimation: (**A**) IMT measurements yield sum-normalized quantitative peptide reporter ion profiles. (**B**) The reporter ion profiles are subjected to hierarchical clustering using an appropriate simplicial distance measure. The number of clusters is determined using a DLRT based on the observed peptide reporter ion profile distributions on the *n*-dimensional simplex. (**C**) Given the clustering, the quantitative measurements are grouped on the protein level, yielding a peptide cluster distribution for each protein. (**D**) The protein signatures are used to determine Mallows distances between proteins, taking into account the fact that the underlying clusters differ in their similarity. (**E**) The resulting distance matrix is subsequently evaluated to yield a shortlist of coregulation candidates.

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 analysis of compositional data (Aitchison, 1982, 1983, 1994) and introduce a novel approach to propagate quantitative profile information obtained from peptide measurements to the protein level. The proposed procedure creates a similarity-ranked shortlist of proteins in an automated and user-independent manner. Such shortlists are easily tested by biochemical assays and circumvent the laborious screening procedures that are currently used. The proposed method is evaluated on a biologically relevant example: we attempt to identify substrates of the E3 ubiquitin ligase Anaphase Promoting Complex/Cyclosome (APC/C), a master regulator of the cell cycle through M and G1 phases (Peters, 2006). We are interested in proteins that are degraded during mitosis or G1 phase, using quantitative proteomics data from an IMT-based experiment measuring the relative protein abundance at four different points during the cell cycle specifically chosen to profile the activity of the APC/C. Conventional assignment of a particular substrate to its unique E3 ubiquitin ligase is a laborious task involving the biochemical screening of hundreds or thousands of cloned and expressed proteins in biochemical assays. Consequently, computational screening procedures that help to prioritize among the candidates contribute to significantly reducing the biochemical effort.

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 LC/MS 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/balance/peptide compound breaks in three and yields K absolute reporter ion abundance measurements $\boldsymbol{x}=(x_1,x_2,\ldots,x_K)^T$, for each of the N peptide species. Given a protein, the vector $\boldsymbol{x}$ holds the respective reporter ion profile of observed abundances.

2.2.2 Normalization An absolute reporter ion profile $\boldsymbol{x}$ is subjected to variable interpeptide ionization efficiency (Song *et al.*, 2008; Turck *et al.*, 2007) and is dependent on the MS/MS sampling mode. Especially for data-dependent acquisition (DDA) schemes, MS/MS sampling depends on the sample complexity and there is no guarantee that MS/MS quantitation is carried out at the apex of peptide elution. In order to remove these effects, peptide reporter ion profiles need to be normalized. Commonly applied approaches include reference- or sum normalization, i.e. element-wise division by the abundance of a designated reporter ion or by the sum of all abundances, respectively. In both cases, the normalization eliminates 1 degree of freedom and a covariance/dependency structure is imposed on the measurements x_i (Supplementary Material). The following presentation studies the mathematically more tractable idea of sum normalization. It yields normalized abundance reporter ion profiles $\boldsymbol{x}^* = (x_1^*, x_2^*, \ldots, x_K^*)^T$, where $x_i^* = x_i / \sum_{j=1}^{K} x_j$. The loss of a degree of freedom is illustrated by the property that the relative intensity of any marker i can be recovered from the remaining normalized reporter ion intensities, i.e. $x_i^* = 1 - \sum_{j \neq i} x_j^*$.

