BIOINFORMATICS ORIGINAL PAPER
Vol. 27 no. 6 2011, pages 844–852
doi:10.1093/bioinformatics/btr027

Systems biology

Baking a mass-spectrometry data PIE with McMC and simulated annealing: predicting protein post-translational modifications from integrated top-down and bottom-up data

Stuart R. Jefferys[1] and Morgan C. Giddings[2,*]

[1]Curriculum in Genetics and Molecular Biology, University of Chapel Hill, Chapel Hill, NC and [2]Biomolecular Research Center, Boise State University, Boise, ID 83725, USA

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Post-translational modifications are vital to the function of proteins, but are hard to study, especially since several modified isoforms of a protein may be present simultaneously. Mass spectrometers are a great tool for investigating modified proteins, but the data they provide is often incomplete, ambiguous and difficult to interpret. Combining data from multiple experimental techniques—especially bottom-up and top-down mass spectrometry—provides complementary information. When integrated with background knowledge this allows a human expert to interpret what modifications are present and where on a protein they are located. However, the process is arduous and for high-throughput applications needs to be automated.

Results: This article explores a data integration methodology based on Markov chain Monte Carlo and simulated annealing. Our software, the Protein Inference Engine (the PIE) applies these algorithms using a modular approach, allowing multiple types of data to be considered simultaneously and for new data types to be added as needed. Even for complicated data representing multiple modifications and several isoforms, the PIE generates accurate modification predictions, including location. When applied to experimental data collected on the L7/L12 ribosomal protein the PIE was able to make predictions consistent with manual interpretation for several different L7/L12 isoforms using a combination of bottom-up data with experimentally identified intact masses.

Availability: Software, demo projects and source can be downloaded from http://pie.giddingslab.org/

Contact: morgan@giddingslab.org.

Supplementary information: Supplementary data are available at *Bioinformatics* online

Received on September 29, 2010; revised on December 22, 2010; accepted on January 13, 2011

1 INTRODUCTION

A cell needs to modify proteins with post-translational adducts for the same reason that an automobile needs gears: without them the operational range is severely limited. If cells relied on fixed proteins, biological reaction times would be on the order of minutes and only a limited set of functional groups would be available to build an active protein. Cells use ≥ 300 known post-translational modifications (PTMs) (Creasy and Cottrell, 2004) to allow dynamic 'functional shifting', providing for rapid and flexible responses to changing local conditions. These PTMs affect protein functions in many ways, including inducing conformation changes (Huse and Kuriyan, 2002), modifying protein–protein interactions (Seet *et al*., 2006), and affecting lifespan (Giglione *et al*., 2003; Hochstrasser, 1996). When PTM regulation breaks down, it is much like a broken transmission in a car—proteins become non-functional (Banerjee and Gerondakis, 2007; Giannopoulos, 2009; Minamoto *et al*., 2001; Shi, 2007). Teasing apart where PTMs occur, when they are used, and how they are modulated is of great interest in biological and biomedical research.

Studying PTMs requires examining proteins, but without the equivalent of polymerase chain reaction (PCR) amplification techniques, proteomics methods are significantly more difficult and indirect than genomic methods. Mass spectrometry (MS) is the star player in the proteomics field, and with the advent of new instrumentation such as orbitrap mass analyzers (Perry *et al*., 2008) combined with the continued maturation of established techniques such as electrospray ionization (Maxwell and Chen, 2008), it is an excellent tool for the investigation of PTMs. Current MS methods measure protein or peptide mass with such accuracy that it is possible to differentiate between two distinct modification states (isoforms) of a protein based only on a mass shift. For example, methylation adds +14.0 Da to the total mass of the protein. If it is possible to measure the intact mass of that protein within ~1 Da or less, then this methylation can be readily detected.

However, in practice, PTM analysis is not simple. Even putting aside sample preparations issues (Fang *et al*., 2010), analytical difficulties remain, including: (i) achieving sufficient accuracy to determine PTMs for large proteins; (ii) working with obstinate, insoluble proteins (Mirzaei and Regnier, 2006); (iii) decoding isobaric masses, when multiple combinations of PTMs give the same mass shift (e.g. both one acetylation and three methylations cause a 42.0 Da mass shift); and (iv) determining precise positioning of PTMs. Several approaches have been devised that address these challenges, including bottom-up, top-down and combined methods. We briefly describe these here; more detailed reviews are available (e.g. Bogdanov and Smith, 2005; Domon and Aebersold, 2006; Kelleher 2004; Yates *et al*., 2009).

Bottom-up MS uses a divide-and-conquer strategy that reduces proteins to constituent, short peptides that are more readily

[*]To whom correspondence should be addressed.

analyzed. The procedure begins by digesting a protein into peptides, generally with an enzyme such as trypsin that cleaves at predictable sites. These peptides are then separated based on distinct chromatographic and/or chemical properties and analyzed by MS to infer PTMs.

Bottom-up methodologies simplify the analysis of PTMs by producing smaller, more accurately measurable peptides and by decoupling modifications that fall on different peptides. However, although reduced in scope, the basic problem of determining modification locations and resolving isobaric masses remains. To overcome this, we can apply another round of divide-and-conquer, breaking a peptide into a set of constituent fragments. Given enough fragments, the amino acid sequence of the peptide can be reconstructed. This process, termed tandem MS (MS/MS), can be used to precisely locate a PTM on a given residue in the peptide, since it will cause a discernible shift at that site.

Though the combined bottom-up and MS/MS strategy is quite powerful, the decoupling of PTMs into separate peptides makes it difficult to resolve which specific combinations of PTMs were present on the original protein isoform. For example, if a protein has two phosphorylation sites, each on a separate peptide, it is not possible using a bottom-up strategy to tell the difference between a sample containing a mix of an unmodified and a doubly modified isoform variant, versus a sample containing a mix of two singly phosphorylated isoforms (each phosphorylated at one of the two sites). The digestion step used by the bottom-up approach converts either protein mix into identical peptide sets, each containing a modified and an unmodified version of each phosphorylation site.

Another challenge with the bottom-up MS/MS approach is the high frequency of missing peptides. Complete peptide coverage requires that each peptide has appropriate concentration, solubility and MS ionization. If a peptide is not observed, all information regarding its PTMs is lost.

Top-down MS starts by determining the intact mass of a protein and then applies a fragmentation step directly to the protein species being analyzed. This is also a divide-and-conquer approach, but it is applied rapidly and dynamically inside the mass spectrometer. For proteins whose solubility and size allows loading and detection, top-down MS has great potential as a faster and more complete way of analyzing PTMs on proteins, but the resulting spectra can be very complex, and the technology is still being developed (Kellie *et al.*, 2010).

Combined top-down and bottom-up (TDBU) MS uses both an intact protein mass and the more readily obtainable bottom-up MS/MS data. Together these data provide more complete and comprehensive information than either alone, and can be used for more comprehensive PTM inference.

