Computational Protein Coregulation Screening for Quantitative Mass Spectrometry Experiments

Marc Kirchner[1,2,4,‡], Bernhard Y. Renard[1,2,‡], Ullrich Koethe[1], Judith A. J. Steen[3], Hanno Steen[2,4,*], Fred A. Hamprecht[1,2]

‡Authors contributed equally.

[1]Interdisciplinary Center for Scientific Computing, University of Heidelberg, Heidelberg, Germany,
[2]Proteomics Center, Dept. of Pathology, Children's Hospital, Boston, MA, USA,
[3]Dept. of Neurobiology, Harvard Medical School and T. M. Kirby Neurobiology Center, Children's Hospital, Boston, MA, USA,
[4]Dept. of Pathology, Harvard Medical School, Boston, MA, USA

ABSTRACT

Motivation: The characterization of enzyme substrate specificity is a key step towards understanding signal transduction and protein interaction in cellular pathways. Exhaustive manual identification and biochemical validation of enzyme-substrate relationships is not feasible. Screening procedures that use quantitative protein reporter ion trace information to identify or computationally enrich specific groups of substrate candidates are necessary to optimize experimental design. This contribution introduces a computational screening procedure that provides enrichment of coregulated substrate candidates directly from endogenous quantitative mass spectrometry protein reporter ion trace measurements. Tailored statistical treatment enables the algorithm to aggregate peptide level information and to conduct protein level inference. It delivers a ranked shortlist that is enriched for specific groups of coregulated proteins and a starting point for targeted biological validation.

Results: The algorithm yields a 46-fold enrichment of anaphase promoting complex/cyclosome (APC/C) substrate candidates in an isobaric mass tagging experiment. Among 2443 identified proteins, seven of 11 known APC/C substrates coregulate with Cyclin-B1 (CCNB1), five of which are reported by the screening procedure.

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

Contact: hanno.steen@childrens.harvard.edu

1 INTRODUCTION

The biochemical processes that govern cell metabolism, proliferation and death are tightly controlled by complex interactions between numerous biomolecules [15]. In order to gain experimental insight into the organization, structure and the functional modules of subcomplexes of interaction networks or reaction pathways, it is necessary to understand the nature and role of the underlying processes. Pathways commonly involve a defined set of subsequent biochemical reactions, in some cases enzymatic post-translational modification (PTM) and subsequent degradation of proteins.

Understanding the relationship between PTM enzymes and the modified protein substrates yields information about the biological processes these substrates are associated with to. Further, it offers insight into conditional substrate specificity of the respective enzymes. Real-world samples often yield highly complex mixtures of proteins, making the exhaustive detection of the substrates of an enzyme non-trivial: biological validation of the substrate properties of all proteins present in a sample is infeasible for practical reasons. It is hence desirable to develop an in silico screening procedure that can output a shortlist of proteins ranked by their correlation with the abundance of a known substrate over a time-course or over a set of environmental conditions [45]. Although coregulated protein abundance levels do not prove common substrate properties, such a ranking can provide a valuable enrichment and prioritization of candidates for biological validation.

Post-translational modifications are not detectable by microarray technologies, and it is in this context that mass spectrometry (MS) proteomics methods provide direct, quantitative [28, 5] measurements of a multitude of peptides and proteins and their post-translationally modified forms at endogenous concentration levels in a single experiment. MS is thus a method of choice for the comprehensive analysis of protein abundance changes and their relationships under changing experimental conditions [12, 34, 32, 7, 39]. Combined with isobaric mass tagging (IMT) techniques [36, 43], quantitative MS enables differential protein expression analysis and biomarker detection in clinical applications. This includes the analysis of cell lysate, human blood serum, plasma, cerebrospinal fluid, tissue, or profiling of cells to identify differentially expressed proteins [40, 1, 19, 10, 45]. Recently, isobaric mass tagging procedures have also been applied to quantitative phosphoproteomic analyses of signaling networks [46, 42]. Within certain limitations [46], time-resolved IMT experiments allow for unbiased analyses of protein abundances at an endogenous level, obliviating the need for tedious precipitation procedures and complex biochemical protocols. IMT delivers rich datasets, and recent computational analyses [14, 27] have developed problem-specific rigorous statistical treatment.

