Optimizing Analytical Depth and Cost Efficiency of IEF-LC/MS Proteomics

Ilona Kifer, Rui M. Branca, Amir Ben-Dor, Linhui Zhai, Ping Xu, Janne Lehtiö, and Zohar Yakhini

Abstract—IEF LC-MS/MS is an analytical method that incorporates a two-step sample separation prior to MS identification of proteins. When analyzing complex samples this preparatory separation allows for higher analytical depth and improved quantification accuracy of proteins. However, cost and analysis time are greatly increased as each analyzed IEF fraction is separately profiled using LC-MS/MS. We propose an approach that selects a subset of IEF fractions for LC-MS/MS analysis that is highly informative in the context of a group of proteins of interest. Specifically, our method allows a significant reduction in cost and instrument time as compared to the standard protocol of running all fractions, with little compromise to coverage. We develop algorithmics to optimize the selection of the IEF fractions on which to run LC-MS/MS. We translate the fraction optimization task to Minimum Set Cover, a well-studied NP-hard problem. We develop heuristic solutions and compare them in terms of effectiveness and running times. We provide examples to demonstrate advantages and limitations of each algorithmic approach. Finally, we test our methodology by applying it to experimental data obtained from IEF LC-MS/MS analysis of yeast and human samples. We demonstrate the benefit of this approach for analyzing complex samples with a focus on different protein sets of interest.

Index Terms—Iso-electric focusing, minimum set cover, greedy heuristics, analytical depth, cost-effectiveness, coverage

1 Introduction

The analysis of proteins in complex samples, termed analytical proteomics, is key to understanding cellular mechanisms as well as of industrial processes that exploit living systems, such as manufacturing protein therapeutics (1; 2; 3; 4).

The most widely used analytical proteomics technique is tandem mass-spectrometry (MS). MS is a sensitive technique used to detect, identify and quantify molecules based on analyte mass and charge (*m*/*z*) and enables comprehensive proteome studies with complex samples. Commonly, proteins are digested into peptides which are then analyzed on MS and the protein level information is inferred from unique peptides identified in the sample. Mass-spectrometry for proteomics is typically coupled with one or more separation techniques, allowing higher proteome analytical depth, enhanced quantitative accuracy and improved PTM characterization. Common separation techniques include:

1. *High performance liquid chromatography (HPLC)*—the most popular separation technique for measuring biological samples by MS or MS/MS (termed **LC-MS** or LC-MS/MS, respectively), as most biological samples are liquid and nonvolatile. LC columns have small diameters (e.g. <0.1 mm) and low flow rates (e.g., 200 nL/min), leading to efficient separation of molecular entities in the sample. A longer duration of an LC gradient in a given measurement generally leads to better performance in identification of the sample compounds. Thus, long LC gradients are used for measuring complex samples. Seamless LC/MS interfaces also greatly increase analysis throughput (5).
2. *Isoelectric focusing (IEF)*—a separation technique based on differences of the isoelectric point (pI) between molecules (typically peptides, in the case of analytical proteomics). The pI of a molecule is the pH at which its net charge is zero (electrically neutral). In IEF the molecules are separated using an electric field, migrating along a slowly increasing pH gradient until their overall charge is neutral. IEF is commonly performed using gel based devices or off-gel systems (6; 7; 8) that separate the analytes into fractions according to their pI driven migration. The standard protocol for achieving maximal analytical depth is to then analyze all fractions using LC-MS/MS. The entire process, in this case, is termed IEF-LC-MS/MS.

In this paper we develop an algorithmic approach that enables more efficient utilization of IEF-LC-MS/MS. Our approach seeks high analytical depth obtained at a cost much lower than that of running LC-MS/MS on all fractions, as in the standard protocol. It is based on algorithmically optimizing the selection of a subset of fractions that provide a complete coverage of a protein set of interest.

In many studies the biological question of interest hinges upon identification and quantification of a particular set of proteins rather than on the content of the entire mixture. This protein set can consist of members of a pathway, biological process or protein complex, or of a group of

- *I. Kifer, A. Ben-Dor, and Z. Yakhini are with the Agilent Laboratories, Tel Aviv, Israel.*
E-mail: {Ilona_kifer, amir_ben-dor, zohar_yakhini}@agilent.com.
- *R.M. Branca and J. Lehtiö are wiith the Karolinska Institute, Stockholm, Sweden. E-mail: {Rui.Mamede-Branca, Janne.Lehtio}@ki.se.*
- *L. Zhai and P. Xu are wiith the BPRC, Beijing Institute of Radiation Medicine, Beijing, P.R. China.*
E-mail: zhailinhui@163.com, Xuping_bprc@126.com.

Manuscript received 9 Mar. 2015; revised 19 June 2015; accepted 23 June 2015. Date of publication 6 July 2015; date of current version 22 Mar. 2017.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2015.2452901

biomarker candidates (possibly to be further validated). Despite substantial recent improvements in analytical proteomics, there are considerable difficulties to reproducibly detect and quantify low abundant proteins and PTMs within complex samples. Targeted proteomics using selected reaction monitoring (SRM) has been developed to address these needs (9). However, SRM assay development is time consuming and requires investment in specific assay components.

Alternatively, narrow-range IEF focusing on pre-selected pH ranges has demonstrated utility in capturing predictable sub-proteomes (10; 11; 12). In-silico calculated pI values can be used to predict the experimental fractions in which each peptide ends up (7). We propose to use predicted pI values to guide a cost effective technique to IEF-LC-MS/MS and thus enable a fast turnaround approach to targeted proteomics. Our algorithmic methodology seeks to retain high analytical depth at a cost much lower than that of analyzing all fractions, as in the standard protocol. It is based on optimizing the selection of a subset of fractions that provide a complete coverage of a protein set of interest. We translate the task of selecting fractions for LC-MS/MS analysis to the well-studied Minimum Set Cover (MSC) problem, formally defined in Section 2.2. MSC is NP-hard (13) and a greedy approach yields an $O(\log\ n)$ approximation ratio, n being the size of the universal set (14), which is also tight for the general case (15). Approximation approaches that exploit bounds on the scarcity of elements (number of sets in which they occur) were also proposed in literature.

To address optimal fraction selection we examine several heuristic approaches and compare their performance, considering both effectiveness (how close to optimal is the produced cover) and efficiency (running time in practice). The algorithms are evaluated using two experimentally measured datasets, one from yeast and one from a human cancer cell line. The performance of three workflows is compared – a long LC gradient (no IEF fractionation), standard all-fraction LC-MS/MS, and LC-MS/MS performed only on the subset of fractions selected by our algorithm (termed the *fraction-cover*). Focusing on several different protein sets of interest we evaluate the number of selected fractions predicted to cover each protein set, as well as the implied total cost and the actual obtained coverage, based on experimentally measured data of complex samples. Our streamlined approach obtains much higher analytical depth than that of the long gradient. The analytical depth is close to that obtained by the standard all-fraction LC-MS/MS but is significantly less expensive and requires considerably less instrument time. Our method can also be applicable to selecting IEF fractions for targeted proteomics (SRM/MRM).

We expect the method presented in this paper to make higher analytical depth for a set of interest proteins broadly affordable, for core facilities as well as for individual laboratories.

2 Methods

2.1 Experimental Methodology

Sample Preparation—Human A431 cancer cell line samples were acquired as described in (7). Yeast strain JMP024 samples were acquired and handled as described in (16).

IEF Pre-fractionation by Peptide Isoelectric Focusing—High resolution isoelectric focusing (HiRIEF) was used as described in (7). Briefly, peptide samples were dissolved in 225 μl rehydration solution containing 8 M urea, and applied to the gel bridge. For reswelling of the IPG strip, 1 percent IPG pharmalyte pH 2.5-5.0 (GE Healthcare) was used. 24 cm linear gradient IPG strips (pI 3.0-10.0, GE Healthcare) were incubated overnight. Samples were applied to the IPG strips by the gel bridge (pH 3.7) at the cathode end and run. After focusing, the peptides were passively eluted into 72 contiguous fractions with MilliQ water using an in-house constructed IPG extractor robotics (GE Healthcare Bio-Sciences AB, prototype).