2.3 Clustering peptides on the simplex

2.3.1 Hierarchical clustering on the simplex In a first step, we group peptides that exhibit similar peptide reporter ion profiles using a hierarchical clustering procedure (Johnson, 1967). 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 $\boldsymbol{x}^*$ add to 1, i.e. $\sum_{i=1}^{n} x_i^* = 1$. This defines a hyperplane in K dimensions and every vector $\boldsymbol{x}^*$ lies on a K-dimensional simplex. Standard distance measures like the Euclidean distance cannot account for such dependency structures and we thus resort to the natural measure of distance on the simplex (Aitchison, 1983) given by

$$\Delta_S(\boldsymbol{x}^*, \boldsymbol{y}^*) = \left[\sum_{i=1}^{K} \left(\ln \frac{x_i^*}{g(\boldsymbol{x}^*)} - \ln \frac{y_i^*}{g(\boldsymbol{y}^*)} \right)^2 \right]^{\frac{1}{2}}, \tag{1}$$

where $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$ are $K \times 1$ vectors of sum normalized reporter ion profiles and $g(\boldsymbol{x}^*) = \left(\prod_{i=1}^{K} x_i^* \right)^{1/K}$ denotes the geometric mean of $\boldsymbol{x}^*$. For the calculation of agglomerative distances during the clustering procedure, we use average linkage (Cortés *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 $\boldsymbol{x}^*$ as a realization drawn from a Dirichlet distribution

$$p(\boldsymbol{x}^* | \boldsymbol{\alpha}) = \mathcal{D}(\alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \left(x_i^* \right)^{\alpha_i - 1}, \tag{2}$$

where Γ is the Gamma function, $x_i^* > 0$, $\sum_{k=1}^{K} x_i^* = 1$ and Dirichlet parameters given by $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K), \alpha_i > 0$.

For the determination of statistical significance, we derive a likelihood ratio test (Casella and Berger, 2001) for the Dirichlet distribution. Assume we have two sets of observations $\mathcal{X}$ and $\mathcal{Y}$. We test whether the observations of the two groups stem from the same underlying Dirichlet distribution with parameter vector $\boldsymbol{\alpha}^{\mathcal{X} \cup \mathcal{Y}}$. In other words, we evaluate if the null hypothesis $H_0 : \boldsymbol{\alpha}^{\mathcal{X}} = \boldsymbol{\alpha}^{\mathcal{Y}}$ can be rejected in favor of the alternate hypothesis $H_1 : \boldsymbol{\alpha}^{\mathcal{X}} \neq \boldsymbol{\alpha}^{\mathcal{Y}}$.

Wilk's λ (Casella and Berger, 2001) is a measure of how well the data can be explained under H_0 versus H_1 and is given by

$$\lambda(\mathcal{X}, \mathcal{Y}) = \frac{L^{H_0}\left(\hat{\boldsymbol{\alpha}}^{\mathcal{X} \cup \mathcal{Y}} | \mathcal{X}, \mathcal{Y} \right)}{L^{H_1}\left(\hat{\boldsymbol{\alpha}}^{\mathcal{X}}, \hat{\boldsymbol{\alpha}}^{\mathcal{Y}} | \mathcal{X}, \mathcal{Y} \right),} \tag{3}$$

with the likelihoods L^{H_0} and L^{H_1} given by the products of the individual likelihoods of the observations

$$L^{H_0}\left(\hat{\boldsymbol{\alpha}}^{\mathcal{X} \cup \mathcal{Y}} | \mathcal{X}, \mathcal{Y} \right) = \prod_{i=1}^{|\mathcal{X}|} p\left(\boldsymbol{x}_i^* | \hat{\boldsymbol{\alpha}}^{\mathcal{X} \cup \mathcal{Y}} \right) \prod_{i=1}^{|\mathcal{Y}|} p\left(\boldsymbol{y}_i^* | \hat{\boldsymbol{\alpha}}^{\mathcal{X} \cup \mathcal{Y}} \right) \tag{4}$$

$$L^{H_1}\left(\hat{\boldsymbol{\alpha}}^{\mathcal{X}}, \hat{\boldsymbol{\alpha}}^{\mathcal{Y}} | \mathcal{X}, \mathcal{Y} \right) = \prod_{i=1}^{|\mathcal{X}|} p\left(\boldsymbol{x}_i^* | \hat{\boldsymbol{\alpha}}^{\mathcal{X}} \right) \prod_{i=1}^{|\mathcal{Y}|} p\left(\boldsymbol{y}_i^* | \hat{\boldsymbol{\alpha}}^{\mathcal{Y}} \right). \tag{5}$$

