nature biotechnology

Computational prediction of proteotypic peptides for quantitative proteomics

Parag Mallick[1–3], Markus Schirle[4], Sharon S Chen[3], Mark R Flory[1], Hookeun Lee[1,5], Daniel Martin[1], Jeffrey Ranish[1], Brian Raught[1], Robert Schmitt[4], Thilo Werner[4], Bernhard Kuster[4] & Ruedi Aebersold[1,5]

Mass spectrometry–based quantitative proteomics has become an important component of biological and clinical research. Although such analyses typically assume that a protein's peptide fragments are observed with equal likelihood, only a few so-called 'proteotypic' peptides are repeatedly and consistently identified for any given protein present in a mixture. Using >600,000 peptide identifications generated by four proteomic platforms, we empirically identified >16,000 proteotypic peptides for 4,030 distinct yeast proteins. Characteristic physicochemical properties of these peptides were used to develop a computational tool that can predict proteotypic peptides for any protein from any organism, for a given platform, with >85% cumulative accuracy. Possible applications of proteotypic peptides include validation of protein identifications, absolute quantification of proteins, annotation of coding sequences in genomes, and characterization of the physical principles governing key elements of mass spectrometric workflows (e.g., digestion, chromatography, ionization and fragmentation).

The interpretation of genomic information in the context of the structure, function and control of biological systems is a fundamental goal of contemporary biology[1]. Several mass spectrometry–based quantitative proteomics methods attempt to comprehensively identify and quantify constituent proteins in complex mixtures[2,3]. Differences in the abundance of proteins in distinct samples have identified cellular functions and pathways affected by perturbations and disease[4,5], revealed new components and changes in the composition of protein complexes and organelles[6–8], and enabled detection of putative disease biomarkers[9].

In common mass spectrometry (MS)-based proteomic pipelines[1], protein samples are first partially purified or separated by chromatographic or electrophoretic methods and then digested with trypsin, often resulting in highly complex peptide mixtures. These are further separated by one or more stages of capillary liquid chromatography (LC) and analyzed using a tandem mass spectrometer. Peptides and proteins are subsequently identified by correlating spectra with a protein sequence database. Despite the success of these approaches, the complexity of biological peptide mixtures often overwhelms even the most modern tandem mass spectrometers. The efficiency and depth of the analysis is limited by the instrument's bias towards repeatedly sequencing the peptide species with the most intense MS signal. An alternate strategy that selectively and nonredundantly identifies one or a few peptides that uniquely identify a protein a priori would thus represent a significant advance in proteomics.

For quantitative proteomic analysis, label-free methods[10] simply use signal intensities to estimate peptide ratios between analyses and thus typically require complex data normalization. For label-based approaches, protein mixtures are modified to include an isotopic signature that identifies their sample of origin upon mixing and provides the basis for accurate quantification[11]. Using labeled synthetic peptide standards[12], it is possible to improve throughput in protein identification and quantification for those peptides expected to be present in a mixture. However, as any protein gives rise to tens or even hundreds of possible tryptic peptides, it is neither tractable, nor desirable, to synthesize all candidates. One must select a small panel that is likely to be observable and unique for their protein of origin.

Only a handful of a protein's possible tryptic peptides are consistently observed[13]. To illustrate this phenomenon, we first considered a carefully curated data set[14]. Briefly, 18 proteins were mixed into two sets of mixtures, A and B, with concentrations ranging from 4 nM (e.g., chicken ovalbumin) to 1,000 nM (e.g., bovine carbonic anhydrase) and subjected to 22 LC-tandem MS (MS/MS) analyses (14 for mixture A, and 8 for mixture B). **Figure 1a** shows tic-plots of the frequency with which peptides were observed for the eight most frequently identified proteins (remaining observed proteins are provided in **Supplementary Fig. 1** online). It is evident that even in such simple, well-defined samples, some peptides are preferentially identified whereas others are identified rarely or not at all, despite being within the mass range observable by the mass spectrometer. These frequently observed peptides are not necessarily derived from the most abundant proteins. For instance, a comparable number of peptides from glycogen phosphorylase (10, 100 nM) are observed as consistently as those from other proteins such as B-casein (1,000, 100 nM) and alkaline phosphatase (200, 1,000 nM), which are two orders of magnitude more abundant. Normalizing for a protein's length and

[1]Institute for Systems Biology, 1441 N. 34th Street, Seattle, Washington 98103, USA. [2]Cedars-Sinai Medical Center, 8750 W. Beverly Blvd, Los Angeles, California 90048, USA. [3]University of California, Los Angeles, 607 Charles E. Young Drive East, Box 951569, Los Angeles, California 90095-1569, USA. [4]Cellzome AG, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [5]Institute of Molecular Systems Biology, ETH Zurich and Faculty of Science, University of Zurich, Switzerland. Correspondence should be addressed to R.A. (rudolf.aebersold@imsb.biol.ethz.ch) or B.K. (Bernhard.kuester@cellzome.com).