LC-MS analysis—LC/MS protein extraction and digestion were performed as previously described in (7). Peptide samples were separated using an Agilent 1200 nano-LC system with a Zorbax 300SB-C18 trap column and a NTCC-360/100-5-153 (Nikkyo) analytical picofrit column. The gradient of mobile phases A (3 percent ACN, 0.1 percent FA) and B (95 percent ACN, 0.1 percent FA) ran from 6 to 40 percent B at a flow rate of 0.4 μl/min. A long gradient of 240 min was used for samples not pre-fractionated by IEF whereas IEF fractions were analyzed using a 45min gradient. The Q-Exactive was operated in a data dependent manner, selecting the top 5 precursors for fragmentation by HCD. The survey scan was performed at 70,000 resolution from 300-1700 m/z, using lock mass at m/z 445.120025, with a max injection time of 100 ms and target of 1 × 10e6 ions. For generation of HCD fragmentation spectra, a max ion injection time of 500 ms and AGC of 1 × 10e5 were used before fragmentation at 30 percent normalized collision energy, at 17,500 resolution. Precursors were isolated with a width of 2 m/z and put on the exclusion list for 60 s. Single and unassigned charge states were rejected from precursor selection. Proteome discoverer 1.3 with Sequest-percolator was used for protein identification. Precursor mass tolerance was set to 10 ppm and fragment mass tolerance to 0.02 Da. Oxidized methionine was set as dynamic modification, and carbamidomethylated cysteine as static modification. Spectra were matched to a yeast Ensembl (R64-1-1 with 6,610 target sequences) or human Ensembl database (GRCh37.63 with 76501 target sequences), and results were filtered to 1 percent FDR on PSM level.

2.2 Computational Methodology

Algorithm—Denote the set of proteins of interest residing within a complex sample by $\mathrm{P} = \{\mathrm{P}_1, \mathrm{P}_2, \ldots, \mathrm{P}_n\}$. In our setting, each protein expected to appear in the sample is cleaved into peptides using a selected protease such as trypsin. For each protein $\mathrm{P_i} \in \mathrm{P}$, we consider a peptide q as a representative of $\mathrm{P_i}$ (also termed proteotypical) if and only if:

a) *q* uniquely identifies $\mathrm{P_i}$—namely, no other protein in the collection of proteins expected to be in the mixture has a peptide, under the used protease, identical to *q*.

b) *q* does not contain a Methionine—as optional Methionine oxidation leads to ambiguity in peptide identification.

c) *q* does not have an ambiguous trypsin cleavage site—e.g., consecutive R or K residues. Such peptides are generally less reproducible between runs.

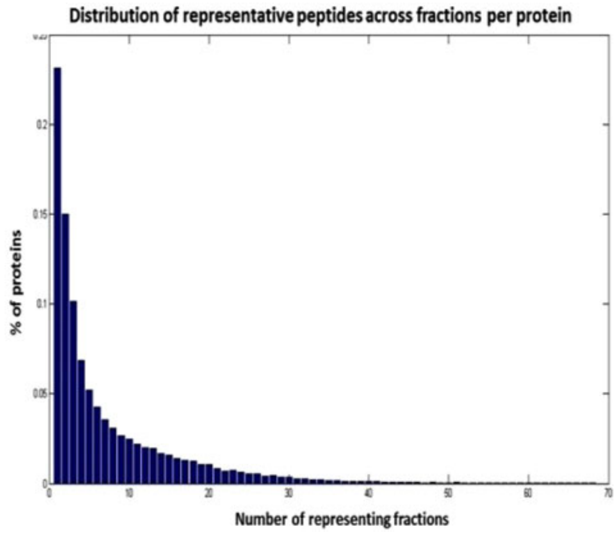

Fig. 1. Distribution of the number of covering fractions per protein on the human A431 cell line. Obtained from a 3-10 pH strip, with 72 fractions. Each fraction is 0.07 pH units wide.

For each peptide obtained by cleavage of the proteins in P, we use the predpI algorithm (7) to predict its iso-electric point (See Fig. 2 for evaluation of the pI prediction algorithm). An IEF fractionation is a partitioning of the range of possible pI values into disjoint intervals. Given an IEF fractionation, each representative peptide q of protein P_i is assigned to a specific fraction, F, based on its predicted pI value. In this case we say that fraction F covers protein P_i. Note that each protein P_i can be separately covered by several fractions, determined by the distribution of its representative peptides among fractions.

We now consider a setting in which there are K fractions (for example, K = 72 for a standard GE IEF, or K = 24 for an Agilent Offgel Fractionator), each fraction consisting of representative peptides of some subset of the proteins in P. We seek to find the minimal set of fractions that is required in order to cover all the proteins in P. If we consider each fraction to be a set and the proteins $\mathrm{P} = \{\mathrm{P}_1, \mathrm{P}_2, \ldots, \mathrm{P}_\mathrm{n}\}$ to be elements, it is straightforward to formulate our optimization task as an instance of the Min-Set-Cover problem (MSC). We now prove that our problem is NP-hard by reduction from MSC.

Reduction from MSC—In the set cover problem we are given a universe U, such that $|\mathrm{U}| = \mathrm{n}$, and sets $\mathrm{S} = \{\mathrm{S}_1, \mathrm{S}_2 \ldots, \mathrm{S}_\mathrm{K}\}$ s.t. $\mathrm{S}_\mathrm{i} \subseteq \mathrm{U}$ and $|\mathrm{S}| = \mathrm{K}$. We seek a collection $\mathrm{C} \ \subseteq \mathrm{S}$ that satisfies $\cup_{\mathrm{S}_\mathrm{i} \in \mathrm{C}} \mathrm{S}_\mathrm{i} = \mathrm{U}$ such that $|\mathrm{C}|$ is minimal. To reduce to our IEF optimization problem, we define each element $\mathrm{e}_\mathrm{j} \in \mathrm{U}$ to be a protein $\mathrm{P}_\mathrm{j} \in \mathrm{P}$. Each set S_i is translated to a fraction F_i, and the elements contained in S_i are reduced to the proteins represented in fraction F_i. More specifically—fraction F_i will have a representative peptide for a protein P_j iff $\mathrm{e}_\mathrm{j} \in \mathrm{S}_\mathrm{i}$. This reduction takes $O(nk)$ operations as we construct the set of representative peptides by scanning all K subsets of S, where $|\mathrm{S}_\mathrm{i}| \leq n$. Solving the above stated IEF optimization problem in polynomial time will lead to a polynomial time solution to the MSC problem. Therefore our IEF optimization problem is also NP-hard.

Heuristic Approaches—The most commonly used heuristic approach to MSC is the Greedy algorithm outlined below:

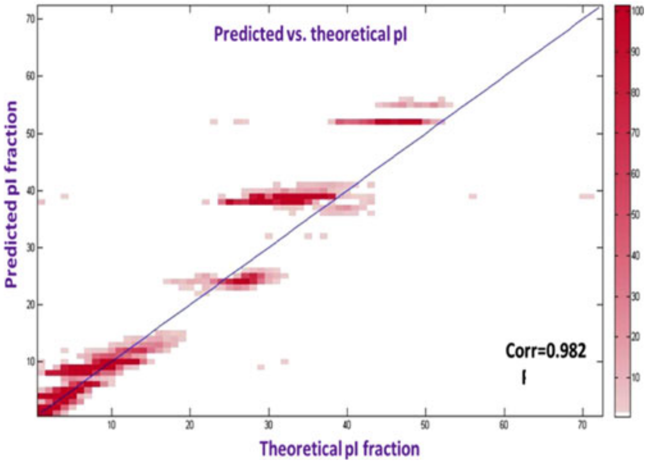

Fig. 2. A Heat map comparing the theoretical IEF fraction, prediced by predpI, with to the experimentally measures IEF fraction. Analysis was performed on the nr90 yeast proteome (>6,000 proteins). Trypsination and selection of representative peptides resulted in a dataset of 17,005 peptides, for which pI values were predicted (predpI algorithm (7)), and experimentally measured with IEF. Color intensity correlates with the concentration of peptides residing at each pixel.