Our study introduces an unbiased coregulation screening procedure for the accurate analysis of quantitative mass spectrometry measurements of IMT datasets. Regarding the statistical evaluation, we investigate the consequences of data normalization, which,

*to whom correspondence should be addressed

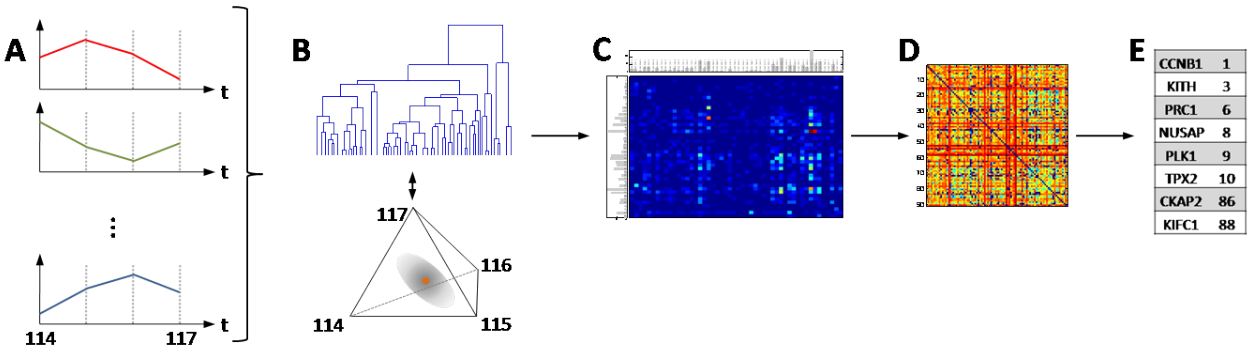

Fig. 1. Data analysis workflow for protein coregulation estimation: (A) isobaric mass tagging (IMT) measurements yield sum-normalized quantitative peptide reporter ion traces. (B) The reporter ion traces are subjected to hierarchical clustering using an appropriate simplicial distance measure. The number of clusters is determined using a Dirichlet Likelihood Ratio Test (DLRT) based on the observed peptide reporter ion trace 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.

if not accounted for, can jeopardize standard testing procedures. As a consequence, we establishm the connection between IMT series and the analysis of compositional data [2, 3, 4]. Finally, we introduce a novel approach to propagate information obtained from peptide level measurements to the protein level. The proposed procedure creates a ranked shortlist of co-substrate candidates in an automated and user independent manner. The small number of candidates renders biological validation feasible even even for complex mixtures of proteins.

Section 2 of the manuscript 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 kinase-substrate relationships 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 coregulation from isobaric mass tagging (IMT) analyses of proteomic time series experiments. Given a set of normalized IMT peptide reporter ion traces (figure 1, A), we deploy hierarchical clustering (figure 1, B) 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 (figure 1, C). The Mallows distance then provides a suitable measure for the inference of protein coregulation (figure 1, D). In the final step, a list of statistically significantly coregulated proteins is extracted (figure 1, E).

2.2 Statistical properties of IMT time-series measurements

Isobaric mass tagging: IMT labels generally consist of three parts: a reactive group which binds to the peptide, a reporter group and a balancer group. The sum of the three parts is isobaric however the reporter and balance groups are different combinations of heavy and light isotopes [36, 43]. 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 trace of observed abundances.

Normalization: An absolute reporter ion trace $\boldsymbol{x}$ exhibits inter-peptide ionization efficiency variability [44, 40] 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, a peptide reporter ion trace $\boldsymbol{x}$ needs 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 one degree of freedom and a covariance/dependency structure is imposed on the measurements x_i (see supplementary material). The following presentation studies the mathematically more tractable idea of sum normalization. It yields normalized abundance reporter ion traces $\boldsymbol{x}^* = (x_1^*, x_2^*, \ldots, x_K^*)^T$, where $x_i^* = x_i / \sum_{j=1}^K x_j$. The lost degree of freedom manifests itself in 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

Hierarchical clustering on the simplex: In a first step we group peptides which exhibit similar peptide reporter ion traces using a hierarchical clustering procedure. We use hierarchical clustering [17] as an unsupervised, agglomerative approach, which gradually

merges small, diverse, but internally homogeneous groups of peptide reporter ion traces into increasingly general, heterogeneous groups and eventually yields a binary clustering tree with a single root. 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 trace $\boldsymbol{x}^*$ 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 [3] given by

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

where $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$ are $K \times 1$ vectors of sum normalized reporter ion traces 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 [9].

Dirichlet Likelihood Ratio Test (DLRT): Hierarchical clustering is driven by sequential merging of subclusters. Given the resulting clustering tree, it is necessary to determine whether each merging can be statistically justified by the dissimilarity of the respective clusters. We approach this problem with a statistical hypothesis test for differences between groups of observations. Although this is a standard question in statistics, it is not possible to apply standard multivariate testing procedures since the normalized underlying data violate the necessary independence assumptions. Consequently, standard statistical tests would either yield unreliable p-values or a substantial decrease in test power. As an alternative, we interpret a normalized peptide reporter ion trace $\boldsymbol{x}^*$ as a realization drawn from a Dirichlet distribution, defined as

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

with Gamma function Γ, $x_i^* > 0$, $\sum_{k=1}^{K} x_i^* = 1$ and a parameter vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K), \alpha_i > 0$. The Dirichlet can be understood as a multivariate generalization of the beta distribution and is the natural distribution for independent measurements which are afterwards constrained to the simplicial domain.