Received 18 August; accepted 6 November; published online 31 December 2006; doi:10.1038/nbt1275

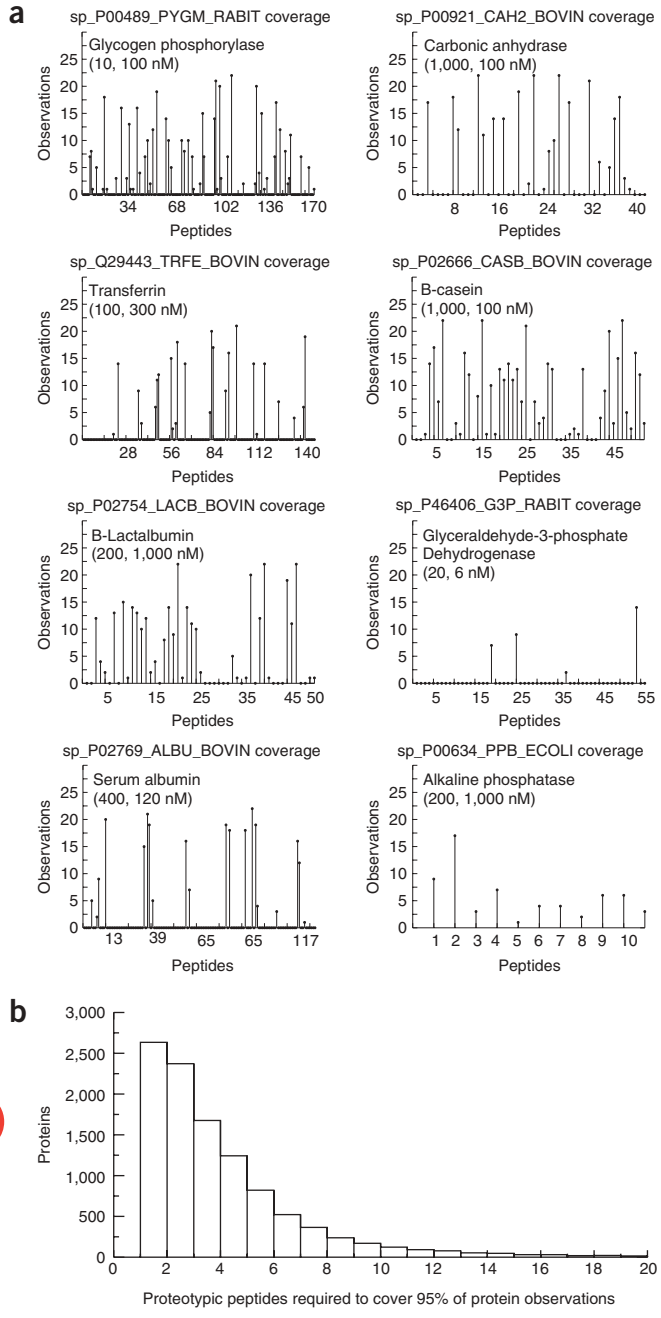


Figure 1 Proteomic data sets allow the identification of preferentially observed (proteotypic) peptides. (**a**) Tick plots showing the number of peptide observations in repeated LC-MS/MS analysis ($n = 22$) for 8 proteins out of a total set of 18 mixed at concentrations ranging from 4 nM to 1,000 nM[14]. Two mixtures, A and B, were constructed. The concentration of each analyte in mixtures A and B is provided in parenthesis as ([A],[B]). It is evident that some peptides are much more frequently observed than others and that this is not merely a function of the amount of protein present in the mixture or the lengths of the proteins. Plots for the remaining proteins are provided in **Supplementary Figure 1** online. (**b**) Analysis of a large PAGE-ESI data set comprising >10,000 human and murine proteins that have been identified >10 times reveals that the majority of all individual protein identifications are represented by the detection of one to a few repeatedly identified peptides (median = 3).

number of theoretically possible tryptic peptides only partially explains the disparity in the number of consistently observed peptides, suggesting that some peptides are more easily identified than others.

The bias in peptide observability is equally pronounced in complex mixtures. Using a large data set comprising >10,000 human and murine proteins that had been identified >10 times using LC-MS/MS, we assessed the minimum number of peptides required to cover 95% of a protein's observations (**Fig. 1b**). Very few (median = 3) distinct peptides are required to cover 95% of each protein's observations. Moreover, for more than 25% of the data set, only a single peptide was required. Consequently, we infer that estimates of protein identification probabilities and quantification may be based on the detection of one or a few preferentially observed, or proteotypic, peptides.