Unfortunately, data integration for TDBU is hard. Experience in our laboratory trying to put together intact mass, peptide and MS/MS data for a study of ribosome modification (Ramkisoon,K. and Giddings,M.C., unpublished data) revealed the difficulty of manual integration. Due to incomplete bottom-up information and the multiple isobaric PTM configurations, there are many possible interpretations for any dataset. Nevertheless, with substantial human effort it is often possible to produce a clear good answer. Generally, these answers are not absolute, but an expert can generally provide a concordant argument for their choices. This kind of reasoning is difficult to turn into an effective and practical computer algorithm. Yet, the flood of new MS data does not allow for human experts to examine every output, necessitating computational tools that make the process significantly more efficient.

The advent of combined approaches such as TDBU is very recent, so not many supporting tools exist to aid in this integration problem. A number of programs exist that can determine PTMs from either bottom-up or top-down data alone. To interpret bottom-up data, most use an alignment algorithm to compare experimentally acquired peptide MS/MS spectra against a database of known spectra. Interpreting the difference allows identification of PTMs. Some examples are cross-correlation (Yates *et al.*, 1995), Mascot (Perkins *et al.*, 1999), TANDEM (Craig and Beavis, 2004) or InsPecT (Tanner *et al.*, 2005). A few algorithms or programs can do *de novo* interpretation of the MS/MS spectra, e.g. Spectral Dictionary (Kim *et al.*, 2009) or an Integer programming-based algorithm, PILOT_PTM (Baliban *et al.*, 2010). Fewer programs exist that interpret top-down (intact mass) data in isolation, as each has to handle the large number of isobaric masses produced from various combinations of PTMs. Proclame (Holmes and Giddings, 2004) is one, using fuzzy logic to constrain the search space. Additionally, there are a growing number of programs that make specific, focused PTM predictions based on the sequence data, without considering the MS data [e.g. SignalP (Bendtsen *et al.*, 2004) and NetPhos (Blom *et al.*, 1999)]. Unfortunately these programs work in almost total isolation, unaware of and unable to include results from each other, and they tend to be limited in scope or in the type of modifications they can consider.

Since the data integration problem is so great, recent effort has focused on solving this problem. While various scripts have been created to search constrained subsets of modification possibilities, few of those are yet in the published literature. Two that we are aware of are PTMSearchPlus (Kertesz *et al.*, 2009), wherein they perform a bounded search constrained by the statistics of the most likely numbers of modifications to occur, and a high-throughput data analysis pipeline from the Kelleher group (Durbin *et al.*, 2010). The pipeline combines bottom-up data with intact mass data of different types (isotope information from high resolution FTMS and change information from lower resolution ion-trap MS) to identify proteins and their simply modified isoforms. Though these approaches have moved the ball forward in this domain, the basic challenge of integrating arbitrary additional data sources in a way that has no hard bounds regarding the PTM scenarios considered is difficult to solve. A human expert uses many other data sources (expressed as prior knowledge) in determining a final solution, and has the capability to consider novel situations, outside of the bounds of a constrained search. With Protein Inference Engine (PIE) we have not created another algorithm to interpret mass spectra, but a way to combine the results from other programs that do so. The PIE takes a holistic approach to the data integration problem, treating each piece of data as a separate ingredient. Using Markov chain Monte Carlo (McMC) and SA allows integrating data together, not in the sense of assembling a puzzle from its pieces but in the sense of baking a pie, where the whole becomes a different entity, different from and greater than the sum of its parts. We are unaware of any other application of McMC or SA to integrate proteomics data.

2 METHODS

We set out to determine whether we could effectively solve this challenge using a directed stochastic search algorithm. Our goal was to develop a platform that could potentially accommodate any type of data or information that might assist in efficiently finding the best answer from this large search space. We focused on the initial problem of attempting to integrate intact mass measurements from top-down analysis, peptide data from bottom-up analysis, prior statistics about the likelihood of any given modification and prior human knowledge into a program that can rapidly determine and score the most likely PTM scenario.

Inspiration for a solution to the data integration problem comes from Plato's allegory of the cave (Bloom, 1991). As denizens of the cave, we cannot directly observe the world, but instead can only see the shadows of the true reality outside projected on a cave walls. In this analogy, a modified protein and its PTMs are the truth we are trying to uncover, but it is impossible to directly observe that truth. We can only see shadows—data revealed by the light of experiment, i.e. MS and related methods. From these shadows we can make inferences, guessing at the nature of the underlying, unobservable truth.

We can formalize this as a collection of data, D, for a protein of interest and a set of all possible guesses, G, about what the underlying 'truth' is—the protein's PTM configuration. Our goal is then to identify the guess $g \in G$ most consistent with the data. To evaluate candidates for the best guess, we can specify a scoring function $S(g|D)$ that assigns to each guess g a score based on the available data. In this formulation, the best guess is the one with the highest value and is the truth we seek:

$$\text{Best guess for 'truth'} = \text{argmax}_g[S(g|D)]$$

To apply this description of PTM inference and build a prediction engine for PTMs, we needed to define the solution set G, specify the scoring function S, and provide a method for finding argmax$_g$ for this D.

2.1 Defining the search space *G*

Unfortunately, the solution set G is rather large. If we allow for only 10 different modification types, a protein of just 100 residues has a googol (10^{100}) possible modification states, which is much, much larger than the age of the universe, in picoseconds (around 10^{30}, Bolte and Hogan, 2002). It is truly impossible to check each possible scenario to find the best one. However, by arranging the potential guesses of the set G into a space G where distances between solutions can be defined and nearby answers have similar scores, we can then use a heuristic method to search only likely places. Our search space is represented in Figure 1.

Our solution space G is defined so that guesses (PTM isoforms) that are close together are similar and hence have similar 'goodness'. This provides a rough continuity for the scoring function $S()$ with respect to G and creates a functional landscape over which we can hunt for the best scoring guess.

2.2 Finding argmax$_g$(S) by Markov chain Monte Carlo and SA

The ultimate goal is to find the truth T which is the modified protein variant underlying the data. The closest we can come to this is argmax$_g$ $S(g|D)$, so our goal is to efficiently seek that without loss of generality. Metropolis McMC (Metropolis *et al.*, 1953) is a heuristic method for sampling from the solution space G using a guided random walk. The walk is directed by a ratio computed for the scoring function S for two neighboring points: g, where we currently are and p, a neighboring solution selected randomly as a possible next step. If where we want to step to has a higher score [the ratio $R = S(p|D)/S(g|D)$ is ≥ 1], we always take this step to a better scoring guess, if not then we only sometimes take this step, with probability R.

Repeated, this walk results in generally climbing toward the highest scoring (best) points in the space, with the ability to occasionally traverse the valleys to avoid being permanently stuck at a local maxima. After walking around for a while we can stop and report the point we are on. This McMC process samples from the landscape as if it were a normalized probability distribution. If we repeat this process many times, we will sample the highest scoring point most often.