Start by marking all elements in U as uncovered. Then:

a) Pick the set that covers the maximal number of uncovered elements

b) Mark all elements in the selected set as covered

Repeat (a) and (b) until all elements of U are covered.

Note that in this algorithm the next set to be added to the cover depends on the previously selected sets since they determine the identity and number of uncovered elements in each of the remaining sets. This approach therefore requires updating the set-to-element adjacency matrix after every iteration. Moreover, Greedy is a generic algorithm that disregards input-specific characteristics such as the distribution of number of covering fractions per protein. For our data, we observe this distribution to be skewed (Fig. 1), as most proteins are covered by very few fractions (or even just one). We thus contemplated that a suitable adaptation to the standard greedy algorithm could reduce the calculated cover size.

We call a set (fraction) *critical* if it is the only set that covers a particular element (protein). A simple adaptation of Greedy is the following: initially all critical sets are identified and added to the set cover. Subsequently, the greedy algorithm proceeds as standard. We call this approach the Critical-First Greedy algorithm (in short, CF-Greedy). The rationale behind this approach, also briefly mentioned in (17), is that since every set cover must add all critical sets at some point, we might as well add them at the beginning and possibly gain coverage of additional elements that may cost more to cover if addressed at a later stage of the algorithm. In Fig. 3A we show that while this adaptation seems very intuitive, it can also lead to the selection of a larger cover than that of Greedy. However, the simulations in Section 3.1 show that CF-Greedy is significantly more compatible with our data than the standard greedy approach.

A natural extension to the CF-Greedy algorithm, potentially more appropriate for IEF-type data distribution (as depicted in Fig. 1), adds sets to the set-cover in an order that is influenced by the scarcity of the elements they consist of. Formally:

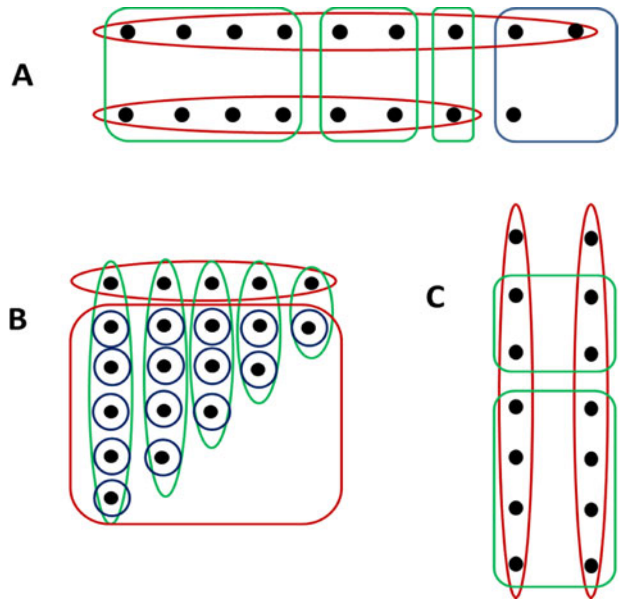

Fig. 3. Examples of performance differences between Greedy variants (A) Adaptation of the classical proof of Greedy approximation ratio: Greedy outperforms CF-Greedy by selecting three sets (two reds followed by blue), similar to OPT, while CF-greedy starts with the blue critical set, followed by *o(log n)* green sets. (B) SB-Greedy chooses $\sqrt{n}$ sets (green), Greedy chooses two sets (red), similar to OPT. (C) Greedy ends up with *o(log n)* green sets, in addition to the two red sets. CF-Greedy and SB-Greedy both choose two sets (red), similar to OPT.

1. Define the scarcity value of element $e_j \in U$ to be the number of sets in which it appears. For example, a critical set is now defined as a set that contains an element with scarcity value of 1. Partition the elements into classes according to their scarcity values.
2. Traverse the scarcity classes in increasing order of scarcity value. For each class with scarcity value r, consider the collection of unused sets $C_r \subseteq S$ which are the sets that have an uncovered element with a scarcity value of r.
3. Add the sets in C_r to the cover in a greedy manner—i.e., at each stage pick the set of C_r covering the largest number of uncovered elements of U. Continue until there are no more sets in C_r with uncovered elements.
4. Move on to the next scarcity class, until all elements in U are covered.

We call this variant Scarceness-Based Greedy (in short, SB-Greedy).

Theoretical effectiveness of the heuristic approaches—It has previously been shown that the greedy algorithm achieves an approximation ratio of $O(log\ n)$ (14). Moreover, it has been demonstrated that Greedy is essentially the best-possible polynomial time approximation algorithm for MSC, under plausible complexity assumptions (15; 18; 19). Hence, neither CF-Greedy nor SB-Greedy can achieve a better than $O(log\ n)$ approximation ratio. Fig. 3B presents a family of cases where SB-Greedy achieves an approximation ratio of $O(\sqrt{n})$, indicating that its general case theoretical bound is inferior to that of Greedy. A different problem setting, illustrated in Fig. 3C, demonstrates superiority of the scarceness based variants over Greedy, when selecting an optimally-sized set cover contrary to Greedy's $O(log\ n)$ sized selection. We conclude that neither of the presented variants is consistently superior to the other two.

In Section 3.1 we compare the practical performance of the three heuristic approaches on our IEF optimization problem for different values of n and K. We find that in practice, for our data and problem setting, all three algorithms converge to the optimal solution in the vast majority of cases. Moreover, for the relatively small problem sizes that were computationally feasible to test, we find that CF-Greedy is significantly closer to OPT than Greedy, and SB-greedy even more so.

Efficiency of the heuristic approaches—With some variation dependent on implementation details and data structures, it is clear that all heuristics presented above can be implemented in polynomial time with a low coefficient. More precisely:

- In the greedy algorithm, an addition of a fraction *f* to the fraction-cover requires a traversal over all elements in all remaining fractions to mark all of *f*'s elements as covered. This step is followed by a reassessment of the next fraction to be added to the cover, which is determined by the number of elements still uncovered in each fraction. The worst-case running time of Greedy is thus $O(K^2n + K^2) = O(K^2n)$.
- For CF-Greedy and SB-greedy the worst case running time is similar to Greedy since in certain cases all elements may appear in exactly X of the K fractions for some constant X. However, in practice running time is expected to be smaller since the process of reassessing the order of addition of fractions to the cover occurs separately within each scarcity class. This observation is further demonstrated in Fig. 5.

In Section 3 we demonstrate that the three proposed Greedy algorithms are extremely fast even for larger problem sizes, despite the exponential growth in search space of the exhaustive calculation.

3 Results

3.1 Efficiency Analysis for Heuristic Approaches

As our fraction selection problem is NP-hard, it very quickly becomes infeasible to calculate the optimal solution for increasing values of K (number of fractions). We thus wish to apply the heuristics described in Section 2.2 namely Greedy, CF-Greedy and SB-Greedy to this task. First we validate their relevance to our problem by comparing their performance to that of the exhaustive MSC calculation on our biological data. In this comparison, we evaluate the algorithms by two criteria—effectiveness (size of the calculated MSC) and efficiency (running time). Since calculation of the exact MSC is exponential in *K* due to enumeration over all possible 2^K subsets of S, we use relatively small values of *K* and n for the evaluation of heuristics with respect to the optimal set cover denoted by OPT.