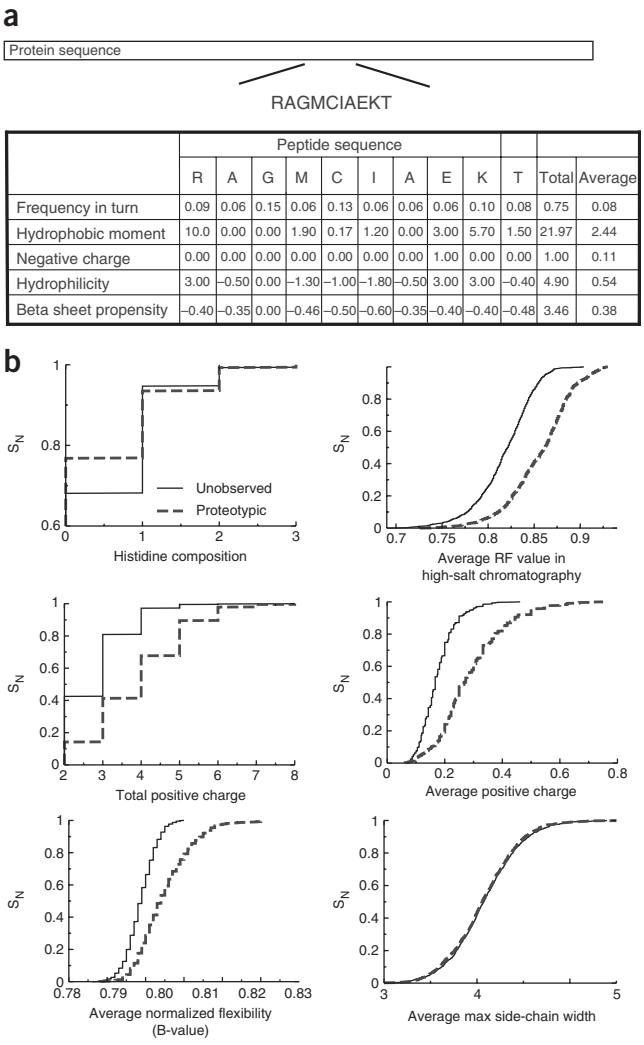
Although proteotypic peptides may be collected empirically in databases such as PeptideAtlas, GPM, SBEAMS and PRIDE[15–18], such collections are typically not available for many recently sequenced genomes, including some bacteria and archaea that are important model organisms in systems biology. Therefore, it is desirable to be able to predict proteotypic peptides for any protein from any organism whether or not these have previously been reported.

The existence of proteotypic peptides motivates several questions. For instance, what properties distinguish frequently observed peptides from peptides that are present in the sample but remain unidentified? What parts of the proteomic experimental process select for these peptides? Furthermore, given a protein's sequence, is it possible to predict which of its theoretical constituent peptides are likely to be proteotypic? We reasoned that prediction of proteotypic peptides might be achieved by first determining the physicochemical peptide properties that distinguish proteotypic peptides from less frequently or unobserved peptides. Towards this goal, we extracted the proteotypic peptides from four large and well-curated archives of yeast proteomic data representing four of the commonly used proteomic platforms (**Supplementary Table 1** online) and containing >600,000 peptide identifications covering 4,030 distinct proteins (61% of the yeast proteome, which contained 6,604 proteins as of this writing, http://www.yeastgenome.org). A peptide was classified as proteotypic if observed in >50% of all identifications of the corresponding protein (for lists of proteins and proteotypic peptides, see **Supplementary Tables 2** and **3** online).

To identify the properties governing a peptide's proteotypic propensity, we evaluated 494 numeric physicochemical property scales for amino acids, including charge, secondary structure propensity and hydrophobicity (**Supplementary Tables 4–7** for complete list)[19]. Each amino acid of a given peptide was replaced by a numerical value for each property and each property description string was summed and averaged for each peptide, resulting in a ~1,000-dimensional property vector per peptide (sum and average for each property plus the amino acid composition and length of the peptide) (**Fig. 2a**).

After describing the empirically derived proteotypic peptides as property vectors, we used classical pattern discovery methods to identify the smallest subset of properties that best distinguished a training set of proteotypic peptides from a training set of unobserved peptides.

Table 1 lists the selected set of discriminatory properties for each of the four experimental approaches, their cumulative prediction accuracy and area under curve (AUC) values. The cumulative prediction accuracy describes the ratio (total number of correct predictions/total number of predictions) using the threshold where the false-positive rate equals the false-negative rate.



	Peptide sequence											
	R	A	G	M	C	I	A	E	K	T	Total	Average
Frequency in turn	0.09	0.06	0.15	0.06	0.13	0.06	0.06	0.06	0.10	0.08	0.75	0.08
Hydrophobic moment	10.0	0.00	0.00	1.90	0.17	1.20	0.00	3.00	5.70	1.50	21.97	2.44
Negative charge	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.11
Hydrophilicity	3.00	−0.50	0.00	−1.30	−1.00	−1.80	−0.50	3.00	3.00	−0.40	4.90	0.54
Beta sheet propensity	−0.40	−0.35	0.00	−0.46	−0.50	−0.60	−0.35	−0.40	−0.40	−0.48	3.46	0.38

Figure 2 Peptide description by a numerical matrix of physicochemical properties identifies properties that discriminate between proteotypic and unobserved peptides. (**a**) Tryptic peptide sequences are described by summed and averaged numerical values of a total of 494 amino acid–associated physicochemical property scales. (**b**) Kolmogorov-Smirnov distributions for proteotypic peptides (broken lines) and unobserved peptides (solid lines) derived from the PAGE-MALDI data set highlight peptide properties that allow proteotypic peptides to be distinguished from unobserved peptides. Dissimilar distributions (e.g., average positive charge) imply that this property can discriminate between proteotypic peptides and unobserved peptides. Indistinguishable distributions (here, average max side-chain width) imply no such discrimination ability.