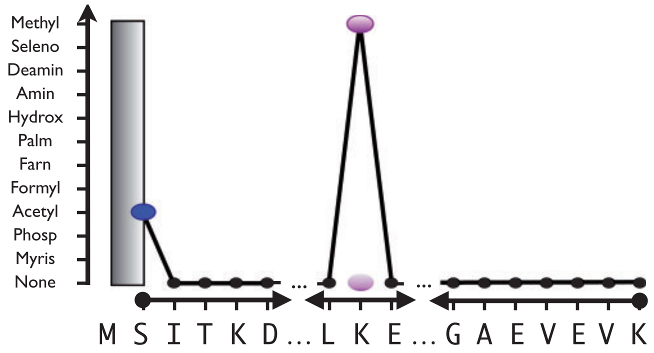

Fig. 1. Solution space. Every possible modified protein that PIE can propose as an answer can be visualized as a jagged line from left to right on the graph shown. The canonical protein sequence of the target being investigated is aligned along the x-axis, and the set of modifications searched for makes up the y-axis. The abbreviated protein sequence shown is from the L7/L12 ribosomal protein. The 11 modifications shown are those used in all experiments in this article. Only answers that can be constructed with modifications from this set can be proposed by PIE. Any modifications from the set can be used, with a linear drop in speed as the number increases. A different, unique line is drawn for each possible modification scenario, formed by connecting the points defined by each adduct modification on the protein. In the example shown, most adducts are 'none' (small points), but there is a N-terminal acetylation (on leftmost S) and an internal K-methylation. To allow for cleavages, the left and right ends of the line do not have to be a the first or the last AA of the protein. An N-terminal methionine truncation is shown by the line starting on the S and not the initial M (emphasized by a vertical gray block). This provides a simple way to visualize the multidimensional solution space where points are answers. Points that are close together in the solution space are here lines that are mostly overlapping and represent similar answers.

Unfortunately, the near infinite number of low scoring solutions outweighs the few high scoring ones, so that the best solution is reported infrequently. To guide the random walk more efficiently to the best answers, we use simulated annealing (SA; Kirkpatrick *et al.*, 1983). SA modifies the McMC walk by scaling down R using a coefficient that decreases gradually with each step, so that near the end of the run it approaches zero. This has the effect of gradually bounding the McMC walk, preventing it from crossing ever shallower valleys, until at the end it can only go uphill. McMC and SA have long been popular in the physical sciences, and have been successfully used to explore very difficult biological search spaces. Two examples are Mr Bayes (Huelsenbeck, 2001) and Rosetta Design (Kaufmann, 2010).

If run long enough, SA will always converge to the highest scoring answer, but we cannot determine in advance how long that will take. To address this, we run the algorithm repeatedly to sample from the space of solutions, providing an empirical distribution showing the frequency with which a given answer is obtained. The true maxima will be found more and more frequently with longer and longer searches. Specifically, we sample 10 times, increase the run length and repeat until we observe convergence.

Although the highest scoring answer is our best guess given the data and scoring models, there may be other guesses that score nearly as well. By sampling suboptimal answers, we can determine how consistently the algorithm and data support a single candidate above all others and reveal difference between situations where the data support several nearly equivalent solutions versus ones where data support only a single best answer. To investigate this empirical distribution of possible answers and their scores, we again profile answers by sampling over 100 runs at a step length estimated to converge ~20% of the time.

Table 1. Data-specific scoring modules

Term	Model	Data type	Model type
S_1	Intact mass	Experimental	$1/x$
S_2	Peptide mass	Experimental	$1/x$
S_3	MS/MS sequence	Experimental	a^n
S_4	Adduct frequency	Prior	$\prod_i f_i$
S_5	Adduct location	Prior	$\prod_i f_i$
S_6	Adduct count	Prior	$1/x$
S_7	N-cleavage	Prior	a^n
S_8	C-cleavage	Prior	a^n
S_9	Rules	Prior	$\prod_i f_i$

2.3 Specifying the scoring function *S()*

In order to specify a flexible scoring function compatible with McMC search, we require the following properties: it must be defined and non-negative for all possible guesses, it must model data such that better supported guesses have higher scores, and the ratio between the scores of any two guesses should reflect the relative support for those guesses given the data.

We considered that available data D would consist of some varying combination of up to k different data types including an intact mass measurement (d_1), a set of matched peptide masses (d_2), a set of MS/MS sequence data (d_3) and multiple prior data types. Defining a complete joint probability distribution $P(G|D)$ is not possible, so we make the assumption that each data type is independent. As with naive Bayesian classifiers, the error this introduces should at least partially cancel through the use of multiple data types (Zhang, 2004). This allows us to express $P(G|D)$ as the product of the individual probabilities for each prior data type

$$P_1(G|d_1)\cdot P_2(G|d_2)\cdot\cdots\cdot P_k$$

During its search, McMC only evaluates answers by the ratios of their probabilities, not their absolute probability of occurrence. This allows us to take a substantial shortcut by representing prior data types through a non-normalized scoring function. Expressed for a single guess $g \in G$ this is:

$$S(g|D)=S_1(g|d_1)\cdot S_2(g|d_2)\cdot\cdots\cdot S_k$$

The requirements for each factor in this scoring function are the same as those we first outlined, but applied to each type separately. We can add arbitrary new data types in the future by developing and adding new scoring terms that meet these requirements. The data types and associated scoring terms used in this article are summarized in Table 1. Three type of data models are used for scoring. The $1/(\Delta x)$ models have a maximum score when the guess matches numerical data exactly, decreasing as the difference increases. For example, the experimental intact mass is data, and its difference from the theoretically calculated intact mass for a guess is a measure of the quality of that guess with respect to the intact mass data. The a^n models have a maximum score when there are a minimum number of mismatches ($n=0$), decreasing as the number of mismatches increases. a is a constant <1. For example, the MS/MS model counts the number of amino acids and modifications that do not match provided MS/MS sequence information, and is a measure of the quality of the guess with respect to the MS/MS data. The $\prod f$ models use data that defines some frequency distribution ($f \leq 1$) of individual modification events then multiples the frequencies together to get a total score. For instance, phosphorylation of serines are about five times more common than phosphorylation of threonine (Lee *et al.*, 2006), and hence guesses that matching this distribution will score higher. This is only a general outline of the scoring methodology, additional details are provided in Supplementary Material.

2.4 Implementation

We implemented the PIE in Java 1.5 using 'best practices' development methods including incremental growth and unit testing. To fully evaluate, one intact mass variant takes about 150 runs of PIE, a total of about 20 min of computer time on a 2010 computer system. Data and scoring are modularized with each type of data evaluated by a separate scoring module. This provides flexibility to add new data types in the future through additional modules. All complex input data are read from simple table-like delimited text files; output is similarly presented. Rather than complicated command-line parameters, control information (along with some simple input data) is provided via a standard java properties file. The program has the ability to sample multiple times at a given step length. Scripts written in the '*R*' programming language are included to produce graphical views of the output similar to the result figures in this article. The program can be download from http://pie.giddingslab.org/ and a step-by-step tutorial walkthrough is available (Jefferys and Giddings, 2011).