Table 1 summarizes the effectiveness of the three heuristic approaches as compared to the optimal solution. For the purpose of statistical soundness, we performed 1,000 simulations for each pair of (K, n): first, we partition the pH region of [3], [4], [5], [6], [7], [8], [9], [10] into K equally spaced fractions. Subsequently, in every simulation the element universe U is constructed from *n* proteins randomly drawn from the human proteome. A fraction F is defined as

TABLE 1
Effectiveness Comparison for the Three Heuristic Approaches with Respect to the Exact MSC Solution, for Different Values of n (Number of Proteins of Interest) and K (Number of Fractions)

		%simulations with optimality ratio=1			
K	n	Greedy	CF-Greedy	SB-Greedy	Min (variants)
12	5	0.978	0.996	1.0	1 0
12	10	0.942	0.992	0.992	0.993
12	20	0.931	0.995	0.996	0.998
12	50	0.96	1.0	1.0	1.0
24	5	0.969	0.992	0.995	0.995
24	10	0.923	0.985	0.988	0.994
24	20	0.872	0.99	0.99	0.996
24	50	0.926	0.995	0.996	0.998
48	5	0.965	0.991	0.997	0.998
48	10	0.928	0.985	0.985	0.992
48	20	0.897	0.979	0.979	0.98
72	5	0.972	0.99	0.995	0.996

See description in text.

covering some protein $\mathrm{p} \in \mathrm{U}$ if the calculated pI of one of its representative peptides falls within F's pH range. The optimal fraction cover is then computed, as well as the fraction covers resulting from the three greedy variants. The values in Table 1 indicate the percentage of simulation instances for which the ratio of $\frac{\mathrm{MSC(OPT)}}{\mathrm{MSC(ALG)}} = 1$, which means the corresponding heuristic computed a solution of optimal size. We call this ratio the optimality ratio of the algorithm. The Min (Greedy-variants) column is a hybrid algorithm which selects the minimal solution out of the three heuristics. It is evident from Table 1 that, for the examined problem sizes all three approximation algorithms compute an optimal solution in the vast majority of cases. Also, the relevance of a scarcity-based approximation approach to our problem is demonstrated by the superior optimality ratio achieved by the two scarcity-based variants, where amongst the two SB-Greedy shows better performance.

Fig. 4 further demonstrates the superiority of the scarceness based approach by comparing the fraction cover size calculated by SB-Greedy to the ones calculated by Greedy and CF-Greedy by examining their ratio. The comparison is performed over a set of 3,000 simulations—1,000 for each value of $\mathbf{n} \in \{10, 60, 1{,}000\}$, where k (number of fractions) is fixed to 72. For this problem size the calculation of the optimal solution becomes computationally infeasible (or at least entirely impractical) and thus only the heuristic calculations are performed. Clearly, and as is demonstrated in Table 1, the fraction cover size computed by all three heuristics is frequently identical. However, some differences are illustrated in favor of the scarceness-based approaches. Specifically, for the smaller (and often practically more relevant) protein group sizes SB-Greedy obtains a smaller fraction cover than that of Greedy. The superiority over CF-Greedy is more subtle and is best demonstrated for the case of $\mathrm{n} = 10$.

Fig. 5 summarizes an efficiency analysis for the exhaustive and greedy variants that is based on the simulation data described above. For each pair of (K, n) we divide the 1,000 simulation instances into groups according to the size of their optimal set-cover. For every such group of simulations we plot the average running time of each algorithm over all relevant instances, in $\log(\mathrm{seconds})$, with error bars representing two standard deviations. Fig. 5 focuses on five entries from Table 1, specifically $(\mathrm{K} = 72, \mathrm{n} = 5)$ and $(\mathrm{K} = 24, \mathrm{n} = \{5, 10, 20, 50\})$, since these K values represent standard fractionations. Results from other (K, n) pairs show similar tendencies. The exponential growth in OPT time complexity for increasing set-cover sizes is most evident. In contrast, all three heuristic approaches show a very moderate, if any, increase in running time for larger

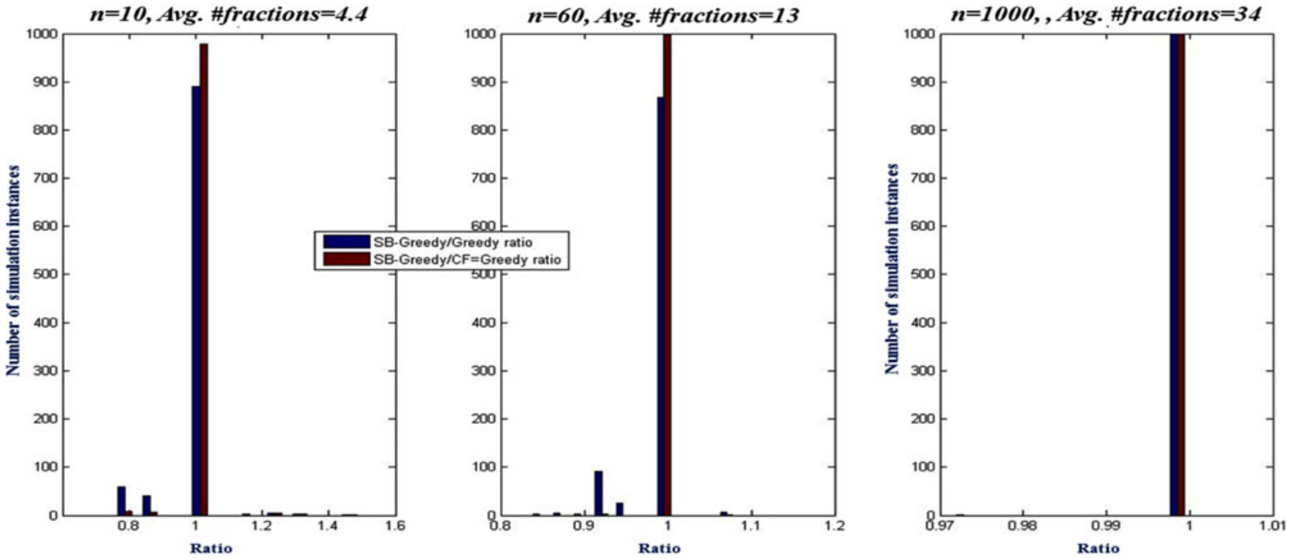

Fig. 4. Comparison of number of fractions calculated by the Scarceness-Based greedy approach to the fraction set size as calculated by Greedy and CF-Greedy. Calculations were performed on simulated data with number of fractions k set to 72, and the protein group size n varying between 10, 60, and 1,000. Clearly, SB-Greedy is able to calculate smaller cover sets in a larger percent of simulations. This is especially true for the smaller protein group sizes, in which the MSC size ratio of SB-greedy to the other algorithms is often below 1.

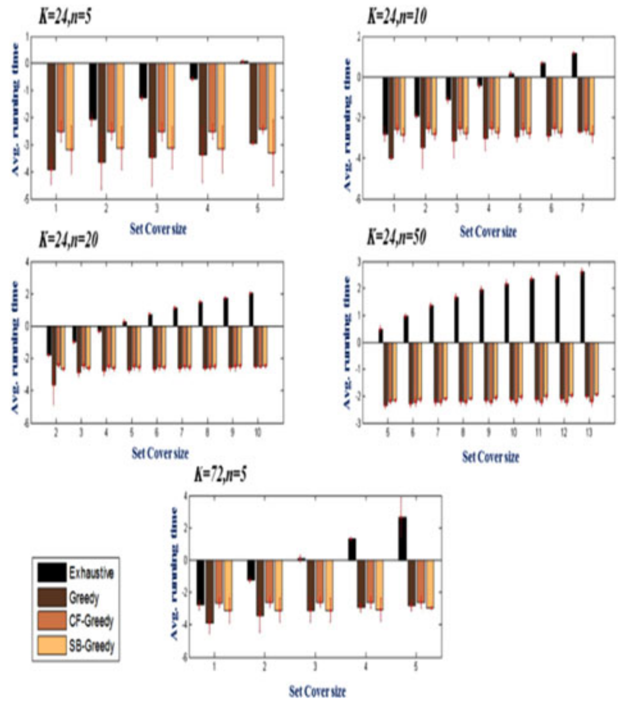

Fig. 5. Comparison of set-cover calculation time in *log(#seconds)* for the exhaustive approach and for the three heuristics. For each pair of (*K*,*n*), the experiment was repeated 1,000 times. Running time is averaged over all instances with a similar optimal fraction cover size.

set-cover sizes, and the calculation remains within the time range of $10^{-4} - 10^{-2}$ *sec*.

