Optimal Decharging and Clustering of Charge Ladders Generated in ESI-MS

ARTICLE *in* JOURNAL OF PROTEOME RESEARCH · MARCH 2010

Impact Factor: 4.25 · DOI: 10.1021/pr100177k · Source: PubMed

4 AUTHORS, INCLUDING:

Chris Bielow
Max-Delbrück-Centrum für Molekulare Me…
16 PUBLICATIONS **177** CITATIONS
SEE PROFILE

Christian G Huber
University of Salzburg
189 PUBLICATIONS **5,012** CITATIONS
SEE PROFILE

Reinert Knut
Freie Universität Berlin
134 PUBLICATIONS **21,105** CITATIONS
SEE PROFILE

research articles Journal of proteome research

Optimal Decharging and Clustering of Charge Ladders Generated in ESI−MS

Chris Bielow,*[,†,‡] Silke Ruzek,[§] Christian G. Huber,[§] and Knut Reinert[†]

Institute of Computer Sciences, Free University Berlin, Berlin, International Max-Planck Research School, MPI for Molecular Genetics, Berlin, and Department of Molecular Biology, Division of Chemistry and Bioanalytics, University of Salzburg, Salzburg

Received February 27, 2010

In electrospray ionization mass spectrometry (ESI−MS), peptide and protein ions are usually observed in multiple charge states. Moreover, adduction of the multiply charged species with other ions frequently results in quite complex signal patterns for a single analyte, which significantly complicates the derivation of quantitative information from the mass spectra. Labeling strategies targeting the MS1 level further aggravate this situation, as multiple biological states such as healthy or diseased must be represented simultaneously. We developed an integer linear programming (ILP) approach, which can cluster signals belonging to the same peptide or protein. The algorithm is general in that it models all possible shifts of signals along the *m*/*z* axis. These shifts can be induced by different charge states of the compound, the presence of adducts (e.g., potassium or sodium), and/or a fixed mass label (e.g., from ICAT or nicotinic acid labeling), or any combination of the above. We show that our approach can be used to infer more features in labeled data sets, correct wrong charge assignments even in high-resolution MS, improve mass precision, and cluster charged species in different charge states and several adduct types.

Keywords: ESI • charge • decharging • cluster • adduct • ILP

Introduction

Data generated in quantitative proteome measurements is usually called a *map* (as illustrated in Figure 1). A map is constructed by conducting many MS measurements in short time intervals, typically at a frequency of 1−10 spectra per second. Each mass spectrum creates a snapshot of the peptide ions eluting from the high-performance liquid chromatography (HPLC) system at a certain time point. The time intervals are usually chosen such that the eluting peptides cover several measurements. A map contains thousands of MS scans since usually separation times of one to several hours are used.

Peptides eluting from the HPLC column are ionized in the ion source of the mass spectrometer and then the mass spectrometer measures their *mass over charge ratios* (*m*/*z*). For example, Figure 1A shows the raw map of two peptides of similar mass but different retention time (RT). Usually, the raw data is subjected to algorithms for data reduction and the measured signal is reduced to a single data point—called *feature*—representing the average retention time, the monoisotopic *m*/*z*, and the integrated intensity of the signal.

For quantitation in HPLC-based proteomics, two paradigms are prevalent. In *label-free* quantitation each biological state is measured separately, resulting in multiple maps containing the signals of the eluting peptides. To compare the signals, they first have to be identified in the corresponding maps and then grouped together (applying suitable data reduction and normalization methods). In *labeled* quantitation approaches different biological states are measured in a *single* map concurrently. To distinguish the states, they can be labeled with a fixed mass label shifting the peptide along the *m*/*z* axis (see for example Figure 1A for measured data and Figure 1B for the respective features of two peptides in two different labeling states—indicated in the figure by filled or empty symbols). Labeling techniques have the disadvantage of requiring yet another biochemical step (the labeling itself) but have the advantage of measuring the different states in one measurement, thereby circumventing the technical variation introduced by several measurements. In both paradigms the ratio of the assigned pairs of signals can be used for subsequent data analysis (e.g., for the detection of biomarkers).

The major problem in the mass spectra generated for peptides or proteins generated upon ESI[1] (and to some extent also matrix-assisted laser desorption/ionization, MALDI) rests within the fact that the peptides or proteins can be represented by a number of different molecular species, such as different charge states or ions adducted with cations or anions. Obviously, it is crucial to assess the *complete* signal(s) generated by the peptide(s) in the specific states. Unfortunately, this task is more difficult in practice than implied in Figure 1. For example Figure 1C shows the (realistic) case of 2 and 3 charge variants of each peptide with the addition of a sodium adduct.