Two measures to quantify the strength of each property's differentiation power were assessed: the Kullbach-Leibler (KL) distance and the Kolmogorov-Smirnov (KS) distance. **Figure 2b** shows KS curves for the properties discovered as best distinguishing proteotypic peptides from unobserved peptides for the PAGE- matrix-assisted laser desorption ionization (MALDI) data set as well as an example for a nondiscriminating property. In addition, we randomly permuted the properties to compute the likelihood of achieving each KS-distance at random (P values uniformly $<$ 1e-5).

Having discovered the most discriminating physicochemical properties for each experimental platform, we applied these predictors to score the proteotypic propensity of each protein's theoretical tryptic peptides. Each protein's tryptic peptides were converted into a property vector, using only those properties found previously to be discriminating for the analytical platform under investigation. Using our training set of proteotypic peptides and unobserved peptides, we estimated a Gaussian mixture likelihood function to score the likelihood for a peptide to be proteotypic. We applied this scoring function to property vectors for peptides that had been excluded from training to derive receiver operator characteristics (ROC) curves (**Fig. 3a**), cumulative accuracy figures and AUCs (**Table 1**). As the number of nonproteotypic peptides exceeds the number of proteotypic peptides by several orders of magnitude, we used the positive-predictive value {true positives/(true positives + false positives)}, instead of the false-positive rate {false positives/(true negatives + false positives)}, because it allows a more accurate assessment of the impact of false positives in our calculation (see **Supplementary Discussion** online for a detailed explanation). Each of the classifiers was able to discriminate the vast majority of proteotypic peptides from nonproteotypic peptides with a cumulative accuracy nearing 90%. The predictors do not appear to be overtrained as the prediction accuracies with the training and testing sets are nearly identical (**Fig. 3b**). However, application of each predictor to other proteotypic peptide sets (e.g., the MALDI-PAGE predictor to the three electrospray ionization (ESI) peptide sets) substantially reduced prediction performance (**Fig. 3c**). We expect that the exponential growth of proteomics data repositories will enable the further refinement of predictors and enable development of predictors for experimental designs not covered in this study.

To test the generality of the predictors developed on the yeast data sets, we applied and compared them to a human PAGE-ESI data set comprising 500 repeatedly identified proteins and 19,732 peptides generated by the same experimental protocol used for the yeast training data. Predictions were validated by computing the coverage and specificity of the predictions on the human data. The application of the yeast PAGE-ESI proteotypic peptide predictor to a human PAGE-ESI data set resulted in no performance degradation for proteins as distant in phylogeny as yeast and humans (the ROC curves show a chi-squared similarity $P < 0.001$ for high-confidence predictions between yeast and human proteomes) (**Figure 3d**). This indicates that the predictors are not overtrained and that amino acid usage in proteins between yeast and human is not radically different. For example, there is good agreement between empiric and predicted proteotypic peptides even for difficult-to-analyze proteins such as human γ-secretase, a transmembrane multiprotein complex that is associated with the development of amyloid plaques in brains of Alzheimer patients[20] (**Supplementary Table 8** online). Most of the frequently observed peptides were also predicted to be proteotypic. Also note that the number of proteotypic peptides for a given protein is not merely a function of protein length. Although large proteins tend to have more proteotypic peptides, factors such as amino acid composition and the presence of transmembrane domains can be of significant influence.

We next applied our predictor to the entire yeast and human genomes to determine the distribution of proteotypic peptides per protein. Each method identified at least one high-confidence (>99%) proteotypic peptide for >60% of yeast proteins (62% of human proteins). For yeast, we predict two high-confidence proteotypic peptides on average per protein and experimental design (**Supplementary Figure 2** online). By allowing lower, but still reasonable confidence thresholds ($P > 0.9$, >30× better than choosing at random), our PAGE-MALDI predictor is able to identify a proteotypic

Table 1 Properties with the greatest capacity to discriminate between proteotypic peptides and unobserved peptides for different experimental data sets

Experiment type	Cumulative accuracy	ROC AUC	KS distance	KL distance
PAGE-ESI selected properties	90	94		
Average normalized flexibility parameters (B-values)			6.2	0.2
Histidine composition			3.1	0.2
Total normalized van der Waals volume			5.5	0.3
Total positive charge			14.3	1.2
Average positive charge			12.7	0.9
PAGE_MALDI Selected properties	89	94		
Histidine composition			17.8	0.01
Total positive charge			11.6	0.5
Average positive charge			14.8	0.8
Average RF value in high salt chromatography			13.7	0.6
Average normalized flexibility parameters (B-values)			12.1	0.6
MUDPIT-ESI Selected properties	87	91		
Average isoelectric point			12.7	1.2
Total propensity to be buried			8.9	0.2
Average net charge			12.8	1.1
Atom-based hydrophobic moment			10.9	0.9
Average positive charge			11.8	1.0
MUDPIT-ICAT Selected properties	88	96		
Retention coefficient at pH 2			8.3	0.3
Total surrounding hydrophobicity			7.4	0.1
Total positive charge			11.8	0.6
Total atom-based hydrophobic moment			12.3	0.6
Histidine composition			12.8	0.02