3 RESULTS

3.1 Analysis of L7/L12 theoretical data

We first tested the PIE program on synthetic data based on a real protein we had studied in the lab: *Escherichia coli* ribosomal protein L7/L12. Manual interpretation of experimental data (Ramkisoon,K. and Giddings,M.C., unpublished data) suggested the presence of several isoforms, one of which had three modifications: an N-terminal methionine loss, an N-terminal acetylation (on 2S) and a lysine methylation of (82K). Using a theoretical lysC enzyme digest from PeptideCutter (Gasteiger *et al.*, 2005), we generated a complete, ideal set of experimental data matching this isoform, consisting of all bottom-up peptides, complete tandem MS/MS sequence data, and exact top-down mass. Several datasets with various levels of error were then produced from this ideal set by removing peptides and MS/MS sequence, and by adding error into the intact mass. An estimate for the intact mass accuracy is also needed, so we choose errors that are near or possibly larger than the estimated error to show how PIE performs with less than ideal data. Predictions are summarized in Table 2; complete results are available as Supplementary Material.

To correctly characterize modification isoforms from typical proteomic experiments requires obtaining enough data to determine what modifications are present and where they are located. For this target isoform, just two peptides (with sequence) serve to identify the location of the acetyl and methyl adducts, and a moderately accurate intact mass, within 0.5 Da of the actual value, provides evidence that the only other modification is a loss of methionine. At the target's intact mass of 12 220, 0.5 Da is about 40 ppm. Any greater intact mass error would support the addition of an amidation or deamidation modification (±1 Da). The program converged to the correct answer with a few minutes of runtime for all theoretical L7/L12 datasets where there was enough data to localize the modifications (sets 1, 2, 3). By using prior scoring modules, PIE was able to obtain consistent answers even when either the intact mass (7 and 8) or the peptide data (4, 5, 6, 9 and 10) did not contain enough information, i.e. when the intact mass error was large, or when MS/MS data or peptides are missing. This includes leaving out all peptide and MS/MS data for one or more modified peptides. In general, the lack of experimental localization information leads to multiple equal scoring answers different only in the position of modifications, but prior scoring modules help to order subsets by probability, rule out many unlikely answers (i.e. a phosphorylated arginine), and in some cases obtain the correct localizations (i.e. an N-terminal acetylation). For the remaining two datasets (11 and 12),

Table 2. L7/L12 predictions given varying theoretical MS datasets

Theoretical L7/L12 data	Imposed intact error/window (ppm)[a]	Peptide/MS-MS coverage[b]	Steps	Top answer(bold matches expected)[c]	Second answer (bold matches expected)[c]	Score ratio	Why?[d]
1. ideal	+ 0.5/1	**1–14** = 100/100%	50 k	**1M-x, 2S-Acet, 82K-Meth**	**1M-x**, **2S-Acet,** 76A**-Meth**	5.72	MS/MS
2. good	−23/20	**1**,3,**5**,**8**,**12**,14 = 50/25%	25 k	**1M-x, 2S-Acet, 82K-Meth**	**1M-x**, **2S-Acet**, 76A-**Meth**	5.72	MS/MS
3. min	+40/50	**1, 8** = 10/10%	25 k	**1M-x, 2S-Acet, 82K-Meth**	1M-Acet, **2S-Acet**, 16S-Acet, **82K-Meth**, 86K-Meth, 120V+121K-x	5.33	Intact
4. no tandem	+40/50	1, 8 = 10/0%	60 k	**1M-x, 2S-Acet, 82K-Meth**	**1M-x, 2S-Acet**, 76**A-Meth**	1.43	Mod AA
5. no acetyl	+40/50	**8** = 5/5%	60 k	**1M-x, 2S-Acet, 82K-Meth**	**1M-x, 82K-Meth**, K-Meth, K-Meth, K-Meth	1.16	Mod count
6. no methyl	+40/50	**1** = 5/5%	15 k	**1M-x, 2S-Acet, K-Meth**	**1M-x, 2S-Acet**, E**-Meth**	1.28	Mod AA
7. high intact	+ 75/100	**1, 8** = 10/10%	50 k	**1M-x, 2S-Acet, 82K-Meth**	**1M-x, 2S-Acet**, 76A**-Meth**	1.43	Mod AA
8. low intact	−90/100	**1, 8** = 10/10%	35 k	**1M-x, 2S-Acet, 82K-Meth**	**1M-x, 2S-Acet**, 76A**-Meth**	1.43	Mod AA
9. no mod	−23/20	3, **4**, **5**, **11**, **12**, 14 = 50/25%	25 k	**1M-x, 2S-Acet, K-Meth**	**1M-x, 2S-Acet, E-Meth**	1.28	Mod AA
10. intact	−23/20	0%	250 k	**1M-x, 2S-Acet, K-Meth**	**1M-x, 2S-Acet**, 39A-**Meth**	1.43	Mod AA
11. hi intact	+75/100	0%	350 k	**1M-x, 2S-Acet,** P-Oxid**,** A-Amid	**1M-x, 2S-Acet, K-Meth**, 65N-Deam	1.04	Mod type
12. low intact	−90/100	0%	50 k	**1M-x, 2S-Acet,** K-Meth, A-Amid	**1M-x, K-Acet, K-Meth**, A-Amid	1.91	Rules

[a] The actual error in the simulated experimental intact mass and the simulated experimenter's estimate for the error range.

[b] Modified peptides: 1 has 1M-x (N-terminal methionine loss) + 2S-Acet (n-terminal acetylation), 8 has 82K-Meth (lysine monomethylation). Bold peptides have MS/MS data.

[c] Modifications: #X, amino acid X, at position #; -x, terminal amino acid loss; Acet, acetylation; Met, methylation; Oxid, oxidation; Amid, amidation; Deam, deamidation. Answers in bold match expected results. Where modifications are given without numbers, multiple equal scoring answers were present with the given modification on the specified AA, but at different positions.

[d] 'Why?' gives the scoring module contributing the most to differentiating the 2 answers (full profiles in Supplementary Material).

the error-adjusted intact mass was off by >40 ppm and no peptide data were used. Here, the answers obtained by the PIE unsurprisingly do not match those expected for this target isoform, but instead are more consistent with the given data and are within 10 ppm of the intact mass provided.

All datasets were profiled to characterize the quality of the proposed answer. Full profiles are provided in Supplementary Material, but the second best answer and the ratio of the top two answer scores are given in Table 2. Score ratios are consistent with the ability of PIE to provide greater discrimination between answers when more data are provided. The highest ratio's are obtained for datasets 1, 2 and 3, which are the only ones containing all the minimum required information for complete characterization of the isoform. For each of these, the second best answer scores lower due to its contradiction of experimental data, as indicated in the 'why' column of Table 2. This column also shows that, for the remaining datasets, where not enough experimental information is available, PIE is using prior expectations to select the best answer, but this is accompanied by lower score ratios.

3.2 Analysis of H23C theoretical data

As an independent test of the PIE, we developed a synthetic dataset based on two theoretical isoforms of the human h23c histone protein. We chose this protein because biologically it is highly modified and presents a more complex target than L7/L12. Considering two different isoforms simultaneously allows testing how the PIE handles conflicting data. One isoform, H5, was generated with two methylations, two acetylations, one phosphorylation and an N-term met loss; the other, H7, has two additional phosphorylations (Fig. 2). From these scenarios, we generated four theoretical datasets: one set containing peptides, tandem sequence and intact mass consistent with H5; one set consistent with the additional phosphorylations present in H7; and the remaining two sets having combined bottom-up data consistent with a mix of the two isoforms, but using either the H5 or the H7 intact mass. Predictions are summarized in Table 3; complete results are available as Supplementary Material.