* To whom correspondence should be addressed. E-mail: bielow@inf.fu-berlin.de. Phone: +49 (0)30 838 75137. Fax: +49 (0)30 838-75218.
† Free University Berlin.
‡ International Max-Planck Research School.
§ University of Salzburg.

10.1021/pr100177k

Published on Web 03/04/2010

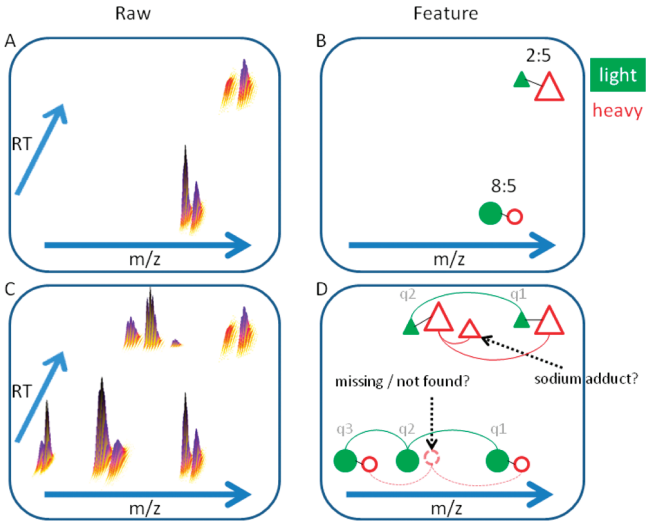

Figure 1. Schematic illustration of ESI spectral cluttering due to multiple charge states and adduct formation in experiments involving light and heavy-labeled species. (A) Two ideal peptide signals (top and bottom) eluting from a chromatographic system, each showing a light and heavy analogue. (B) Features identified from raw signal on the left with the resulting intensity ratio between labeled and unlabeled compounds. (C) Same peptides as in subfigure A but spread across several charge states. Even an adduct can be observed, which is related to the high-intensity heavy peptide at higher retention time. (D) Charge ladders and charge states indicated at the feature level for subfigure C.

Figure 1D then shows the respective feature map in which one feature is missing, which is also quite common in practice because of noisy data or algorithms not being able to detect all signals correctly.

The process of electrospray ionization has been investigated in considerable detail. Protonation sites are usually attributed to accessible basic residues (Arg, Lys, His, and N-terminus).[2,3] The number of charges that is taken up by a peptide/protein during ESI is highly complex and depends on a number of factors, for example, number of basic and acidic residues,[4,5] solution pH, solvent system,[6] presence of proton sponges,[7,8] supercharging additives[9] and instrumental factors. Suggestions have been made to experimentally shift and/or compress the charge distribution by ESI[5] to facilitate disentanglement of spectra. Recently, Kaltashov et al.[10] reviewed the information hidden in charge state distributions to infer macromolecular structure.

The problem of grouping differently charged species from the same compound in ESI spectra is often referred to as deconvolution (although this is misleading—mathematically speaking), decharging (although experimentalists usually interpret this as reducing the average charge state[4,8]), or simply disentanglement. Deconvolution is also sometimes used synonymously for deisotoping[11] or resolving overlapping shapes.[12] We thus suggest the name *decharging* for reducing multiple (deisotoped) species of the same analyte with different charge adducts into a single zero-charge signal.

In labeling approaches (such as stable isotope labeling with amino acids in cell culture (SILAC), isotope coded affinity tagging (ICAT), labeling with nicotinoyloxy succinimide (nic-NHS)) usually no decharging is applied. Instead, signals of different labeling states with equal charge are grouped and compared directly, which results in redundant information if multiple charge states are present.

In both label and label-free approaches, the quantitation is further aggravated by the presence of adducts with ubiquitous ions, like sodium and potassium, whose occurrence depends on experimental conditions, for example, usage of salts during HPLC or capillary electrophoresis (CE). Peptide signals incorporating such adducts are usually low in abundance, but will nevertheless reduce the ion count of proton-only signals.