For each experiment, the cumulative prediction accuracy, ROC AUC values and most discriminating properties are shown. For each discriminating property, the Kolmogorov-Smirnov (KS) distance and the Kullbach-Leibler (KL) distance are shown. The KL distance (or relative entropy) is a measure of the 'distance' between two probability distributions (e.g., Gaussian). A large KL distance implies the distribution of proteotypic peptides is strongly different from the distribution of unobserved peptides and the given property thus a good predictor of proteotypic propensity. The KS distance is defined as the maximum value of the absolute difference between two cumulative distribution functions. A concurrent large KS and small KL can occur if the mean and s.d. of two distributions are similar, but the distributions are otherwise shaped differently. Plots visualizing the KS distance are provided in **Figure 2b**.

peptide for every protein in the yeast proteome (98% human proteome). Likewise our Multidimensional Protein Identification Technology (MudPIT) ESI, PAGE ESI and MUDPIT isotope-coded affinity tags (ICAT) predictors were able to identify peptides for >95% of yeast proteins (>85% of human proteome). Yeast and human peptide predictions are available for download at http://www.peptideatlas.org/ and are listed in **Supplementary Tables 9–12** online.

Predictive models based on experimental data have proven to be extremely powerful in biological research. Examples include the prediction of splice sites to annotate coding regions in genome sequences[21], description of functional aspects of proteins by sequence domains and motifs[22], prediction of the sequences likely to be presented by the major histocompatibility complex[23], prediction of protein phosphorylation motifs[24] and the prediction of oligonucleotides suitable for RNA interference experiments[25]. We anticipate that proteotypic peptides also have the potential to significantly affect the way proteomic experiments may be conducted.

In particular, it has often been noted that different sets of peptides are observed when using different MS-based approaches. Our data confirm that experimentally determined proteotypic peptides for MALDI and ESI are not synonymous (**Fig. 3c**, **Supplementary Fig. 3** online, **Supplementary Table 13** online and **Supplementary Table 14** online). Given the difference in the training sets, it is not surprising that the application of the PAGE-MALDI predictor to predict ESI proteotypic peptides resulted in significant performance degradation (**Fig. 3c**). Although the experimental bias is evident in our prediction, we note that (i) 95% of all yeast proteins have at least one proteotypic peptide predicted with 99% confidence, (ii) 2,394 yeast proteins (36% of the proteome) have a high-confidence predicted proteotypic peptide in all experimental designs, (iii) 770 proteins have a proteotypic peptide predicted in one experimental condition only and (iv) 304 proteins have no predicted high-confidence proteotypic peptide for any experimental design, suggesting that these proteins might be more difficult to analyze by current proteomic methods (**Supplementary Fig. 3**, **Supplementary Table 15**, **Supplementary Table 16** online). However, our data suggest that it is possible to get 99% coverage of the yeast proteome by integrating several experimental approaches. Furthermore, by allowing slightly lower confidence proteotypic peptides, it is possible to improve this coverage. The ability to predict which peptides are likely to be observed should be valuable for improving protein identification algorithms (**Supplementary Discussion** online).

The discovered properties governing proteotypic propensity differ between experimental designs (**Table 1** and **Supplementary Tables 4–7** online). Although the PAGE-MALDI and PAGE-ESI data sets share four of the top five properties (albeit leading to different outcomes as shown in **Figure 3c**), there is no overlap of top properties between MUDPIT-ESI and MUDPIT-ICAT data sets, possibly reflecting the different strategies used for peptide isolation. Among the discriminatory properties, several are related to charge. This may not be surprising as the detection of peptides and their fragments in a mass spectrometer are intimately linked to charge effects, for example, for initial peptide ionization, fragmentation and detection. Other properties may relate more specifically to a particular experimental approach. For example, it is possible that structural properties may be discriminatory because of bias in tryptic digestion. Hydrophobicity parameters may be related to the behavior of peptides in reversed-phase chromatography. Likewise, average isoelectric point may be related to the cation exchange chromatography used in MUDPIT experiments and hydrophobic moment may represent aspects of the mechanism of ionization within the ESI droplet. Initially, we had separated the ESI-PAGE experiment according to the instrument used for data acquisition (quadrupole TOF versus ion trap). However, results indicated only minimal differences in proteotypic propensity, suggesting that the influence of the type of mass analyzer used is not significant (ref. 26; data not shown). As the goal of this study was primarily to use physicochemical properties for prediction purposes, we did not require a formal, generative physical model of proteotypic propensity. Nevertheless, our analysis of the ~500 physicochemical properties should provide a basis for future research towards the mechanistic description of the parameters that govern proteomic experiments and how these can be used to improve experimental design and outcome. It is presently impossible to