We assume that the peptide abundance traces in each cluster of peptides are realizations from a single Dirichlet distribution; and we further assume that merging of two clusters is permissible whenever the respective Dirichlet distributions do not differ to a statistically significant extent. To implement this idea, we derive a likelihood ratio test [8] for the Dirichlet distribution: Given two sets of observations $\mathcal{X}$ and $\mathcal{Y}$, we test whether the null hypothesis that the observations of the two groups stem from the same underlying Dirichlet distribution with parameter vector $\boldsymbol{\alpha}^{\mathcal{X}\cup\mathcal{Y}}$ can be rejected, i.e. we evaluate

$$H_0 : \boldsymbol{\alpha}^{\mathcal{X}} = \boldsymbol{\alpha}^{\mathcal{Y}} \quad \text{versus} \quad H_1 : \boldsymbol{\alpha}^{\mathcal{X}} \neq \boldsymbol{\alpha}^{\mathcal{Y}}. \quad (3)$$

Wilk's λ [8] is a measure of how well the data can be explained given the null hypothesis and is given by

$$\lambda(\mathcal{X}, \mathcal{Y}) = \frac{\sup_{H_0} L_0\left(\boldsymbol{\alpha}^{\mathcal{X}\cup\mathcal{Y}}|\mathcal{X}, \mathcal{Y}\right)}{\sup_{H_0\cup H_1} L_1\left(\boldsymbol{\alpha}^{\mathcal{X}}, \boldsymbol{\alpha}^{\mathcal{Y}}|\mathcal{X}, \mathcal{Y}\right)} \quad (4)$$

$$= \frac{L_0\left(\hat{\boldsymbol{\alpha}}^{\mathcal{X}\cup\mathcal{Y}}|\mathcal{X}, \mathcal{Y}\right)}{L_1\left(\hat{\boldsymbol{\alpha}}^{\mathcal{X}}, \hat{\boldsymbol{\alpha}}^{\mathcal{Y}}|\mathcal{X}, \mathcal{Y}\right)}, \quad (5)$$

with the likelihoods L given by the products of the individual Dirichlet distributions

$$L_0\left(\hat{\boldsymbol{\alpha}}^{\mathcal{X}\cup\mathcal{Y}}|\mathcal{X}, \mathcal{Y}\right) = \prod_{i=1}^{m} p\left(\boldsymbol{x}_i|\hat{\boldsymbol{\alpha}}^{\mathcal{X}\cup\mathcal{Y}}\right) \prod_{i=1}^{m} p\left(\boldsymbol{y}_i|\hat{\boldsymbol{\alpha}}^{\mathcal{X}\cup\mathcal{Y}}\right) \quad (6)$$

$$L_1\left(\hat{\boldsymbol{\alpha}}^{\mathcal{X}}, \hat{\boldsymbol{\alpha}}^{\mathcal{Y}}|\mathcal{X}, \mathcal{Y}\right) = \prod_{i=1}^{m} p\left(\boldsymbol{x}_i|\hat{\boldsymbol{\alpha}}^{\mathcal{X}}\right) \prod_{i=1}^{m} p\left(\boldsymbol{y}_i|\hat{\boldsymbol{\alpha}}^{\mathcal{Y}}\right). \quad (7)$$

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 parameter estimates for the respective Dirichlet distributions which need to be estimated from the observations. This is complicated by the fact that there is no closed form solution for the maximum likelihood estimator (MLE) of the Dirichlet parameter vector $\boldsymbol{\alpha}$. We follow previous approaches [35, 47, 25] and estimate $\boldsymbol{\alpha}$ based on a Newton-Raphson approximation scheme with a method of moments initialization.

To allow 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 $t \sim \chi^2_K$, and are thus able to compute (one-sided) p-values. The DLRT can be shown to be the uniformly most powerful test [8] for the problem at hand.

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. The DLRT computes the p-value of the null hypothesis in equation (3) with $\mathcal{X}$ and $\mathcal{Y}$ being the sets of peptide reporter ion traces associated with the left and right branches of the 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 assigned p-value of 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 [6].

2.4 Estimating Protein Similarity

Protein Signatures: To determine which proteins show similar reporter ion traces over a set of K experiments, the aggregation of peptide-level information is required. The peptide-level clustering identifies peptides with similar behavior and groups them into clusters. We represent each of the P proteins observed in the MS/MS experiments by a peptide signature vector s with C components (with C the number of peptide clusters that result from the DLRT-truncated clustering). The element s_l^k holds the ratio of peptides observed for protein k which fall into cluster l. The peptide cluster representation for proteins eliminates intra-cluster 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. Using protein-normalized counts

removes the dependency on the absolute number of peptides that have been identified for a protein.

The rationale behind this approach is that IMT peptide reporter ion traces are susceptible to effects from post-translational modifications: in the presence of PTMs, peptides of a protein may exhibit very diverse reporter ion traces. Different types of reporter ion traces 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 abundance level regulation similarity.