The program was run assuming each intact mass was in errors by +10 ppm at ±20 ppm error, and with 3/4 the protein covered by MS/MS peptide data (including all modifications). For the combined peptide datasets, two regions of the protein have both a correct and an incorrect peptide assigned to them, the incorrect peptides originating from a different isoform. The PIE was able to correctly identify all modifications and their positions for the H5 and H7 pure isoform data and for the H5 mixed case; for the H7 mixed case it correctly identified all modifications, misplacing just one of the phosphorylations with conflicting data. For the H5 mixed case, this means PIE correctly identified and localized all modifications, and was able to ignore peptide data indicating two additional phosphorylations that did not apply to H5 isoform. For the H7 mixed case, this means PIE is reaching the limit of its understanding of the data. To correctly place this modification, the PIE needs additional information or modified scoring: see Section 4 for more information.

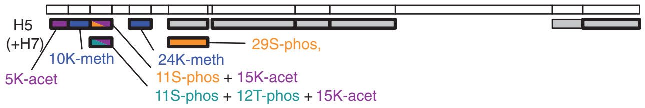

Fig. 2. H23C theoretical peptide data. The top white bar shows the theoretical peptide digest of H23C, blocks aligned below indicate theoretical peptides provided to the PIE. Those with thick edges indicate MS/MS sequence data were also provided. Gray boxes indicate unmodified peptides, other boxes are colored based on the modifications as shown. Some regions of the protein have no aligned peptide blocks, simulating missing peptides in the bottom-up data. The second row of peptides are labeled (+H7) to indicate that in the H7 bottom-up dataset they replace the matching unmodified peptides from the H5 bottom-up dataset. Both modified and unmodified peptides are present in the combined bottom-up datasets.

3.3 Analysis of L7/L12 ribosomal extracts

After validating the PIE on theoretical data (3.1 and 3.2, above), we applied the PIE to L7/L12 ribosomal extracts collected during an investigation of the role and extent of ribosomal PTMs in *E.coli* K-12 (Ramkisoon,K. and Giddings,M.C., unpublished data). L7/L12 is particularly complicated and was chosen because there are multiple isoforms simultaneously present in the sample, testing the PIE's ability to handle heterogeneity.

Table 3. H23C predictions with multiple theoretical isoforms

Theoretical H23C data	Imposed intact error/window (ppm)[a]	Peptide set[b]	Steps	Top answer (bold matches expected)[c]	Second answer (BOLD matches expected)[c]	Score ratio	Why?[d]
H5	−10/20	H5 set	250 k	**1M-x, 5K-Acet, 10K-Meth, 11S-Phos, 15K-Acet, 24K-Meth**	**1M-x,** 2A-**Acet**, **9K-Meth, 10S-Phos, 14K-Acet, 23K-Meth**	2.08	Peptide
H7	−10/20	H7 set	375 k	**1M-x, 5K-Acet, 10K-Meth, 11S-Phos, [12T-Phos], 15K-Acet, 24K-Meth, [29S-Phos]**	**1M-x, 5K-Acet, 10K-Meth, 11S-Phos, *12T-Phos*, 15K-Acet, 24K-Meth,** *32T-Phos*	4.35	Peptide
H5both	−10/20	H5 plus H7	75 k	**1M-x, 5K-Acet, 10K-Meth, 11S-Phos, 15K-Acet, 24K-Meth**	**1M-x,** 2A-**Acet**, **9K-Meth, 10S-Phos, 14K-Acet, 23K-Meth**	2.08	Peptide
H7both	−10/20	H5 plus H7	300 k	**1M-x, 5K-Acet, 10K-Meth, 11S-Phos, 15K-Acet, 24K-Meth**, **[29S-Phos]**, [S-**Phos]**	**5K-Acet, 10K-Meth, 11S-Phos, 15K-Acet, *K-Meth,** *K-Meth, *K-Meth	1.89	Intact

[a]The actual error in the simulated experimental intact mass and the simulated experimenter's estimate for the error range.
[b]Peptide sets are described in Section 3.
[c]Modifications: #X, amino acid X, at position #; ()X, amino acid X, no unique position; -x, terminal amino acid loss; Acet, acetylation; Met, methylation; Phos, phosphorylation. Where modifications are given without numbers, multiple equal scoring answers were present with the given modification on the specified AA, but at different positions Answers in bold match expected results, those in square brackets match modifications present only in the H7.
[d]'Why?' gives the scoring module contributing the most to differentiating the two answers (full profiles in Supplementary Material).

Table 4. L7/L12 predictions using integrated top-down and bottom-up experimental data

ID	Intact mass	Intact error window (ppm)	Manual interpretation[a]	Steps	Top answer (bold consistent with manual analysis)[b]	Intact mass delta (ppm)[c]	Second answer (bold consistent with manual analysis) [b]	Score ratio	Why?[d]
220-2H	12220.3	50	1M-x, 2S-Acet, 82K-Meth	50 k	**1M-x, 2S-Acet, 82K-Meth**	−17.3	**1M-x,** K**-Acet, 82K-Meth**	1.78	Rule
220-1H	12220.1	50	1M-x, 2S-Acet, 82K-Meth	50 k	**1M-x, 2S-Acet, 82K-Meth**	−0.9	**1M-x,** K**-Acet, 82K-Meth**	1.78	Rule
207-1L	12206.9	150	1M-x, K-Meth (x2), 82K-Meth	75 k	**1M-x, K-Meth (x2), 82K-Meth**	−65	**1M-x, K-Meth, 82K-Meth, 101K-Meth**	1.12	Peptide
207-2H	12206.8	50	1M-x, K-Meth (x2), 82K-Meth	75 k	**1M-x, K-Meth (x2), 82K-Meth**	−57	**1M-x, K-Meth, 82K-Meth, 101K-Meth**	1.12	Peptide
207-1H	12206.5	50	1M-x, K-Meth (x2), 82K-Meth	75 k	**1M-x, K-Meth (x2), 82K-Meth**	−32	**1M-x, K-Meth, 82K-Meth, 101K-Meth**	1.12	Peptide
206-0H	12206.1	50	1M-x, K-Meth (x2), 82K-Meth	100 k	**1M-x, K-Meth (x2), 82K-Meth**	+0.5	**1M-x, K-Meth, 82K-Meth, 101K-Meth**	1.12	Peptide
205-0L	12205.5	150	1M-x, K-Meth (x2), 82K-Meth	50 k	**1M-x, K-Meth (x2), 82K-Meth**	+50	**1M-x, K-Meth, 82K-Meth, 101K-Meth**	1.12	Peptide
175-1M	12174.5	100	1M-x, 82K-Meth	25 k	**1M-x, 82K-Meth**	+292	**1M-x**, 55F-Amid, **82K-Met**h, 115-Myr, 118V-Meth, 120V+121K-x	8.96	Peptide
163-1M	12162.9	100	1M-x	10 k	**1M-x**, 82K-Meth	+1246	1M-x, 2S-x, 81K-Meth, 11E-Acet, 115A-Acet, 116E-Acet	5.56	MS/MA

[a]Manual interpretation taken from investigation of modification isoforms of all ribosomal proteins (Ramkisoon,K. and Giddings,M.C., unpublished data).
[b]Modifications: #X, amino acid X, at position #; -x, terminal amino acid loss; Acet, acetylation; Met, methylation; Myr, myrystolation; Amid, amidation. Where modifications are given without numbers, multiple equal scoring answers were present with the given modification on the specified AA, but at different positions. Answers in bold match manual interpretation. For the 205, 206 and 207 isoforms, PIE's answer caused a partial revision of the potential manual results.
[c]Intact mass difference gives the difference in theoretical intact mass of the guess relative to the experimental intact mass of the isoform.
[d]'Why?'gives the scoring module contributing the most to differentiating the two answers (full profiles in Supplementary Material).