3.2 Application to Biological Sets of Interest

This study proposes an efficient approach for obtaining high LC-MS analytical depth for a subset of proteins of interest within a complex biological sample, while minimizing experimental cost and duration. We thus test our methodology on several well-studied groups of proteins, specifically:

- The human Anaphase Promoting Complex (APC), consisting of 11 proteins. APC is responsible for marking target cell cycle proteins for degradation.
- The human MAPK/ERK cascade, consisting of 62 proteins. These proteins form a signaling pathway from cell surface receptors to the nucleus DNA.
- The 945 human glycosylation-related genes, retrieved from the site of the Consortium for Functional Glycomics (CFG).
- The 130 yeast Ubiquitination-related genes described in (20).

For running calculations on the human and yeast biological groups we used the human and yeast Ensembl FASTA sequence files, respectively. The sequences were in-silico trypsinated and the peptides' pI calculated (7). In parallel, samples from the human A431 cell line and from yeast were fractionated into 72 fractions and LC-MS/MS was performed separately for each fraction. Since the number of fractions (sets) is relatively large (K = 72), it is infeasible to calculate the optimal set cover. We thus employ the above-described heuristics to search for a minimum-fraction-cover of the biological protein groups mentioned above within the given complex samples. As we have found CF-Greedy to obtain intermediate effectiveness, we focus the remainder of our analysis on Greedy and SB-Greedy.

Table 2 summarizes the MS2 protein identification rate for each of the four biological groups described above. We compare our minimum-fraction methodology to two LC-MS/MS based measurement procedures commonly applied for analyzing a focus set of proteins within a complex sample. The three compared protocols are as follows:

- Long gradient—No fractionation is used; a single slow-gradient LC-MS/MS experiment is performed for the duration of 5 hours. This procedure indicates the expected coverage for a single LC-MS/MS analysis of the entire complex sample.
- Full fraction—IEF is run on the entire sample, distributing peptides with a pI value in the range of [3], [4], [5], [6], [7], [8], [9], [10] into 72 fractions, LC-MS/MS is performed on each fraction separately.
- Fraction subset—fractionation is performed as before, but LC-MS/MS is run only on the fractions in the minimum-fraction-cover as calculated by either heuristic.

The substantial difference in coverage obtained by the long gradient and by the two fractionation-based approaches, as demonstrated in Table 1 suggests that the single-gradient procedure is not suitable for obtaining high analytical depth in complex samples. Importantly, the coverage obtained by the fraction-cover approach is remarkably close to the coverage obtained by running LC-MS/MS on all 72 fractions, despite a considerably smaller number of analyzed fractions. We thus propose that our methodology offers an efficient way to significantly reduce cost and instrument time, while maintaining high coverage of the proteins of interest.

To further validate the statistical significance of our results on biological groups of interest, we perform simulations that mimic the experimental setting of each of the three human protein groups described in Table 2. Contrary to previous

TABLE 2
Comparison of Analytical Depth Acquired by Several LC-MS/MS Based Techniques, for Different Protein Groups

	size	Coverage (# proteins identified in MS2)				Fraction-Cover Size	
		Lone Grad.	All Frac.	Greedy	SB-Greedy	Greedy	SB-Greedy
APC	11	O	10	6	6	5	5
MAPK	62	14	45	43	44	29	29
Glyco	945	164	408	400	397	41	39
Yeast Ub*	130	N/A	113	80	79	8	8

For each group we present the coverage, i.e., the number of proteins identified using LC-MS/MS, using each methodology, and also the number of fractions required to obtain this coverage, as computed by the relevant heuristic. (*) A long gradient was not performed on the yeast sample.

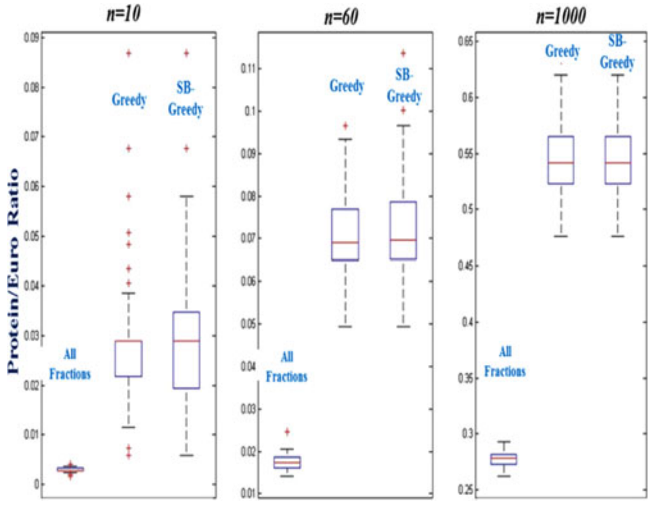

Fig. 6. The Proteins/Euro ratio calculated for different sizes of protein groups chosen at random (1,000 repeats). Based on experience, a single-fraction LC-MS/MS it considered to run for 69 min., and an LC-MS/MS hour to cost 30 EU. Note the differences in *Y*-axis scaling.

simulations, we leave the number of fractions constant ($K = 72$). For each protein group size $n \in \{10, 60, 1{,}000\}$ we perform 1,000 simulations. Each simulation consists of drawing n proteins at random from the human proteome and calculating their minimal fraction cover using Greedy and SB-Greedy. We now consider two criteria for evaluating simulation results based on our actual measurement data. One is the coverage obtained for the randomly drawn set of interest proteins. The other is the cost-effectiveness of the experiment. Fig. 7 depicts the distribution of coverage across simulations for each of the three protein group sizes. Most evident is the lack of analytical depth obtained by the long gradient experiment, in agreement with the results of Table 2. Also, as expected, the best analytical depth is demonstrated by the full-fractionation approach. Nevertheless and as also stands out from Table 2, limiting LC-MS

MS analysis to the fractions computed by our proposed methodology results in good coverage (for small protein groups) to excellent coverage (for larger protein groups), while allowing significant reduction inexperimental cost and complexity.