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 [37, 21]: 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_i^k and s_j^l of proteins k and l. We then identify the distribution $\boldsymbol{F}^*$ that minimizes the expected cost d_{ij}:

$$\boldsymbol{F}^*(\boldsymbol{s}_k, \boldsymbol{s}_l) = \arg\min_{\boldsymbol{F}} \left\{ \sum_{i=1}^{C} \sum_{j=1}^{C} d_{ij} f_{ij} \left(\boldsymbol{s}^k, \boldsymbol{s}^l \right) \right\}. \quad (8)$$

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

$$f_{ij}^* \left(\boldsymbol{s}^k, \boldsymbol{s}^l \right) \geq 0, \quad \text{and} \quad \sum_i \sum_j f_{ij}^* \left(\boldsymbol{s}^k, \boldsymbol{s}^l \right) = 1, \quad (9)$$

and their marginals must correspond to the signature vectors,

$$\sum_j f_{ij}^* \left(\boldsymbol{s}^k, \boldsymbol{s}^l \right) = \boldsymbol{s}^k, \quad \text{and} \quad \sum_i f_{ij}^* \left(\boldsymbol{s}^k, \boldsymbol{s}^l \right) = \boldsymbol{s}^l. \quad (10)$$

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

$$d_{ij} = \frac{1}{N_i N_j} \sum_{u=1}^{N_i} \sum_{v=1}^{N_j} (\boldsymbol{x}^{u*} - \boldsymbol{y}^{v*})^2, \quad (11)$$

where $\boldsymbol{x}^{u*}$ with $u \in \{1, \ldots, N_i\}$ represents all normalized reporter ion traces of peptides in the ith cluster and $\boldsymbol{y}^{v*}$ with $v \in \{1, \ldots, N_j\}$ all normalized reporter ion traces of peptides in the jth 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

$$m_{kl} = m(\boldsymbol{s}^k, \boldsymbol{s}^l) = \sum_{i=1}^{C} \sum_{j=1}^{C} d_{ij} f_{ij}^*(\boldsymbol{s}^k, \boldsymbol{s}^l). \quad (12)$$

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

2.5 Identifying Coregulated Proteins

In order to analyze the substrate properties of a specific protein, we must generate a shortlist of coregulation candidates from the protein distance matrix $\boldsymbol{M}$. Given a known substrate protein p, the elements of the column vector $\boldsymbol{m}_p = (m_{1p}, m_{2p}, \ldots, m_{Pp})^T$ 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.

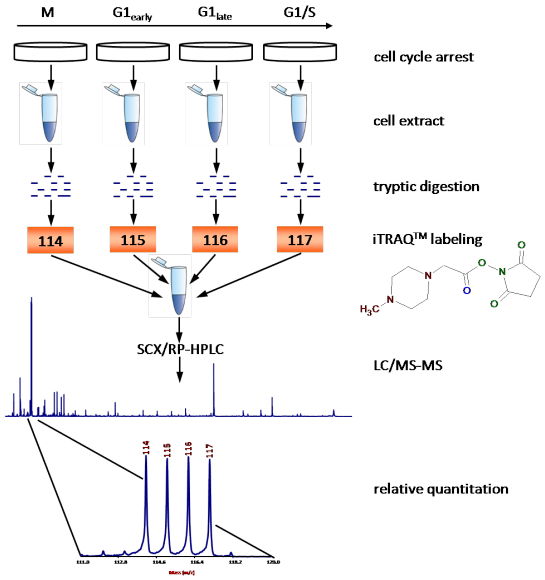

Fig. 2. Experimental setup: Lysates from HeLa S3 cells were arrested in different states of the cell cycle. Samples were digested, iTRAQ-labeled, combined and analyzed by LC-MS/MS. Reporter ion traces were acquired by subsequent quantitation and normalization.

3 EXPERIMENTS

We evaluated our method on an iTRAQ isobaric mass tagging MS experiment of the *Anaphase Promoting Complex/Cyclosome* (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, IPI00745793, P14635) [20].

3.1 Experimental Background

The data are from lysates of HeLa S3 cells arrested in four time points in the cell cycle: prometaphase, M/G1, G1 and G1/S (fig. 2). During the selected time course cells undergo division and the observed protein changes also reflect changes induced by APC/C activity. 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 [36] 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