Top-down (intact mass) measurements were collected from several ribosomal extracts analyzed on two different mass spectrometers: a Bruker BioTOF II time-of-flight MS and a Fourier transform ion cyclotron resonance (FTICR) MS. Mass resolution for the BioTOF typically runs around 20 ppm and for the FTICR around 1 ppm.

A total of nine intact masses were selected from the MS data as corresponding to isoforms of L7/L12. The intact scoring model requires some estimate for the accuracy of these masses, although a precise estimate is not needed. We could simply have used the expected accuracy for the analyzing instruments, but the presence of internal standard analogs provides the opportunity for a second estimate. We calculated the mass error for all other apparently unmodified ribosomal proteins identified in the extract. Misidentification of one or more protein as unmodified is possible as these are not true internal standards, but this only makes our error estimate more conservative. The intact error windows used in Table 4 are those that would contain most data points, excluding outliers.

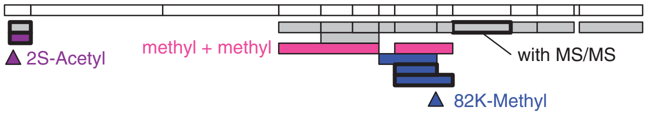

Fig. 3. *Escherichia coli* L7L12 peptide data. As in Figure 2, the top white bar shows the theoretical peptide digest, blocks aligned below indicate theoretical peptides. Those with thick edges indicate that MS/MS sequence data were provided. Gray boxes indicate unmodified peptides; other boxes are colored based on the modifications as shown. All peptides were provided to PIE, peptides are on separate lines only to show the overlapping and contradictory nature of the data.

Corresponding bottom-up peptide data for L7/L12 were obtained from *E.coli* K-12 ribosomal extracts by digestion with trypsin and analysis on a QSTAR MS/MS Quadrupole time-of-flight. Eighteen unique peptides were identified by precursor masses including six with adduct modifications. MS/MS sequence was obtained for six peptides, including three of those with modifications (Fig. 3).

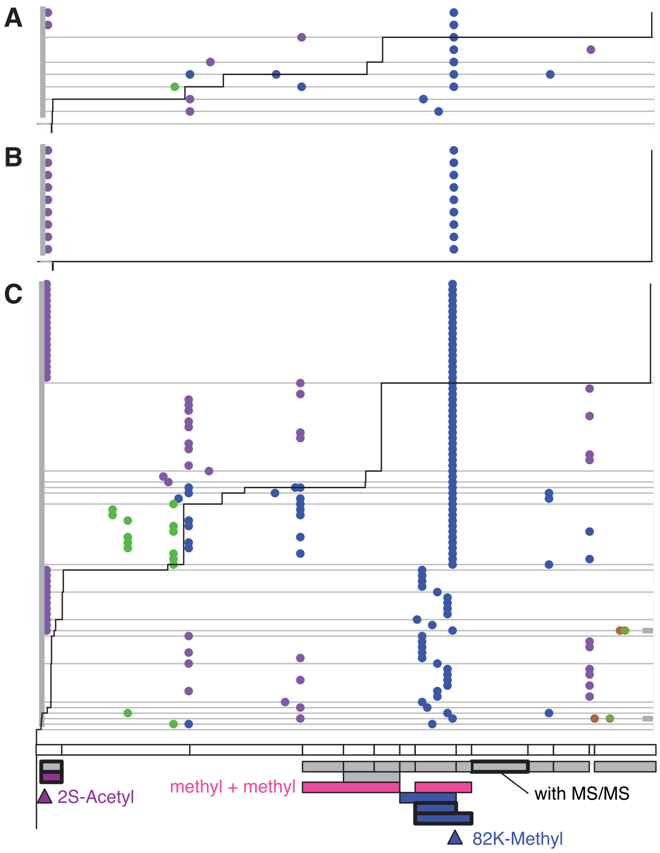

Fig. 4. Convergence and profile sampling, L7/L12 220-1H. Each row represents a different guess proposed by PIE, with guesses in each figure ordered by score from top (hi) to bottom (lo). The peptide dataset is reproduced to scale at the bottom of the figure. Colored dots represent modifications—methyl (blue), acetyl (purple), formyl (green), hydroxyl (brown) and myristoyl (olive)—each aligned at their proposed positions. Gray blocks at the left and right indicate N- or C-terminal truncations. Guesses without horizontal gray lines indicate they have the exact same score as the guess above, generating blocks of identically scoring guesses. The jagged black line running roughly diagonally through the graph indicates the score for each row, like a bar graph, except turned on its side with maximum value to the right. Scanning the images by sliding a straight edge top to bottom or left to right provides an 'animated' display that helps interpret modification alignments. (**A**) 10 samples taken after 50 000 steps. Although the highest score (guess) has been found twice, this is not clear evidence for convergence, so a longer length run set is necessary. (**B**) Ten samples taken after 250 000 steps. The same highest score (guess) is found here as was found in the 50 000 step set, and it is found many more times. This indicates convergence has probably been reached, and this is likely the best answer given the data. (**C**) Profile of 100 samples. To look for alternate interpretations for the data, a profile of guesses is sampled at a step length providing 20% of the answers as the best scoring candidate, and 80% as other suboptimal guesses. Estimated from the convergence plots, (A) shows this to be 50 000 steps. By examining these nearby answers it is possible to see not just how strongly a guess is supported, but how strongly different features of a guess are supported. The best guess with its specifically localized acetylation and methylation is approximately twice as good as the next bess guess (black scoring bar drops to half between them). All high scoring guesses have a predicted N-terminal methionine loss and an 82K methylation, indicating no consistent predictions could be made that did not include these features. Most suboptimal guesses have an acetylation. In the best scoring guess its localization is on the new N-terminus, 2S. If not there, the second best scoring guess places it on a lysine aligned with a missing peptide (see Supplementary Material for specifics). Here, the effect of the prior data module implementing the AA preference of modifications can be seen, as lysine is the most commonly acetylated modification, as well as the effect of the peptide data model, since peptides provide information that any covered lysine is probably not modified. Unfortunately, no other information is available to distinguish between these three positions, so there are three different equally likely predictions made as the second best guess. Other less likely answers include different positions for the acetyl modification, a tetra-methyl species, and a dimethyl + formyl species. Two rare, bad guesses suggest methyl + acetyl + hydroxyl + myristoyl, along with a cleavage of 1 N-term and 2 C-term amino acids, but even these rare guesses are in agreement with the intact mass data. They are approximately isobaric to the best guess, but score much lower due to multiple conflicts with peptide, MS/MS and prior data modules.