The vectors $\hat{\boldsymbol{\alpha}}^{\mathcal{X}}$, $\hat{\boldsymbol{\alpha}}^{\mathcal{Y}}$ and $\hat{\boldsymbol{\alpha}}^{\mathcal{X} \cup \mathcal{Y}}$ denote the maximum likelihood Dirichlet distribution parameters estimated from the observations in $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{X} \cup \mathcal{Y}$. Since there is no closed form solution for the maximum likelihood estimator of the Dirichlet parameter vector $\boldsymbol{\alpha}$, we follow previous approaches (Minka, 2004; Ronning, 1989; Wicker *et al.*, 2008) and estimate $\boldsymbol{\alpha}$ based on a Newton–Raphson approximation scheme with a method of moments initialization. For inference, we take advantage of Wilk's λ and define $t = -2\log(\lambda(\mathcal{X}, \mathcal{Y}))$, where t can be shown to approximately follow a chi-square distribution $t \sim \chi_K^2$, and are thus able to compute (one-sided) P-values. The DLRT is the uniformly most powerful test (Casella and Berger, 2001) for the problem at hand.

2.3.3 Adaptive thresholding for cluster determination With the DLRT, it is possible to use a rigorous statistical testing scheme to determine adaptive thresholds in the clustering tree: starting from the root we conduct a DLRT for each cluster tree node. Given a predefined type-I error rate/alpha level (generally 0.05 or 0.01), we merge all tree leaves into a cluster if the P-value assigned to a node is larger than the alpha-level threshold. This implicitly determines the number of clusters and the top-down scheme circumvents potential multiple testing issues intrinsically related with bottom-up testing procedures (Benjamini and Hochberg, 1995).

2.4 Estimating protein profile similarity

2.4.1 Protein signatures To determine which proteins show similar reporter ion profiles over a set of K experiments, the quantitative peptide-level information needs to be aggregated. The DLRT-based peptide-level clustering identifies peptides with similar behavior and groups them into C clusters. We represent each of the P proteins observed in the MS/MS experiments by a $C \times 1$ peptide signature vector $\boldsymbol{s}_p$ with $p \in \{1, \ldots, P\}$. Hence, the element s_{pq} holds the ratio of peptides observed for protein p which fall into cluster q. Thus, making use of ratios we avoid a dependency on the absolute number of peptides that have been identified for a protein. In addition, the peptide cluster representation for proteins eliminates intracluster variance (which is then regarded as experimental noise) and serves as a data-dependent dimension reduction procedure, effectively projecting the protein onto the peptide clusters.

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.

2.4.2 Mallows distance An intuitive way of comparing two protein signatures $\boldsymbol{s}_k$ and $\boldsymbol{s}_l$ is to determine the least-effort redistribution of the mass of the signature $\boldsymbol{s}_k$ to yield $\boldsymbol{s}_l$, taking into account that the clusters which underlie the signatures exhibit different degrees of similarity. Mathematically, this leads to a discrete version of the Mallows distance

(Levina and Bickel, 2001; Rubner *et al.*, 1998): we define a discrete joint distribution $\boldsymbol{F}(s_k,s_l)=\{f_{ij}(s_k,s_l)\}$ of flows between the signature entries s_{ki} and s_{lj} of proteins k and l. We then identify the distribution $\boldsymbol{F}^*$ that minimizes the expected cost d_{ij}:

$$\boldsymbol{F}^*(\boldsymbol{s}_k,\boldsymbol{s}_l)=\arg\min_{\boldsymbol{F}}\left\{\sum_{i=1}^{C}\sum_{j=1}^{C}d_{ij}f_{ij}\left(\boldsymbol{s}_k,\boldsymbol{s}_l\right)\right\}. \qquad (6)$$