Notably the two heuristic variants obtain similar analytical depth, indicating that the actual choice of algorithm variant will not, on average, affect coverage. This is despite some differences demonstrated in Tables 1 and in 2. For example, SB-Greedy is shown to sometimes calculate a smaller fraction cover for the proteins in the biological group of interest. Thus, given a perfect agreement between the theoretical and experimental peptide fraction assignment, SB-greedy is expected to on occasion produce a less expensive solution than Greedy. In practice however, not all peptides considered in the calculation necessarily appear in their predicted fractions (see Section 3.3 for further discussion). Thus the choice of a larger fraction cover (as sometimes calculated by Greedy) may potentially lead to slightly higher coverage due to the unpredicted inclusion of proteo-typical peptides that are either unaccounted for by the algorithm, or theoretically reside in a different fraction. Fig. 6 compares the distribution across simulations of cost-effectiveness, which we quantify by the ratio $\frac{\#id{-}ed\ proteins}{cost\ in\ EURO}$, for the three protein group sizes. As

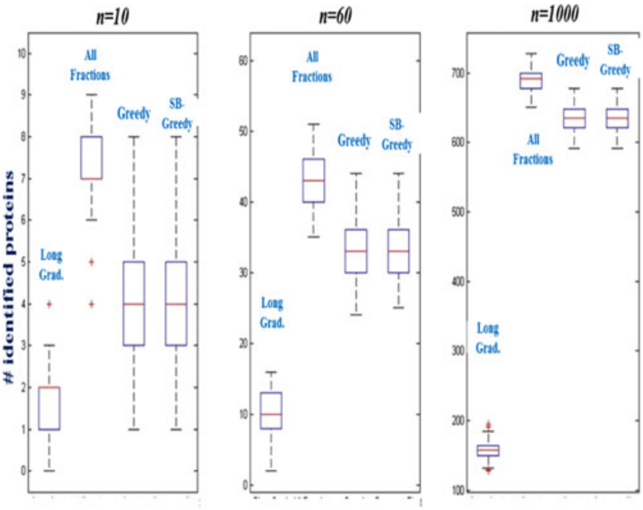

Fig. 7. Comparison of MS2 coverage distributions for three protein group sizes $n \in \{10, 60, 1, 000\}$ (1,000 random repeats each). In each box: central mark is the median, box edges are the 25th and 75th percentiles, whiskers extend to the most extreme data points not considered as outliers, and outliers are plotted individually. Note the differences in *Y*-axis scaling.

Fig. 7 clearly illustrates the incompatibility of the long-gradient measurement approach for achieving high analytical depth, we compare cost-effectiveness for the fractionation-based approaches only. Fig. 6 demonstrates the substantial improvement in cost-effectiveness gained from utilizing the minimum-fraction-cover approach rather than the all-fraction methodology. This is an expected outcome given the significantly reduced number of fractions submitted to LC-MS/MS analysis, together with the maintained high coverage.

Notably, some superiority of SB-Greedy over Greedy is evident, especially for the smaller group sizes. This reflects the fact that SB-Greedy's ability to produce a more compact solution that can sometimes reduce LC-MS/MS time and cost, more than compensates for the marginal loss in identified proteins. We thus propose that considerations related to working with SB-Greedy versus Greedy should be taken on a per-case basis, factoring in the importance of cost (biasing towards SB-Greedy) versus coverage (biasing towards Greedy).

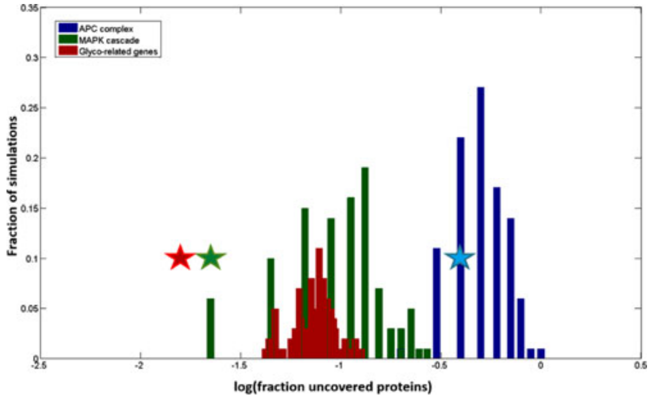

Fig. 8. Performance Evaluation against similar sized random fraction covers. Coverage was evaluated for the APC complex (11 proteins, blue), MAPK cascade (62 proteins, green) and glyco-related genes (945 proteins, red). The histograms depict the distribution of portion of uncovered proteins per focus set, in log space. Asterisks indicate this portion for the fraction cover selected by our algorithm.

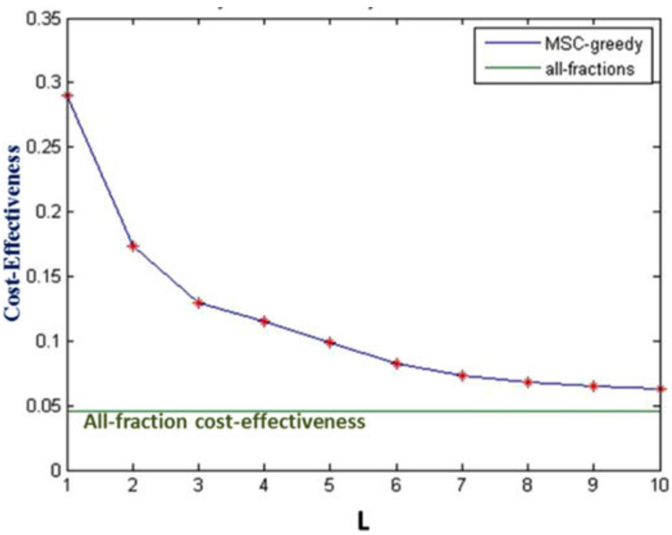

Fig. 9. Effect of increasing values of L on cost-effectiveness ratio (proteins/Euro). The green line indicates the cost-effectiveness obtained by performing LC-MS/MS on all 72 fractions.

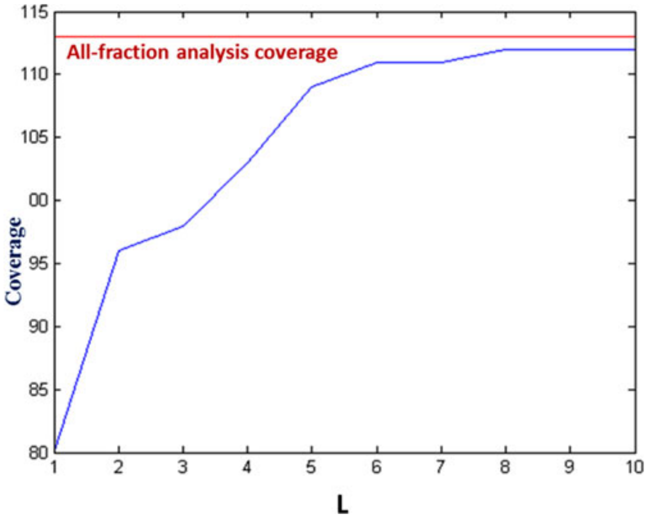

Fig. 10. Effect of increasing values of L on coverage size. The red line indicates maximal obtainable coverage (LC-MS/MS of all 72 fractions).

3.3 Statistical Validation of Significance

Thus far we have validated our method's robustness in the sense that it does not depend on a specific choice of a protein focus set (reproducibility). Herein we asses our method's dependence on the specific selection of a fraction cover, i.e., we test whether employing the minimum-fraction-cover approach as computed by our algorithm contributes to coverage and cost-effectiveness over an arbitrary selection of a fraction-cover of similar size. To do this we examine the protein focus sets as described above. Given the set C of fractions selected by the Minimum Fraction Cover algorithm, we perform 100 simulations in which $|C|$ fractions are chosen at random and the coverage attained for the corresponding focus set is measured. Fig. 8 presents the distribution of the portion of uncovered proteins for every simulation, in log space, for each of the three tested protein focus sets. The asterisks indicate the portion of uncovered proteins when fractions are chosen according to the Min-Fraction-Cover algorithm. In all three cases the advantage of our approach is evident, as the portion of interest proteins that remain uncovered with the Min-Fraction-Cover approach is either smaller than for all randomly chosen fraction sets (MAPK cascade and glyco-related proteins), or very close to that (for the APC complex only 12 percent of the simulations result in higher coverage). This result indicates that an arbitrary selection of a fraction cover of similar size will almost always result in poorer coverage and thus reduced cost-effectiveness. Moreover, we note that the naïve approach of arbitrary fraction selection can only be applied for a predefined parameter of fraction cover size—a factor which is in itself extremely important in maintaining the balance between high coverage and low cost.