The PIE was applied to each of the nine intact mass targets, scoring the combined intact mass and bottom-up experimental data along with all available prior data models using default parameters. Figure 4A and B show two of the run sets used to identify convergence for one of the targets. The PIE converged to a best prediction for each of the nine targets, describing three different isoforms. The results summarized in Table 4 were consistent with a manual interpretation for eight of the nine intact targets (complete results are available as Supplementary Material). Three separate isoforms were identified, 12 206, 12 175 and 12 220. These were each consistent with prior manual analysis, including localization of one to three modifications and an N-terminal methionine cleavage. One of the nine intact masses, putatively representing a fourth isoform at 12 163 Da, was incorrectly predicted to be mono-methylated; manual interpretation suggests this isoform is unmodified except for N-terminal methionine loss.

Each of the nine targets was profiled, and the second best guess along with its relative score ratio are also included in Table 4. Figure 4C shows the profile for one of the targets, 12 220-H1. For the tri-methylated prediction, there is not enough data to localize two of the methyl adduct, so there are many nearby answers differentiated only by placement of the modifications. When the nearby best answers have similar scores, the score ratio is near 1, indicating multiple answers are supported by the data.

4 DISCUSSION

To investigate the feasibility of McMC/SA-based data integration, we sought answers to questions such as: Could we avoid being lost in endless sea of equal-mass answers, particularly given only incomplete data? Can the system handle datasets that have dependencies, errors or contradictions? Because McMC and SA are computationally intensive, can we find answers in a reasonable time? Can we develop the approach into a practical and modular program that works well now, but can readily accommodate additional data types in the future?

Data integration in proteomics is difficult because data are usually incomplete and the solution space is ambiguous, especially when proteins have multiple modifications. Given the results obtained above, McMC and SA appear to be a useful way to approach this problem. Predictions for simply modified proteins are easy to obtain. For data representing a single protein isoform with only one or two modifications, interpretation is easy and many methods should work well. However, the situation becomes rapidly more complex with just a few isoforms, each with several modifications, especially given missing or contradictory data and differentially modified peptides. In these more complex cases, the combinatoric explosion of possible answers require algorithms—like SA—that scale well. These are also the cases where we need the most interpretive help. The PIE's surprising ability to obtain useful results from intact data alone (as in Table 2, dataset 10) encouraged us to try bottom-up data with peptides from several isoforms, relying on the idea that the intact mass would allow distinguishing the relevant peptides. The PIE did so well interpreting these complex mixtures—even coming up with an answer not previously considered during manual analysis—that we chose to focus this article on these more complicated datasets.

Having multiple interpretations for the same dataset makes mass spectrometer data difficult to analyze. For example, the data for the L7/L12 isoform 12 206 supports either tri-methylation or acetylation. Which is correct? Given the available mass spectrometer data, it is impossible for human or software to tell. PIE's tri-methyl prediction results from the multiple methylated peptides in the bottom-up data. Which prediction is correct is indeterminate, and it is our belief that both isoforms may be present. It is interesting that before using the PIE we had not previously considered tri-methylation since a manual interpretation of an acetylation was so 'obvious'. Also, without assuming completeness, there is no reason to believe all the intact masses have been found, and hence no reason to believe the di-methylated peptides (Fig. 3) have to belong to any of the intact masses. The only answer is to acquire more data, but how and what data? This is essentially a resource availability question, and depends on the specifics of the experiment and data. If, for example, ultra-high mass accuracy instrumentation was available (i.e. an FTICR-MS), it would be possible to resolve the ∼0.03 Da difference between a tri-methylation and an acetylation.

This lack of specificity in MS data makes it critical for automated analysis software to integrate a wide range of data. Without the flexibility to incorporate new techniques, any such software will quickly become obsolete, whereas software supporting a wide variety of data allows the experimenter to choose the fastest, cheapest and most accurate methods to produce data. Modularity is central to the PIE, allowing simultaneous integration of a variety of both MS and non-MS-based data. New modules can be added with relative ease, and it is easy to run 'what-if' experiments including or excluding data.

One important aspect of the way we used bottom-up data was to specify what not to look for. Modifications not detected in bottom-up data are, less likely to be present. We use this 'negative information' to help the PIE apply Occam's Razor and favor simpler answers. For example, without peptide data, the intact data model causes PIE to add extra amidations or deamidations (±1 Da modifications) to better match any deviation in intact mass >0.5 Da, even if such deviation is due to measurement error and not the presence of a modification (as in Table 2, dataset 12). Where bottom-up data was available and no such modifications were seen, this chance that such a modification would be proposed is reduced.

4.1 Model and data accuracy

Scoring is affected both by the accuracy and completeness of the available data as well as the accuracy of the model itself. Prior modules are based on information obtained from databases that suffer from ascertainment bias. The modifications present in the protein data bank (PDB), for example, are not independently sampled from all proteins, but are more like an applause meter, where popular or interesting proteins and their modifications are overrepresented. We chose simple priors to allow for fast calculation; given the bias in the underlying data, additional effort to refine them to provide highly accurate prior models seems unproductive.

The intact mass model is dependent entirely on the accuracy of the measurement. As shown in the results, if the intact mass has enough error, PIE will find a consistent answer that is also in error. Although PIE performed surprisingly well even with wide mass tolerances, narrower windows increase discriminating power.

The peptide model had difficulty dealing with isoform mixtures due to data conflicts inherent in a bottom-up shotgun approach. We tried several models to allow discrimination of multiple isoforms, but no simple model worked to our satisfaction as the peptide with the most variants can override other information. Compared with the consistent tri-methylated proposal obtained for the L7/L12 12,175 isoform, the methylation proposed for the 12 163 variant is not supported by the intact mass. Here, matching to the intact mass data by the intact mass model is outweighed by stronger matching to the peptide data as evaluated by the peptide model. This is due to the large number of methylated peptides, and might be avoided with an improved scoring model. Increasing accuracy in intact mass would also eventually reverse this, producing the expected manual answer. Additional data on the relative abundance of peptides could help identify the most prevalent isoform, but would decrease the ability to identify all others. The underlying McMC and SA algorithms can optimize continuous values as well as discrete ones, and it is possible that PIE could be extended to include 'guesses' with quantification estimates for an isoform. However, we have no immediate plans to do so.

PIE uses a score ratio derived from the answer profile in lieu of a formal error model (e.g. *P*-value). The answer profile samples directly from the empirical distribution, with the ratio of any two scores giving the ratio of their probabilities. It also represents how unique a given answer is. As the incomplete and ambiguous nature of MS data supports multiple similar answers, it is important to determine if other good answers are likely. It is not clear how to generate a more meaningful error model. Bootstrapping can have difficulty with extreme values (Kysely, 2009) and pure McMC sampling (without SA) is computationally expensive. Additionally, these or similar error models only provide probabilities or confidence intervals with respect to the model used. The modular data framework used by PIE is designed to allow change models easily.