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

$$f_{ij}^*\left(\boldsymbol{s}_k,\boldsymbol{s}_l\right)\geq 0, \quad \text{and} \quad \sum_i\sum_j f_{ij}^*\left(\boldsymbol{s}_k,\boldsymbol{s}_l\right)=1, \qquad (7)$$

and their marginals must correspond to the signature vectors,

$$\sum_j f_{ij}^*\left(\boldsymbol{s}_k,\boldsymbol{s}_l\right)=\boldsymbol{s}_k, \quad \text{and} \quad \sum_i f_{ij}^*\left(\boldsymbol{s}_k,\boldsymbol{s}_l\right)=\boldsymbol{s}_l. \qquad (8)$$

The costs of changes d_{ij} are defined as the average squared distance between the peptide clusters i and j, i.e.

$$d_{ij}=\frac{1}{N_iN_j}\sum_{u=1}^{N_i}\sum_{v=1}^{N_j}(\boldsymbol{x}^{u*}-\boldsymbol{y}^{v*})^2, \qquad (9)$$

where $\boldsymbol{x}^{u*}$ with $u\in\{1,\dots,N_i\}$ represents all normalized reporter ion profiles of peptides in the i-th cluster and $\boldsymbol{y}^{v*}$ with $v\in\{1,\dots,N_j\}$ 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 $\boldsymbol{s}_k$ and $\boldsymbol{s}_l$ is then given by

$$m_{kl}=m(\boldsymbol{s}_k,\boldsymbol{s}_l)=\sum_{i=1}^{C}\sum_{j=1}^{C}d_{ij}f_{ij}^*(\boldsymbol{s}_k,\boldsymbol{s}_l). \qquad (10)$$

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

2.5 Identifying similar proteins

It is now possible to derive a shortlist of proteins that exhibit similar abundance profiles from the distance matrix $\boldsymbol{M}$. Given a known substrate protein p, the elements of the column vector $\boldsymbol{m}_p=(m_{1p},m_{2p},\dots,m_{Pp})^T$ are constrained to the interval [0,1] and approximately follow a beta distribution. The parameters α_p and β_p 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 t with a Mallows distance m_{tp} 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 χ^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).

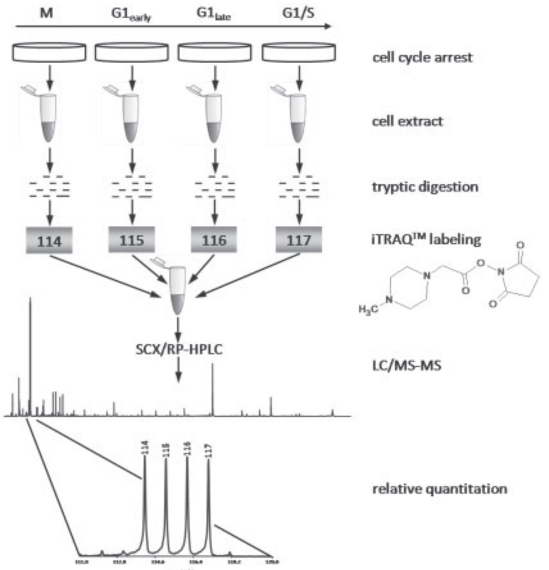

Fig. 2. Experimental setup: lysates from HeLa S3 cells were arrested in different states of the cell cycle. Samples were digested, iTRAQ-labeled, combined and analyzed by LC-MS/MS. Reporter ion profiles were acquired by subsequent quantitation and normalization.

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*-hydroxysuccinimide, 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 χ^2 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, DNMNT1 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 binding and 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 ranks is $5/24 = 20.8\%$ compared with $10/2443 = 0.41\%$ in the original unranked data. For PCP, we observe an enrichment factor of 20.4, corresponding to a likelihood of 8.3%. The fraction of confirmed proteins present in the top 1% ranks is $5/10 = 50\%$ for PSS and $2/10 = 20\%$ for PCP.