3.4 L-Fraction-Cover

The results presented in Table 2 and Fig. 7 may bring one to notice the somewhat incomplete coverage obtained by the heuristically-calculated fraction-covers for the groups of proteins subjected to optimization. It is of importance to understand why these computationally selected sets of fractions do not include the entire group of interest proteins as identified by MS2 on the 72 fractions. A plausible explanation for this follows from the premises underlying our approach: First, that pI prediction is always precise. While prediction accuracy is relatively high (Fig. 2), some downward bias is observed in pI value prediction of the less acidic fractions. The existence of PTMs and other modifications, usually not addressed by pI prediction programs, can result in changes to pIs of peptides and are expected to be more prevalent for proteomes of high complexity such as human. In future, we plan to correct pI prediction by training on experimental data, likely improving our method's performance. The second assumption is that every representative peptide is selected for MS2 analysis. In actuality low peak ionization and insufficient separation lead to many peaks not being selected for MS2 analysis. In future we plan to employ more sophisticated means of selecting potentially proteo-typical peptides, to maximize the chance of experimental detection. It is also possible to use additional targeted proteomics approaches to further improve coverage, under the assumption that prediction and separation are sufficiently accurate.

To narrow the gap between the predicted coverage of the computationally selected fractions and the actual coverage obtained experimentally on the same set of fractions we have devised an L-fraction-cover approach, where each element is required to be covered at least L times in the selected fraction-cover. Assigning each interest protein to more than one representative peptide, and often to more than one fraction, decreases the probability of a protein not being identified by MS2. Hence with increasing L values the actual coverage is expected to increase, albeit the fraction-cover size is also expected to grow. We thus evaluate the gain in coverage with respect to the increase in experimental time and cost. Fig. 10 and Fig. 9 demonstrate the effect of increasing the value of L on coverage and cost-effectiveness, for the yeast Ubiquitin-related protein dataset. All fraction covers were calculated using SB-Greedy. Some increase in coverage is observed when requiring a protein to be represented by more than one peptide, the most significant leap being demonstrated for $L = 2$. However, Fig. 9 indicates that this gain in coverage leads to a significant decrease in cost-effectiveness. For example, the 12 percent increase in coverage demonstrated for L = 2 is associated with an almost two-fold decrease in the number of proteins identified per Euro.

These results indicate that when assigning equal importance to analytical depth and to cost-effectiveness, the single representative variant (L = 1) would be the best choice. However, when conducting an LC-MS/MS experiment requiring high emphasis on maximal analytical depth, a larger L value can be used. As a consequence more experimental resources will be required and cost is expected to increase, yet a practically full coverage can be obtained with significantly less resource investment than in an all-fraction LC-MS/MS analysis.

4 Conclusions

In this study we demonstrate that rational selection of IEF fractions can direct LC-MS analysis to aid in the detection of desired groups of proteins with the benefit of shorter analysis time and hence lower cost of the experiment. The algorithm presented herein functions as an experimental design tool for discovery proteomics, where the detection of predefined groups of proteins is important for biological interpretation and for other purposes such as monitoring a protein variant in an industrial process. Furthermore, this method can be used to select experimental IEF fractions for targeted proteomics analysis, hence potentiating antibody free pre-fractionation for SRM. These applications can help proteomics laboratories and core facilities in optimizing use of instrument time with maximal output and lower cost.

Here we have optimized our method for unmodified peptides, but in future we intend to further enhance capability to correctly handle post-translationally modified peptides as well.

In addition, we intend to further test out methodology in the context of quantification-based MS studies. That is, to use iTRAQ data to evaluate the agreement between per-protein ratios inferred by an all-fraction analysis to the corresponding ratios inferred from the fractions obtained by our approach. This can potentially have extensive applications in comparative studies of cancer biomarkers as well as for the study of other diseases.

Finally, we note that our algorithmic approach can be extended to any separation technique, other than IEF, that is based on a molecular characteristic that can be predicted in-silico.

Acknowledgments

This research has been funded, in part, by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 260600 (the GlycoHIT project), by the Swedish Research Council and Swedish Cancer Society, by the Chinese National Basic Research Programs (2011CB910600), and the China-EU International Collaboration Program (Grant No. 20111001). The authors thank Linhui Zhai and Dr. Heng Zhang for preparation and analysis of the yeast samples. They also thank Dr. Israel Steinfeld for fruitful discussions and insightful suggestions.

References

[1] R. Aebersold and M. Mann, “Mass spectrometry-based proteomics,” *Nature*, vol. 422, pp. 198–207, 2003.

[2] J. F. Kellie, J. C. Tran, J. Eun-Lee, D. R. Ahlf, H. M. Thomas, I. Ntai, A. D. Catherman, K. R. Durbin, L. Zamdborg, A. Vellaichamy, P. M. Thomas, and N L. Kelleher, “The emerging process of Top Down mass spectrometry for protein analysis: biomarkers, protein-therapeutics, and achieving high throughput,” *Mol Biosyst.*, vol. 6, pp. 1532–1539, 2010.

[3] C. H. Ahrens, E. Brunner, E. Qeli, K. Basler, and R. Aebersold, “Generating and navigating proteome maps using mass spectrometry,” *Nat. Rev. Mol. Cell Biol.*, vol. 11, pp. 789–801, 2010.

[4] A. F. Altelaar, J. Munoz, A. J. Heck, “Next-generation proteomics: towards an integrative view of proteome dynamics,” *Nat. Rev. Genet.*, vol. 14, pp. 35–48, 2013.

[5] D. Gershon, “Mass spectrometry: gaining mass appeal in proteomics,” *Nat. Methods*, vol. 2, pp. 465–472, 2005.

[6] Z. J. Sahab, Y. Suh, and Q. X. Sang, “Isoelectric point-based prefractionation of proteins from crude biological samples prior to two-dimensional gel electrophoresis,” *J. Proteome Res.*, vol. 4, 2005.

[7] P. Horth, C. A. Miller, T. Preckel, and C. Wenz, “Efficient fractionation and improved protein identification by peptide OFFGEL electrophoresis,” *Molecular Cellular Proteomics*, vol. 5, no. 10, pp. 1968–74, 2006,

[8] R. M. Branca, L. M. Orre, H. J. Johansson, V. Granholm, M. Huss, Å. Pérez-Bercoff, J. Forshed, L. Käll, and J. Lehtiö, “HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics,” *Nat. Methods*, vol. 11, pp. 59–62, 2014.

[9] B. J. Cargile, D. L. Talley, and J. L. Jr. Stephenson, “Immobilized pH gradients as a first dimension in shotgun proteomics and analysis of the accuracy of pI predictability of peptides,” *Electrophoresis*, vol. 25, pp. 936–4, 2004.

[10] H. Eriksson, J. Lengqvist, J. Hedlund, K. Uhlén, L. M. Orre, B. Bjellqvist, B. Persson, J. Lehtiö, and P. J. Jakobsson, “Quantitative membrane proteomics applying narrow range peptide isoelectric focusing for studies of small cell lung cancer resistance mechanisms,” *Proteomics*, vol. 8, pp. 3008–3018, 2008.

[11] J. Lengqvist, K. Uhlen, and J. Lehtio, “iTRAQ compatibility of peptide immobilized pH gradient isoelectric focusing,” *Proteomics*, vol. 7, pp. 1746–1752, 2007.

[12] P. Picotti and R. Aebersold, “Selected reaction monitoring—based proteomics: workflows, potential, pitfalls and future directions,” *Nat. Methods.*, vol. 9, pp. 555–66, 2012.

[13] R. M. Karp, “Reducibility among combinatorial problems,” *Complexity Comput. Comput.*. pp. 85–103, 1972.

[14] V. Chvatal, “A greedy heuristic for the set-covering problem,” *Math. Operations Res.*, vol. 4, pp. 233–235, 1979.