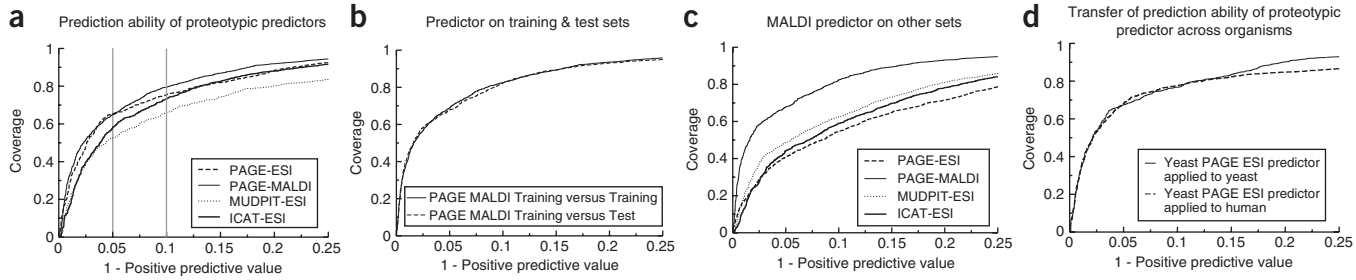


Figure 3 Evaluation of the prediction ability of proteotypic peptide predictors and application of the yeast PAGE-ESI predictor to human proteins. (**a**) ROC plots using the false-positive ratio (1 – positive-predictive value) and false-negative ratio (coverage) are used to evaluate the ability of the algorithms to predict proteotypic peptides for four commonly used proteomic analysis strategies. The four predictors cover 65–80% of the true-positive proteotypic peptides with >90% positive predictive value. (**b**) Applying the PAGE-MALDI predictor to both the training and test data results in almost identical performance, demonstrating minimal if any overtraining of the predictor. (**c**) Application of the PAGE-MALDI predictor to the test data from each of the four experimental strategies substantially reduces performance, demonstrating that different physicochemical peptide properties govern the detection of peptides analyzed by different experimental strategies. (**d**) ROC plot illustrating similar performance of the yeast PAGE-ESI predictor when applied to yeast data (solid line) or to a set of 500 human proteins generated by the same experimental approach (broken line). This suggests that predictors developed for yeast can be generally applied to human proteins and possibly to any protein regardless of species of origin.

generate databases and predictors for all combinations of variations of the common proteomic workflow but with the rapid increase of data deposited in public repositories, this might become possible.

As numerous studies have demonstrated positional effects related to peptide fragmentation and ionization[27,28], additional predictive power for proteotypic propensity prediction might be derived from a model incorporating residue order information. We discuss positional information in more detail in the **Supplementary Results** online. Briefly, we have observed that such effects for the basic amino acids arginine and lysine can be linked to trypsin digestion as well as for cysteine in ICAT experiments (**Supplementary Fig. 4** online).

The excellent performance of our predictors opens up the exciting possibility that these can be applied universally to any protein of any species of origin as long as DNA sequence information for this protein is available. Predictors derived from physicochemical properties of the constituent amino acids of proteotypic peptides should enable rational selection of synthetic peptides for absolute protein quantification as well as improve label-free quantification algorithms that rely on correlating spectra of observed peptides with protein quantity (**Supplementary Results** online). We are making the predictors publicly available and anticipate that the community will find other creative uses for proteotypic peptides.

METHODS

Defining empiric proteotypic peptides. Four experimental approaches were used herein. ICAT-labeling followed by LC-ESI-MS/MS, ICAT-ESI[11], one-dimensional (1D) gel electrophoresis followed by LC-ESI-MS/MS (PAGE-ESI) (T.W. *et al*, unpublished data), two-dimensional LC-ESI-MS/MS (MudPit-ESI)[29] and 1D gel electrophoresis followed by PAGE-MALDI[30]. The experimental databases of peptides and proteotypic peptides used herein are of sufficient size and quality to yield statistically significant results from trained predictors (see below and supplement for details).

Different proteins and peptides were observed by different experimental techniques (**Supplementary Fig. 3**). Given the possibility that the physicochemical properties governing the proteotypic propensity of a peptide are related to the experimental method used, further analyses were performed for each data set separately. In fact, pilot studies of integrated data sets did not generate predictors of the same quality (data not shown). Peptide identification frequencies were determined from repeat identifications of a given protein in multiple experiments. To ensure the studied effects were minimally a function of protein abundance, only proteins identified in multiple experiments by multiple different peptides were included in the study. A peptide was denoted proteotypic if observed in more than 50% of all identifications of the corresponding protein. For each protein for which proteotypic peptides had been identified, *in silico* digests were performed and peptides were grouped into: (i) proteotypic, (ii) rarely observed, (iii) peptide-segment observed (those peptides which may not have been observed themselves, but contain segments that were) and (iv) unobserved. For each experimental design, we constructed three sets of peptides: proteotypic peptides, nonproteotypic peptides that were either a subset or superset of a proteotypic peptide, and distinct nonproteotypic peptides. Adjacent arginine and lysine residues were included in the pattern discovery to partially account for digestion bias. The miscleavage, mass and length distributions of the unobserved peptides in the training and test sets were matched to those of the observed peptides (**Supplementary Figs. 5** and **6**). By matching these distributions we ensured discovered properties were not a trivial artifact of a poorly constructed negative training set. We discuss the importance of this synchronization in Supplementary Information.