Table 1. Results of the CCNB1 PSS

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

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.

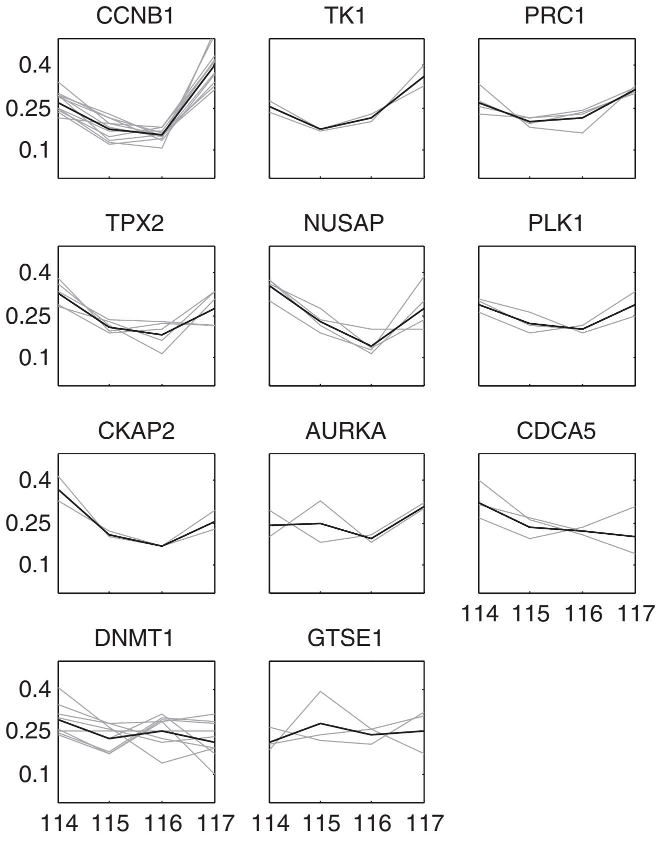

Fig. 3. Peptide reporter ion profile plots for all identified APC/C substrates in the sample: peptide reporter ion profiles are shown in gray, protein-wise geometric means are used as a measure of simplicial central tendency and shown in black.CCBN1 (upper left corner), the reference protein in the analysis, exhibits a U-shaped central tendency of peptide profiles which is shared by the coregulating proteins reported by the proposed screening procedure at the 1% level as well as by CKAP2. In the bottom row, the observed peptide reporter ion profiles and strongly diverging central tendencies support the algorithmic findings that the data do not exhibit detectable coregulation for AURKA, CDCA5, DNMT1 and GTSE1.

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, CDCA5, 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 overall algorithmic setup merely assumes sum-normalized relative quantification measurements, and 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 (Deparrment of Pathology, Children's Hospital, Boston, MA, USA) for comments, suggestions and fruitful discussions.

Funding: DFG (HA4364/2-1 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.); Harvard Medical School (Junior Faculty Award to J.A.J.S.).

Conflict of Interest: none declared.

REFERENCES

Aitchison,J. (1982) The statistical analysis of compositional data. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **44**, 139–177.

Aitchison,J. (1983) Principal component analysis of compositional data. *Biometrika*, **70**, 57–65.

Aitchison,J. (1994) Principles of compositional data analysis. *IMS Lect. Notes Monagr. Ser.*, **24**, 73–81.

Andersen,J.S. *et al.* (2003) Proteomic characterization of the human centrosome by protein correlation profiling. *Nature*, **426**, 570–574.

Bantscheff,M. *et al.* (2007) Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.*, **389**, 1017–1031.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)*, **57**, 289–300.

Bürckstümmer,T. *et al.* (2006) An efficient tandem affinity purification procedure for interaction proteomics in mammalian cells. *Nat. Methods*, **3**, 1013–1019.