Name	UniProt	Gene Name	AccNum	position	# pept.	ref.
G2/mitotic-specific cyclin-B1	P14635	CCNB1	IPI00745793	1	11	
Thymidine kinase TK1, cytosolic	P04183	TK1	IPI00299214	3	2	[18]
∗Protein regulator of cytokinesis 1	O43663	PRC1	IPI00022629	6	4	[16, 26]
Nucleolar and spindle-assoc. protein 1	Q9BXS6	NUSAP	IPI00000398	8	4	[22]
Serine/threonine-protein kinase PLK1	P53350	PLK1	IPI00021248	9	3	[11, 23]
Targeting protein for Xklp2	Q9ULW0	TPX2	IPI00008477	10	5	[41]
Cytoskeleton-associated protein 2	Q8WWK9	CKAP2	IPI00071824	86	2	[38]
Kinesin-like protein KIFC1	Q9BW19	KIFC1	IPI00306400	88	4	[45]
Serine/threonine-protein kinase 6	O14965	AURKA	IPI00298940	670	2	[24]
Sororin	Q96FF9	CDCA5	IPI00061989	1571	3	[33]
DNA (cytosine-5)-methyltransferase 1	P26358	DNMT1	IPI00031519	1951	9	[13]
G2 and S phase-expressed protein 1	Q9NYZ3	GTSE1	IPI00160901	2075	3	[31]

Table 1. Results of the Cyclin-B1 (CCNB1) coregulation screening for APC/C substrates: the table displays the list of known (i.e. chemically validated) APC/C substrates present in the sample. The entries are ordered by the ranking derived from the proposed screening procedure. Including PRC1 (a known coregulator with unclear substrate properties, marked by $*$), the screening procedure identifies 5 out of 11 CCNB1 coregulating proteins at a confidence level of 1%. The screening yields an approximately 46-fold enrichment of CCNB1-coregulation candidates among the first 24 proteins in the shortlist.

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) [30] 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 and matched to identified peptides using in-house software tools. In addition, the quality of the spectra and/or identification matches was also assessed requiring a spectral quality score (SQS) [29] above 1000.

3.2 Computational Analysis

The MS analysis yielded 19619 MS/MS spectra with complete quantitative information and identified 5258 proteins, 2443 of which were identified based on two or more of the 16785 unique peptides. All reporter ion traces were sum-normalized and subjected to the computational analysis described in the previous section. The DLRT significance level was set to 0.01. We selected Cyclin-B1 (CCNB1) as a reference protein, and derived a shortlist for the most similarly regulated proteins in the sample. Protein co-regulation candidates were treated as significant if their dissimilarity measure was below the 1% distance cutoff quantile.

4 RESULTS

After MS/MS analysis, the 11 known APC/C substrates listed in table 1 could be observed in the acquired data (see supplementary table 1 for a full list of all 45 known APC/C substrates). Along with names, accession numbers and references for chemical validation of the APC/C substrate properties of a protein, the table lists the number of observed peptides for a specific protein and its ranking as reported by the proposed screening procedure.

At a 1% confidence level, the screening reports five of the known proteins among the set of coregulators, increasing to seven at 5%. This includes CCNB1, on which the analysis was based. The procedure also reports a confident hit for PRC1 (protein regulator of cytokinesis 1). PRC1 is 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 coregulates with CCNB1 and that it indeed may be an APC/C substrate [16, 26], however, biological validation is still pending.

Consequently, among the set of confident coregulators, we observe (at least) 5 truly coregulating proteins, yielding a true positive ratio of $5/24 = 0.2083$. Remarkably, the 5 true positives are among the top ten hits in the shortlist.

Figure 3 displays the normalized peptide reporter ion traces (gray lines) for the 11 known APC/C substrates found in the sample and for PRC1 and their respective geometric means. The latter 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.

Overall, the screening results on the APC/C iTRAQ dataset yield an approximately 46-fold enrichment of CCNB1-coregulated proteins as compared to 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 $5/24 = 0.2083$ vs. $11/2443 = 0.0045$ in the original unranked data.

5 DISCUSSION

The biologically validated set of top-ranked APC/C substrates includes CCNB1, TK1, NUSAP, PLK1 and TPX2 and potentially PRC1. The examination of the peptide reporter ion traces of the known APC/C substrates (AURKA, CDCA5, DNMNT1 and GTSE1) which were not reported as coregulation candidates shows significant deviations from the CCNB1 reporter ion traces (see figure 3). Consequently, the proteins feature very different protein signatures and the reported results are supported by the measurements.

Of particular interest are CKAP2 and KIFC1 which are reported as coregulators on the 5%, but not on the 1% confidence level. They both exhibit U-shaped peptide reporter ion traces, but with