The necessity for compression. The representation of each peptide as total and average properties discards residue order information, but allows for sufficient sampling for statistical significance. Ten or more orders of magnitude more data would be required to discover a statistically significant classifier that incorporated both physicochemical property and residue order (**Supplementary Discussion** online).

Although our studies have indicated that bulk, whole-peptide properties contain sufficient information, for practical applications, there are numerous experimental studies in progress that attempt to systematically perturb peptide sequence order to uncover its impact on ionization and fragmentation. The results of these studies suggest that additional information could theoretically be gained by considering residue order information as described in supplementary information. However, such global discovery analyses are presently intractable. A discussion of the data requirements of an order-dependent analysis is described in supplementary material as well.

Classification functions distinguishing proteotypic peptides from nonproteotypic peptides. Briefly, to determine the most discriminating set of properties, we first trained a classification function (predictor) for each property individually. Classification functions take a property, or set of properties as input and return a classification score describing how likely a peptide is to be proteotypic. Ideally the distributions of scores for experimentally observed proteotypic peptides will be significantly different than the scores for nonproteotypic peptides. The sensitivity and specificity of each predictor is related to the distinctiveness of these two distributions. We calculate the sensitivity and specificity of each predictor on a set of test peptides; from this information the predictor's ROC AUC was determined. ROC curves describe the relationship

between a classifier's true-positive and false-positive rates. Typically, by decreasing the value of a classification threshold (e.g., the score above which peptides are classified as proteotypic, and below which peptides are classified as not proteotypic) one is able to make more true-positive predictions at the cost of making more false-positive predictions as well. An ideal ROC curve is a horizontal line at 1 indicating 100% of true proteotypic peptides could be classified to be proteotypic (100% coverage) at the same time as 0% of nonproteotypic peptides were classified as proteotypic (0% false positives). Consequently, the maximum AUC possible is 100.

Pattern discovery for properties that distinguish proteotypic peptides from nonproteotypic peptides. Sets of properties discriminating proteotypic peptides from unobserved peptides were discovered using classical pattern discovery approaches. Validation of those properties revealed them to be statistically significantly differential on their own, and as panels. We define the ability of a set of properties to distinguish among our sets of peptides by using those properties and a set of training peptides to train a multivariate-normal mixture classifier and then by applying that classifier to a set of evaluation peptides, and computing the area under a receiver-operator-characteristic plot (AUC) describing their classification performance.

To determine the optimal properties, we used a hierarchical hill-climbing search approach where the prediction ability of each property or set of properties is assessed. We discuss the implications of this style of statistical learning in Supplementary Information. For each property or set of properties under investigation, a Gaussian mixture discriminant function was developed from a training set. This function takes a peptide's property value(s) and generates a score for a peptide's proteotypic propensity. Ideally, the score of this function is high for sets of properties generated from proteotypic peptides and low for property strings generated from unobserved peptides. Property sets were allowed to contain as many as 20 properties. However, training attempted to find the smallest number of properties capable of generating a given classification strength. The properties whose predictors had the highest AUCs at a given round became the starting seeds for the next iteration. For example, once all single properties had been explored, sets containing two properties were explored. We developed predictors for all pairings of the 1,010 properties with the starting seeds, again calculating sensitivity, specificity and AUCs. Next, sets containing three properties were explored and so on. Iteration was continued until there was no gain in prediction ability. Several combinations of properties had comparable prediction performance due to correlation of the corresponding properties, for example, multiple correlated hydrophobicity scales.

In all cases, classifiers containing five properties were selected, however, the algorithm in no way enforced this number of properties. The sensitivity and specificity of each predictor was calculated on a set of test peptides and the predictor's ROC AUC was determined. A pruning operation was performed every two iterations to determine if it is possible to achieve comparable performance with a subset of a predictor's properties, or including a different property instead of one of the three selected. To validate the significance of the properties selected by the ultimate predictor, random sets of properties were used to develop predictors and assess their prediction ability. Approximately 1.7×10^6 different combinations of properties were explored for each peptide set to estimate the distribution of AUCs and thus to estimate the probability of getting an AUC of equal to or better than our ultimate predictor.

KS and KL distances between observed and unobserved peptide properties. The KL distance (or relative entropy) is a measure of the 'distance' from a probability distribution P (proteotypic peptides) to a probability distribution Q (unobserved peptides). A large KL distance implies that the distribution of proteotypic peptides is strongly different from the distribution of unobserved peptides and the given property thus a good predictor of proteotypic propensity. The Kolmogorov-Smirnov distance is defined as the maximum value of the absolute difference between two cumulative distribution functions. It can be visualized by plotting the two empirically estimated cumulative distribution functions S_N. Large differences between the curves (highly discriminatory properties) result in large KS values.

URLs. Predictors are available at http://www.proteomecenter.org/.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

The authors are grateful to Julien Gagneur for fruitful discussions and the Cellzome biochemistry, mass spectrometry and informatics teams for generating and managing data. The work was supported in part with federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, under contract N01-HV-28179.