Casella,G. and Berger,R.L. (2001) *Statistical Inference*. Duxbury Press.

Cortés,J.A. *et al.* (2007) Deciphering magma mixing: the application of cluster analysis to the mineral chemistry of crystal puopulations. *J. Vulcanol. Geoth. Res.*, **165**, 163–188.

Fields,S. and Song,O. (1989) A novel genetic system to detect protein-protein interactions. *Nature*, **340**, 245–246.

Foster,L.J. *et al.* (2006) A mammalian organelle map by protein correlation profiling. *Cell*, **125**, 187–199.

Hill,E.G. *et al.* (2008) A statistical model for iTRAQ data analysis. *J. Proteome Res.*, **7**, 3091–3101.

Jiang,W. *et al.* (1998) PRC1: a human mitotic spindle-associated cdk substrate protein required for cytokinesis. *Mol. Cell*, **2**, 877–885.

Johnson,S.C. (1967) Hierarchical clustering schemes. *Psychometrika*, **32**, 241–254.

King,R.W. (1995) A 20s complex containing cdc27 and cdc16 catalyzes the mitosis-specific conjugation of ubiquitin to cyclin b. *Cell*, **81**, 279–288.

Levina,E. and Bickel,P. (2001) The earth mover's distance is the Mallows distance: some insights from statistics. In *Proceedings of the Eighth IEEE International Conference on Computer Vision, 2001. ICCV 2001.*, Vol. 2, pp. 251–256.

Minka,T. (2004) fastfit.

Mollinari,C. *et al.* (2002) PRC1 is a microtubule binding and bundling protein essential to maintain the mitotic spindle midzone. *J. Cell Biol.*, **157**, 1175–1186.

Oberg,A.L. *et al.* (2008) Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. *J. Proteome Res.*, **7**, 225–233.

Ong,S.-E. and Mann,M. (2005) Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.*, **1**, 252–262.

Parker,K.C. *et al.* (2004) Depth of proteome issues: a yeast isotope-coded affinity tag reagent study. *Mol. Cell Proteomics*, **3**, 625–659.

Perkins,D.N. *et al.* (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.

Peters,J.-M. (2006) The anaphase promoting complex/cyclosome: A machine designed to destroy. *Nat. Rev. Mol. Cell Biol.*, **7**, 644–656.

Puig,O. *et al.* (2001) The tandem affinity purification (tap) method: a general procedure of protein complex purification. *Methods*, **24**, 218–229.

Rigaut,G. *et al.* (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.*, **17**, 1030–1032.

Ronning,G. (1989) Maximum likelihood estimation of Dirichlet distributions. *J. Stat. Comput. Simul.*, **32**, 215–221.

Ross,P.L. *et al.* (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics*, **3**, 1154–1169.

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

Selbach,M. and Mann,M. (2006) Protein interaction screening by quantitative immunoprecipitation combined with knockdown (QUICK). *Nat. Methods*, **3**, 981–983.

Song,X. *et al.* (2008) iTRAQ experimental design for plasma biomarker discovery. *J. Proteome Res.*

Tedford,N.C. *et al.* (2008) Illuminating signaling network functional biology through quantitative phosphoproteomic mass spectrometry. *Brief. Funct. Genomics Proteomics*.

Thompson,A. *et al.* (2003) Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.*, **75**, 1895–1904.

Turck,C.W. *et al.* (2007) The Association of Biomolecular Resource Facilities Proteomics Research Group 2006 study: relative protein quantitation. *Mol. Cell. Proteomics*, **6**, 1291–1298.

White,F.M. (2008) Quantitative phosphoproteomic analysis of signaling network dynamics. *Curr. Opin. Biotechnol.*, **19**, 404–409.

Wicker,N. *et al.* (2008) A maximum likelihood approximation method for Dirichlet's parameter estimation. *Comput. Stat. Data Anal.*, **52**, 1315–1322.