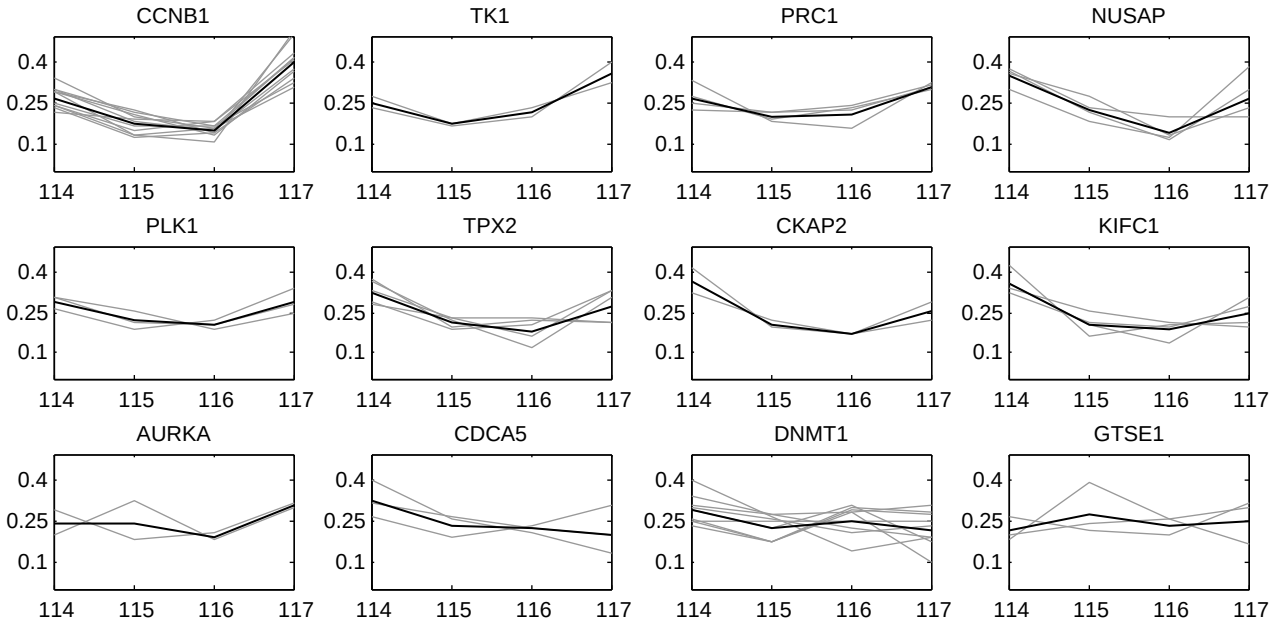

Fig. 3. Peptide reporter ion trace plots for all identified APC/C substrates in the sample: peptide reporter ion traces are shown in gray, protein-wise geometric means are used as a measure of simplicial central tendency and shown in black. Cyclin-B1 (CCBN1, upper left corner), the reference protein in the analysis, exhibits a U-shaped centra5 tendency of peptide reporter ion traces which is shared by the coregulating proteins reported by the proposed screening procedure at the 1% level as well as by CKAP2 and KIFC1 (reported at the 5% level). In the bottom row, the observed peptide reporter ion traces and strongly diverging central tendencies support the algorithms findings that the data do not exhibit detectable coregulation for AURKA, CDCA5, DNMT1, GTSE1.

higher starting and lower ending points compared to CCNB1. In the case of KIFC1, an overall larger variation is visible. For CKAP2, the two observable peptide reporter ion traces are similar to some of the CCNB1 reporter ion traces and the cluster assignment of one of the peptides is close to a CCNB1 cluster (data not shown). However, because only two reporter ion traces 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.

6 CONCLUSIONS

The proposed data analysis procedure enables protein level coregulation screening from series of isobaric mass tagging experiments. The procedure introduces novel statistical methodology for the treatment of IMT abundance reporter ion traces 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.

The proposed procedure identifies coregulation candidates 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, coregulation 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 IMT research questions.

Ultimate validation of substrate relationships has to be carried out in the biochemical domain. However, our findings indicate that high-confidence coregulation candidates reported by the proposed methodology are well-chosen candidates for chemical validation.

Of particular importance for the proposed approach is the fact that the correct metrics with regard to the underlying statistical dependency structures are employed for each analysis step. Thus, the overall approach gains statistical power and is able to generate usable results even with comparatively small sample sizes. The underlying methods, including the DLRT, can also be applied to other fields of application dealing with relatively quantified data, e.g. biomarker discovery.

Future developments in time-resolved isobaric mass tagging 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.

ACKNOWLEDGMENTS

The authors would like to thank Michael Hanselmann, Xinghua Lou (Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Germany), and Flavio Monigatti (Dept. of Pathology, Children's Hospital, Boston, MA, USA) for comments,

suggestions, and fruitful discussions. We gratefully acknowledge financial support by the the DFG under grant no. HA4364/2-1 (M.K., B.Y.R., F.A.H), the Children's Hospital Trust (J.A.J.S. and H.S.), Harvard Medical School for a Junior Faculty Award (JAJS) and Robert Bosch GmbH (F.A.H.).

REFERENCES

[1]F Abdi, J F Quinn, J Jankovic, M McIntosh, J B Leverenz, E Peskind, R Nixon, J Nutt, K Chung, C Zabetian, A Samii, M Lin, S Hattan, C Pan, Y Wang, J Jin, D Zhu, G J Li, Y Liu, D Waichunas, T J Montine, and J Zhang. Detection of biomarkers with a multiplex quantitative proteomic platform in cerebrospinal fluid of patients with neurodegenerative disorders. *J. Alzheimers Dis.*, 9:293–348, 2006.