AUTHOR CONTRIBUTIONS

P.M., data (yeast MUDPIT-ESI), data analysis, idea and concept, wrote most of manuscript. M.S., data (yeast PAGE-MALDI, human PAGE-ESI), data mining, wrote part of manuscript. S.C., data (yeast MUDPIT-ESI), idea and concept. M.F., data (yeast MUDPIT-ICAT). H.L, data (yeast MUDPIT-ICAT), D.M., data (yeast MUDPIT-ESI). J.R., data (yeast MUDPIT-ESI, MUDPIT-ICAT). B.R., data (yeast MUDPIT-ICAT). R.S., computation underlying **Figure 1b**. T.W., data (yeast PAGE-ESI). B.K., idea and concept, wrote part of manuscript. R.A., idea and concept, wrote part of manuscript.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
2. Gerber, S.A., Rush, J., Stemman, O., Kirschner, M.W. & Gygi, S.P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci. USA* **100**, 6940–6945 (2003).
3. Ong, S.E., Foster, L.J. & Mann, M. Mass spectrometric–based approaches in quantitative proteomics. *Methods* **29**, 124–130 (2003).
4. Wright, M.E. *et al.* Identification of androgen-coregulated protein networks from the microsomes of human prostate cancer cells. *Genome Biol.* **5**, R4 (2003).
5. Durr, E. *et al.* Direct proteomic mapping of the lung microvascular endothelial cell surface *in vivo* and in cell culture. *Nat. Biotechnol* **22**, 985–992 (2004).
6. Ranish, J.A. *et al.* The study of macromolecular complexes by quantitative proteomics. *Nat.Genet.* **33**, 349–355 (2003).
7. Blagoev, B. *et al.* A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. *Nat. Biotechnol.* **21**, 315–318 (2003).
8. Andersen, J.S. *et al.* Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* **426**, 570–574 (2003).
9. Marko-Varga, G. *et al.* Discovery of biomarker candidates within disease by protein profiling: principles and concepts. *J.Proteome Res.* **4**, 1200–1212 (2005).
10. Old, W.M. *et al.* Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol.Cell Proteomics* **4**, 1487–1502 (2005).
11. Flory, M.R., Griffin, T.J., Martin, D. & Aebersold, R. Advances in quantitative proteomics using stable isotope tags. *Trends Biotechnol.* **20**, S23–29 (2002).
12. Kirkpatrick, D.S., Gerber, S.A. & Gygi, S.P. The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. *Methods* **35**, 265–273 (2005).
13. Kuster, B., Schirle, M., Mallick, P. & Aebersold, R. Innovation: Scoring proteomes with proteotypic peptide probes. *Nat. Rev. Mol. Cell Biol.* (2005).
14. Keller, A. *et al.* Experimental protein mixture for validating tandem mass spectral analysis. *Omics* **6**, 207–212 (2002).
15. Desiere, F. *et al.* Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* **6**, R9 (2005).
16. Craig, R., Cortens, J.P. & Beavis, R.C. Open source system for analyzing, validating, and storing protein identification data. *J.Proteome Res.* **3**, 1234–1242 (2004).
17. Marzolf, B. *et al.* SBEAMS-Microarray: database software supporting genomic expression analyses for systems biology. *BMC Bioinformatics* **7**, 286 (2006).
18. Jones, P. *et al.* PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.* **34**, D659–663 (2006).
19. Kawashima, S. & Kanehisa, M. AAindex: amino acid index database. *Nucleic Acids Res.* **28**, 374 (2000).
20. De Strooper, B. *et al.* Deficiency of presenilin-1 inhibits the normal cleavage of amyloid precursor protein. *Nature* **391**, 387–390 (1998).
21. Xing, Y. & Lee, C. Alternative splicing and RNA selection pressure–evolutionary consequences for eukaryotic genomes. *Nat. Rev. Genet.* **7**, 499–509 (2006).
22. Eisenberg, D., Marcotte, E.M., Xenarios, I. & Yeates, T.O. Protein function in the post-genomic era. *Nature* **405**, 823–826 (2000).
23. Rotzschke, O. *et al.* Exact prediction of a natural T cell epitope. *Eur. J. Immunol.* **21**, 2891–2894 (1991).
24. Schwartz, D. & Gygi, S.P. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat. Biotechnol.* **23**, 1391–1398 (2005).

25. Marques, J.T. *et al.* A structural basis for discriminating between self and nonself double-stranded RNAs in mammalian cells. *Nat. Biotechnol.* **24**, 559–565 (2006).
26. Schirle, M. *et al. Proceedings of the 52nd ASMS Conference on Mass Spectrometry and Allied Topics*, Nashville, Tennessee, May 23–27, 2004 (American Society for Mass Spectrometry, Santa Fe, NM 2004).
27. Tabb, D.L., Huang, Y., Wysocki, V.H. & Yates, J.R. 3rd Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **76**, 1243–1248 (2004).
28. Breci, L.A., Tabb, D.L., Yates, J.R., 3rd & Wysocki, V.H. Cleavage N-terminal to proline: analysis of a database of peptide tandem mass spectra. *Anal. Chem.* **75**, 1963–1971 (2003).
29. Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J. & Gygi, S.P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2**, 43–50 (2003).
30. Gavin, A.C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).