5 CONCLUSION

The current version of the PIE is only the first step in creating a tool that can integrate MS data and predict PTMs, but already it shows great promise. Using SA allows the PIE to explore the unfathomably large solution space of all possible modifications of a protein, and find the consistent answers. It is surprisingly robust, capable of decomposing an intact mass into a likely combination of modifications, and with the addition of MS/MS data, even a complex mixture of overlapping and conflicting peptides from several isoforms can be used to obtain specific modifications localizations.

The PIE provides an integrated approach for combining TDBU MS data in the context of prior knowledge to automatically determine the PTMs associated with a protein. By starting with few assumptions about the answer needed and using a flexible, modular framework that lets the data provide the constraints, the PIE can be extended and improved as new or better data and data models become available.

ACKNOWLEDGEMENTS

The experimental L7/L12 data analyzed in this article were provided by Kevin Ramkissoon and Hsun-Cheng Su, with FTICR data from Robert Hettich at ORNL. We also thank Chris Maier for programming and project management help and Suzy Vasa and the UNC Bioinformatics and Computational Biology students for their support and good advice.

Funding: National Institutes of Health National Center for Research Resources (R01-RR020823 to M.C.G.).

Conflict of Interest: none declared.

REFERENCES

Baliban,R. *et. al*. (2010) A novel approach for untargeted post-translational modification identification using integer linear optimization and tandem mass spectrometry. *Mol. Cell Proteomics*, **9**, 764–779.

Banerjee,A. and Gerondakis,S. (2007) Coordinating TLR-activated signaling pathways in cells of the immune system. *Immunol. Cell Biol.*, **85**, 420–424.

Bendtsen,J.D. *et al*. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.

Blom,N. *et al*. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.*, **294**, 1351–1362.

Bloom,A.D. (1991) The Republic of Plato translated, with notes, and an interpretive essay. 2nd edn. *Basic Books*, New York.

Bogdanov,B. and Smith,R.D. (2005) Proteomics by FTICR mass spectrometry: top down and bottom up. *Mass Spectrom. Rev*., **24**, 168–200.

Bolte,M. and Hogan,C.J. (2002) Conflict over the age of the Universe. *Nature*, **376**, 399–402.

Craig,R. and Beavis,R.C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, **20**, 1466–1467.

Creasy,D.M. and Cottrell,J.S. (2004) Unimod: protein modifications for mass spectrometry. *Proteomics*, **4**, 1534–1536.

Domon,B. and Aebersold,R. (2006) Mass spectrometry and protein analysis. *Science*, **312**, 212–217.

Durbin,K.R. *et al*. (2010) Intact mass detection, interpretation, and visualization to automate Top-Down proteomics on a large scale. *Proteomics*, **10**, 3589–3597.

Fang,Y. *et al*. (2010) Quantitative analysis of proteome coverage and recovery rates for upstream fractionation methods in proteomics. *J. Proteome Res*., **9**, 1902–1912.

Gasteiger,E. *et al*. (2005) Protein identification and analysis tools on the ExPASy server. In Walker,J.M. (ed.) *The Proteomics Protocols Handbook*. Springer, Heidelberg, Germany, pp. 571–607.

Giannopoulos,P.N. *et al*. (2009) Phosphorylation of prion protein at serine 43 induces prion protein conformational change. *J. Neurosci.,* **29**, 8743–8751.

Giglione,C. *et al*. (2003) Control of protein life-span by N-terminal methionine excision. *EMBO J*., **22**, 13–23.

Hochstrasser,M. (1996) Ubiquitin-dependent protein degradation. *Ann. Rev. Gen.*, **30**, 405–439.

Holmes,M.R. and Giddings,M.C. (2004) Prediction of posttranslational modifications using intact-protein mass spectrometric data. *Anal. Chem*., **76**, 276–282.

Huse,M. and Kuriyan,J. (2002) The conformational plasticity of protein kinases. *Cell*, **109**, 275–282.

Huelsenbeck,J.P. *et al*. (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, **294**, 2310–2314.

Jefferys,S.R. and Giddings,M.C. (2011) Automated data integration and determination of posttranslational modifications with the protein inference engine. In Wu,C.H. and Chen,C. (eds) *Bioinformatics for Comparative Proteomics*. Springer, Heidelberg, Germany, Ch. 17, pp. 255–290.

Kaufmann,K.W. *et al*. (2010) Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry*, **49**, 2987–2998.

Kellie,J.F. *et al.* (2010) The emerging process of top down mass spectrometry for protein analysis: biomarkers, protein-therapeutics, and achieving high throughput. *Mol. BioSyst*., **6**, 1532–9.

Kim,S. *et al*. (2009) Spectral dictionaries: integrating de novo peptide sequencing with database search of tandem mass spectra. *Mol. Cell Proteomics*, **8**, 53–69.

Kelleher,N.L. (2004) Top-down proteomics. *Anal. Chem*., **76**, 196A–203A.

Kertesz,V. *et al*. (2009) PTMSearchPlus: software tool for automated protein identification and post-translational modification characterization by integrating accurate intact protein mass and bottom-up mass spectrometric data searches. *Anal. Chem*., **81**, 8387–8395.

Kirkpatrick,S. *et al*. (1983) Optimization by simulated annealing. *Science,* **220**, 671–680.

Kysely,J. (2009) Coverage probability of bootstrap confidence intervals in heavy-tailed frequency models, with application to precipitation data. *Theor. Appl. Climatol*., **101**, 345–361.

Lee,T.Y. *et al*. (2006) dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res*., **34**, D622–D627.

Maxwell,E.J. and Chen,D.D.Y. (2008) Twenty years of interface development for capillary electrophoresis-electrospray ionization-mass spectrometry. *Anal. Chim. Acta*., **627**, 25–33.

Metropolis,N. *et al*. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys*., **21**, 1087–1092.

Minamoto,T. *et al*. (2001) Distinct pattern of p53 phosphorylation in human tumors. *Oncogene*, **20**, 3341–3347.

Mirzaei,H. and Regnier,F. (2006) Enhancing electrospray ionization efficiency of peptides by derivatization. *Anal. Chem*., 78, 4175–4183.

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

Perry,R.H. *et al*. (2008) Orbitrap mass spectrometry: instrumentation, ion motion and applications. *Mass Spectrom. Rev*., **27**, 661–699.

Seet,B.T. *et al*. (2006) Reading protein modifications with interaction domains. *Nat. Rev. Mol. Cell Biol*., **7**, 472–483.

Shi,Y. (2007) Histone lysine demethylases: emerging roles in development, physiology and disease. *Nat. Rev. Genet*., **8**, 829–833.

Tanner,S. *et al*. (2005) InsPecT: Identification of Posttranslationally Modified Peptides from Tandem Mass Spectra. *Anal. Chem*., **77**, 4626–4639.

Yates,J.R.,3rd *et al*. (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem*., **67**, 1426–1436.

Yates,J.R.,3rd *et al*. (2009) Proteomics by mass spectrometry: approaches, advances, and applications. *Annu. Rev. Biomed. Eng*., **11**, 49–79.

Zhang,H. (2004) The optimality of naive Bayes. In Barr,V. and Markov,Z. (ed.) *Proceeding of 17th International FLAIRS Conference*, Florida AI Research Society, FL, USA, pp. 562–567.