[2]J Aitchison. The statistical analysis of compositional data. *Jounal of the Royal Statistical Society, Series B (Statistical Methodology)*, 44:139–177, 1982.

[3]J Aitchison. Principal component analysis of compositional data. *Biometrika*, 70(1):57–65, 1983.

[4]J Aitchison. Principles of compositional data analysis. *IMS Lecture Notes Monagraph Series*, 24:73–81, 1994.

[5]M Bantscheff, M Schirle, G Sweetman, J Rick, and B Kuster. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry*, 389(4):1017–1031, 2007.

[6]Y Benjamini and Y Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodogical)*, 57:289–300, 1995.

[7]T Bürckstümmer, K L Bennett, A Preradovic, G Schütze, O Hantschel, G Superti-Furga, and A Bauch. An efficient tandem affinity purification procedure for interaction proteomics in mammalian cells. *Nature Methods*, 3(12):1013–1019, 2006.

[8]G Casella and R L Berger. *Statistical Inference*. Duxbury Press, 2001.

[9]J A Cortés, J L Palma, and M Wilson. Deciphering magma mixing: The application of cluster analysis to the mineral chemistry of crystal puopulations. *Journal of Vulcanology and Geothermal Research*, 165:163–188, 2007.

[10]L DeSouza, G Diehl, M J Rodrigues, J Guo, A D Romaschin, T J Colgan, and K W Siu. Search for cancer markers from endometrial tissues using differentially labeled tags iTRAQ and cICAT with multidimensional liquid chromatography and tandem mass spectrometry. *J. Proteome Res.*, 4:377–386, 2005.

[11]D K Ferris, S C Maloid, and C C Li. Ubiquitination and proteasome mediated degradation of polo-like kinase. *Biochemical and Biophysical Research Communications*, 252(2):340–344, 1998.

[12]S Fields and O Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, 1989.

[13]K Ghoshal, J Datta, S Majumder, S Bai, H Kutay, T Motiwala, and S T Jacob. 5-aza-deoxycytidine induces selective degradation of dna methyltransferase 1 by a proteasomal pathway that requires the ken box, bromo-adjacent homology domain, and nuclear localization signal. *Molecular and Cellular Biology*, 25(11):4727–4741, Jun 2005.

[14]E G Hill, J H Schwacke, S Comte-Walters, E H Slate, A L Oberg, J E Eckel-Passow, T M Therneau, and K L Schey. A Statistical Model for iTRAQ Data Analysis. *Journal of Proteome Research*, 7(8):3091–3101, 2008.

[15]O N Jensen. Interpreting the protein language using proteomics. *Nature Reviews Molecular Cell Biology*, 7(6):391–403, 2006.

[16]W Jiang, G Jimenez, N J Wells, T J Hope, G M Wahl, T Hunter, and R Fukunaga. PRC1: A human mitotic spindle-associated cdk substrate protein required for cytokinesis. *Mollecular Cell*, 2(6):877–885, 1998.

[17]S C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(2):241–254, 1967.

[18]P-Y Ke and Z-F Chang. Mitotic degradation of human thymidine kinase 1 is dependent on the anaphase-promoting complex/cyclosome-Cdh1-mediated pathway. *Molecular and Cellular Biology*, 24(2):514–526, 2004.

[19]V G Keshamouni, G Michailidis, C S Grasso, S Anthwal, J R Strahler, A Walker, D A Arenberg, R C Reddy, S Akulapalli, V J Thannickal, T J Standiford, P C Andrews, and G S Omenn. Differential protein expression profiling by iTRAQ-2DLC-MS/MS of lung cancer cells undergoing epithelial-mesenchymal transition reveals a migratory/invasive phenotype. *J. Proteome Res.*, 5:1143–1154, 2006.

[20]R W King, J M Peters, S Tugendreich, M Rolfe, P Hieter, and M W Kirschner. A 20s complex containing cdc27 and cdc16 catalyzes the mitosis-specific conjugation of ubiquitin to cyclin b. *Cell*, 81(2):279–288, 1995.

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

[22]L Li, Y Zhou, L Sun, G Xing, C Tian, J Sun, L Zhang, and F He. NuSAP is ddegraded by APC/C-Cdh1 and its overexpression results in mitotic arrest dependent of its microtubules' affinity. *Cellular Signalling*, 19(10):2046–2055, 2007.

[23]C Lindon and J Pines. Ordered proteolysis in anaphase inactivates PLK1 to contribute to proper mitotic exit in human cells. *Journal of Cell Biology*, 164(2):233–241, 2004.

[24]L E Littlepage and J V Ruderman. Identification of a new APC/C recognition domain, the a box, which is required for the Cdh1-dependent destruction of the kinase Aurora-A during mitotic exit. *Genes & Development*, 16(17):2274–2285, Sep 2002.