[15] U. Feige, “A threshold of ln n for approximating set cover,” *J. ACM*, vol. 45, pp. 634–652, 1998,

[16] P. Xu, D. M. Duong, N. T. Seyfried, D. Cheng, Y. Xie, J. Robert, J. Rush, M. Hochstrasser, D. Finley, and J. Peng, “Quantitative proteomics reveals the function of unconventional ubiquitin chains in proteasomal degradation,” *Cell*, vol. 137, pp. 133–145, 2009.

[17] S. S. Steven, *The Algorithm Design Manual*. New York, NY, USA: Springer, 2008, pp. 621–625.

[18] R. Raz and S. Safra, “A sub-constant error-probability low-degree test, and a sub-constant error-probability PCP characterization of NP,” *Annu. ACM Symp. Theory Comput.*, pp. 475–484, 1997.

[19] N. Alon, D. Moshkovitz, and S. Safra, “Algorithmic construction of sets for k-restrictions,” *ACM Trans. Alg.*, vol. 2, pp. 153–177, 2006.

[20] S. M. Nijman, M. P. Luna-Vargas, A. Velds, T. R. Brummelkamp, A. M Dirac, T. K. Sixma, and R. Bernards, “A genomic and functional inventory of deubiquitinating enzymes,” *Cell*, vol. 125, pp. 773–786, 2005.

Ilona Kifer is a data scientist at Microsoft R&D Israel, and was a staff scientist at Agilent Laboratories while doing much of the work described in this paper. She earned her BSc degree in computer science at the Hebrew University in Jerusalem, 2004 and her PhD degree in computer science at Tel-Aviv University, 2010. Dr. Kifer's early work focused on developing novel fast and efficient protein alignment and prediction methods and software, based on sequential and structural fingerprints. This approach is now broadly used by the protein structure bioinformatics community. In more recent years, Dr. Kifer worked on efficient approaches to proteo-glycomics measurements and on developing and implementing statistical and algorithmic methods for analysing high volume sequencing data. Applying these methods to large high throughput sequencing datasets, she played a key role supporting Agilent's improved nucleic acid synthesis technology, including new generations and variants. Dr. Kifer co-led the informatics aspects of several studies conducted under the framework of the EU GlycoHIT Project, developing and improving glyco-proteomics measurement approaches and their application in cancer research.

Rui Branca is an assistant professor at the Science for Life Laboratory and in the Dept. of Oncology Pathology, Karolinska Institute in Stockholm, Sweden. At the Clinical Proteomics Mass Spectrometry group led by Professor Janne Lehtiö, Dr. Rui Branca is involved mainly in methodology development, particularly in peptide separation techniques and LC-MS for Proteomics. He earned the degree of Licentia in biochemistry at the University of Porto, Portugal, in 2002, and his PhD degree in biophysics at the University of Szeged, Hungary, in 2008. Having worked on structure and function of individual metalloproteins during his PhD studies using a variety of spectroscopic methods, Dr. Rui Branca has since focused his work on the analysis of the Proteome as a whole through the use of Mass Spectrometry and in developing the peptide level separation technique of High Resolution Isoelectric Focusing (HiRIEF).

Amir Ben-Dor is a master scientist at Agilent Laboratories. He received his PhD degree with the direction of Professor B. Chor, at the Technion, Israel Institute of Technology, working on radiation hybrid mapping. Dr. Ben-Dor did a postdoctoral research at the University of Washington, Seattle, under the supervision of Prof. R. Karp., developing algorithmic approaches for the analysis of gene expression data. He joined Agilent Labs in 2000, doing independent research to support the efforts of developing array based hybridization assays and microarray applications such as gene expression profiling. Early work from Dr. Ben-Dor addressed combinatorial design of DNA probes and the analysis of high throughput molecular measurement using DNA microarrays. His work, joint with Dr Yakhini, on clustering, classification, and on bi-clustering won the RECOMB Test of Time Award in 2011, 2012 and 2014, respectively. He then participated in the development of probe design and aberration calling methods and tools for microarray based DNA copy number measurement (aCGH). Other fields of interest for Dr. Ben-Dor include analysis of multi-omics datasets, large dataset visualization methods, and, more recently, Digital pathology.

Linhui Zhai is a research assistant in the Xu Lab at Beijing Proteome Research Center. He received his MS degree in medicinal chemistry from Wuhan University in 2012. His major focus is high coverage proteomics and PTM study.

Ping Xu is the Beijing Sea Poly Talents Program' scholar and the director of the Department of Genomics and Proteomics at the Beijing Institute of Radiation Medicine. He is also the principal investigator in the Lab of Protein Posttranslational Modification at the Beijing Proteome Research Center. He earned his BSc degree in genetics at Wuhan University and his PhD degree in microbiology at Yunnan University, 2004. Right after that, he moved to the Department of Human Genetics at Emory University for his postdoctor training on proteomics. He is working on the biological function of protein post-translational modification, specifically the function of ubiquitin chains and their disregulation in the occurrence and development of liver disease by proteomics approaches. He is also leading the Chromosome Centric Proteome Project in China. His research involves state-of-art technologies for profiling thousands of proteins and posttranslational modifications by quantitative mass spectrometry and large scale data processing. Dr Xu co-chaired China-EU GlycoHIT International Collaboration Project under the 7th framework of the EU, developing and improving glyco-proteomics measurement approaches and their application in cancer research.

Janne Lehtiö has been a principle investigator at Dept. of Oncology and Pathology at Karolinska Institutet (KI), Sweden, since 2008, leading the cancer proteomics research group of 20 researchers. In 2004, Dr. Lehtiö was appointed as director of the Karolinska University Hospital's clinical proteomics facility and since 2010 he has been the platform director of the clinical proteomics mass spectrometry facility at the national research center Science for Life Laboratory. He received a faculty professor position in medical proteomics at Karolinska Institutet in 2015. He received his MS degree in biochemistry from the University of Helsinki, Finland, and his PhD degree in biotechnology in 2001 from the Royal Institute of Technology, Stockholm. After two years in the biotech industry in the USA and Denmark, he moved back to academic research and completed a postdoc period in cancer research in 2003, at Karolinska Institutet. In 2009, he was granted an associate professorship in proteomics at Karolinska Institutet, Sweden. Dr Lehtiö's major research interest is to improve human proteome analysis to obtain detailed molecular phenotype information; and to use this in-depth proteome information to personalize cancer therapy. His group has published a number of novel methods for proteome analysis and applications of them in leading journals such as PNAS, Molecular and Cellular Proteomics, Nature Methods, Nature Communications, and Nature Biotechnology. In a recent external evaluation at Karolinska Institutet, his research group was rated as Excellent. Dr. Lehtiö has attracted significant external grant funding as the PI from Swedish Research Council, European Union, Swedish Cancer Society, and Stockholm Cancer Society as well as pharma and technology Industry.

Zohar Yakhini is a master scientist at Agilent Laboratories and an adjunct faculty in the Computer Science Department, Technion, Haifa. Dr. Yakhini leads a group of computational biologists working on information aspects of genomics, proteomics, and glycomics. He earned his BSc degree in mathematics and computer science at the Hebrew University in Jerusalem and his PhD degree in mathematics at Stanford University, 1997. He has been working in computational biology and bioinformatics since after graduation, with a focus on statistical and algorithmic aspects of microarrays and other high throughput measurement technologies. Dr. Yakhini did data analysis work in several early gene expression studies and then co-developed probe design and data analysis methods, including software tools, for Agilent's aCGH microarray platform. His group developed several data analysis tools that are widely used by the genomics community, including differential expression and statistical enrichment analysis tools, such as GOrilla. Dr. Yakhini co-led the informatics aspects of several studies conducted under the framework of the EU GlycoHIT Project, developing and improving glyco-proteomics measurement approaches and their application in cancer research.