[25]T Minka. fastfit, 2004.

[26]C Mollinari, J-P Kleman, W Jiang, G Schoehn, T Hunter, and R L Margolis. PRC1 is a microtubule binding and bundling protein essential to maintain the mitotic spindle midzone. *Journal of Cell Biology*, 157(7):1175–1186, 2002.

[27]A L Oberg, D W Mahoney, J E Eckel-Passow, C J Malone, R D Wolfinger, E G Hill, L T Cooper, O K Onuma, C Spiro, T M Therneau, and H R Bergen Iii. Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. *Journal of Proteome Research*, 7:225–233, 2008.

[28]S-E Ong and M Mann. Mass spectrometry-based proteomics turns quantitative. *Nature Chemical Biology*, 1(5):252–262, 2005.

[29]K C Parker, D Patterson, B Williamson, J Marchese, A Graber, F He, A Jacobson, P Juhasz, and S Martin. Depth of proteome issues: a yeast isotope-coded affinity tag reagent study. *Mol Cell Proteomics*, 3(7):625–59, 2004.

[30]D N Perkins, D J Pappin, D M Creasy, and J S Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–67, 1999.

[31]C M Pfleger and M W Kirschner. The KEN box: an APC recognition signal distinct from the d box targeted by Cdh1. *Genes & Development*, 14(6):655–665, Mar 2000.

[32]O Puig, F Caspary, G Rigaut, B Rutz, E Bouveret, E Bragado-Nilsson, M Wilm, and B Seraphin. The tandem affinity purification (tap) method: a general procedure of protein complex purification. *Methods*, 24(3):218–229, 2001.

[33]S Rankin, N G Ayad, and M W Kirschner. Sororin, a substrate of the anaphase-promoting complex, is required for sister chromatid cohesion in vertebrates. *Molecular Cell*, 18(2):185–200, 2005.

[34]G Rigaut, A Shevchenko, B Rutz, M Wilm, M Mann, and B Seraphin. A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*, 17(10):1030–1032, 1999.

[35]G Ronning. Maximum likelihood estimation of dirichlet distributions. *J. Statist. Comput. Simulation*, 32:215–221, 1989.

[36]P L Ross, Y N Huang, J N Marchese, B Williamson, K Parker, S Hattan, N Khainovski, S Pillai, S Dey, S Daniels, S Purkayastha, P Juhasz, S Martin, M Bartlet-Jones, F He, A Jacobson, and D J Pappin. Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & Cellular Proteomics*, 3(12):1154–1169, 2004.

[37]Y Rubner, C Tomasi, and L J Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision*, 1998.

[38]A Seki and G Fang. CKAP2 is a spindle-associated protein degraded by APC/C-Cdh1 during mitotic exit. *Journal of Biological Chemistry*, 282(20):15103–15113, 2007.

[39]M Selbach and M Mann. Protein interaction screening by quantitative immunoprecipitation combined with knockdown (QUICK). *Nature Methods*, 3(12):981–983, 2006.

[40]X Song, J Bandow, J Sherman, J D Baker, P W Brown, M T McDowell, and M P Molloy. iTRAQ Experimental Design for Plasma Biomarker Discovery. *Journal of Proteome Research*, 00:00, 2008.

[41]S Stewart and G Fang. Anaphase-promoting complex/cyclosome controls the stability of TPX2 during mitotic exit. *Molecular and Cellular Biology*, 25(23):10516–10527, 2005.

[42]N C Tedford, F M White, and J A Radding. Illuminating signaling network functional biology through quantitative phosphoproteomic mass spectrometry. *Briefings in Functional Genomics and Proteomics*, 2008.

[43]A Thompson, J Schfer, K Kuhn, S Kienle, J Schwarz, G Schmidt, T Neumann, R Johnstone, A K A Mohammed, and C Hamon. Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry*, 75(8):1895–1904, 2003.

[44]C W Turck, A M Falick, J A Kowalak, W S Lane, K S Lilley, B S Phinney, S T Weintraub, H E Witkowska, and N A Yates. The Association of Biomolecular Resource Facilities Proteomics Research Group 2006 study: relative protein quantitation. *Mol. Cell Proteomics*, 6:1291–1298, 2007.

[45]A Tzur, S T Liffers, F Monigatti, M Kirchner, B Y Renard, K C Parker, P Ross, D J Pappin, MW Kirschner, H Steen, and JAJ Steen. A quantitative proteomic analysis of the mammalian cell cycle identifies targets of the anaphase-promoting complex. *submitted*, 2008.

[46]F M White. Quantitative phosphoproteomic analysis of signaling network dynamics. *Current Opinion in Biotechnology*, 19(4):404–409, 2008.

[47]N Wicker, J Muller, Kalathur R K R, and O Poch. A maximum likelihood approximation method for Dirichlet's parameter estimation. *Computational Statistics & Data Analysis*, 52(3):1315–1322, 2008.