Inferring the correct mass of a peptide or protein from charge states and charge ladders has been an active research topic from the onset of application of ESI−MS in proteomics. Early approaches targeting undigested protein samples use "global" information, that is, multiple signals of different charge states to infer the mass and are best suited for mass spectra containing only a few analytes. With the advent of high resolution mass spectrometers it became possible to use "local" information, that is, isotope patterns, to infer charge, which is sometimes the only option to infer mass when an analyte is only present in a single charge species. For protein spectra, Mann et al.[13] proposed an algorithm to fold a spectrum into mass space, thus eliminating charge ladders. Although this greatly improved mass precision, the algorithm can only deal with few analytes in one spectrum, and gives rise to artifact peaks. This algorithm was further improved by Reinhold and Reinhold[14] who reduced artifact peaks by using an entropy based measure at the cost of requiring a model distribution of charge ladders, which is applied to all masses under investigation, and a loss of the peak height-abundance relationship. For broad-range MALDI spectra, a heuristic approach[15] working on single spectra was devised, which can cluster multiple charge states considering only H^+, but relies on MALDI specific rules not applicable to ESI. The widely known ZScore-Algorithm[16] features either local or global decharging, but not both simultaneously. It can deal with complex single (stick) spectra but might also produce artifact peaks due to spectral noise. A similar algorithm along with a brief review was published by Zheng et al.[17] Du et al.[11] infer charge from local isotope peaks, and cluster all species projecting onto the same mass. This, however, is prone to wrong charge assignment during charge estimation and requires a threshold parameter. One algorithm that attempts to make use of global and local information was published by Wehofsky in 2002.[18] It rewards features with charge q when their sibling of charge $q - 1$ is also found. Unfortunately, the algorithm has no notion of retention time, only considers adducts of type H^+ and relies on identifying gap-less charge ladders. Furthermore, if charge and thus mass is estimated incorrectly, the decharged spectrum is neither likely to contain the wrong signal nor the correct one, because wrong charges are not fed back into the input spectrum. MaxQuant[19] creates charge pairs based on retention time correlation and a peptide mass estimate threshold for SILAC-based experiments. ASAPRatio[20] also uses charge pairs in ICAT-type LC/ESI−MS data to improve quantitation results. In addition to the two algorithms above, another tool capable of analyzing labeled data is VIPER.[21] It supports arbitrary mass differences (e.g., from ICAT or $^{16}O/^{18}O$ labeling) and can deal with pairs in multiple charge states.

None of the algorithms mentioned above can model charge ladders with multiple adduct combinations, for example, a combination of pure proton adduct species with a proton/sodium species from the same peptide or protein. And except for the more recent ones they were all designed for undigested

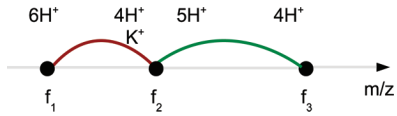

Figure 2. Two conflicting edges inducing a constraint due to inconsistent annotation of feature f_2 (implicit H^+ are shown as well).

analytes producing long charge ladders. In tryptic digests, however, one rarely encounters species of charge five or higher. There exists a solution to cluster undigested protein degradation products based on an EM algorithm, which is however not publicly available.[22] For metabolites an approach using database search accounting for different adducts was recently devised.[23] Additionally, there exists the *CAMERA* software package (http://www.bioconductor.org/packages/bioc/html/CAMERA.html), which groups metabolite mass signals based on rules for mass differences and peak shape comparison.[24]

In this work, we propose a method for identifying groups of signals belonging to the same compound in labeled or unlabeled MS data. The algorithm is general in that it models all possible shifts of signals along the m/z axis. These shifts can be induced by a different charge state of the compound, the presence of adducts (e.g., potassium or sodium), the presence of a mass shift due to isotope labeling, or any combination of the above. It allows for an iterative approach (rerunning feature detection on missing charge states, or missing pairs in labeled experiments) and can deal with missing data (e.g., gapped charge ladders). We show that by applying our algorithm, several types of errors can be corrected in a feature map, for example, wrong charge assignment or missing features. Additionally, we can achieve a reduction of data and improvements in mass precision.

Methods

The input data set is a set of features F (also called a feature map) as generated by a feature finding algorithm, each feature having at least a retention time and m/z (monoisotopic). In addition, it can be advantageous to have an initial charge estimate and an intensity value.

We model our problem first as a graph, which lends itself easily to an ILP formulation. The nodes in the graph correspond to features at a certain RT and m/z and hence to a peptide with a certain charge state, possibly with adducts and/or mass labels. Edges are inserted between pairs of nodes, if a certain combination of adducts and charge assignment of the nodes explains the mass difference between the nodes. Each edge carries information on the potential charge of its adjacent nodes and the adducts which are required to explain the resulting mass difference. For example, in Figure 2, the edge between f_1 and f_2 is inserted, because the mass difference can be explained by the assumption that f_1 has charge 6, f_2 has charge 5, and f_2 has a potassium adduct. The inserted edge hence induces charge states on f_1 and f_2.

For building the graph, we only use the most commonly occurring adducts listed in Table 1. Note that this table can be easily modified by the user if required. Simple protonation is the most common effect (and desirable due to better fragmentation behavior and decreased signal congestion[25]). Nonproton adducts are usually a result of prior prefractionation via capillary electrophoresis (CE) or high-performance liquid chromatography (HPLC).

Table 1. Adducts Commonly Observed in ESI−MS[a]

name	formula	monoisotopic mass (Da)
hydrogen	H	1.0078250319
ammonium	NH_4	18.05
sodium	Na	22.98976928
potassium	K	38.96370668

[a] All adducts occur singly charged, that is, are lacking one electron.

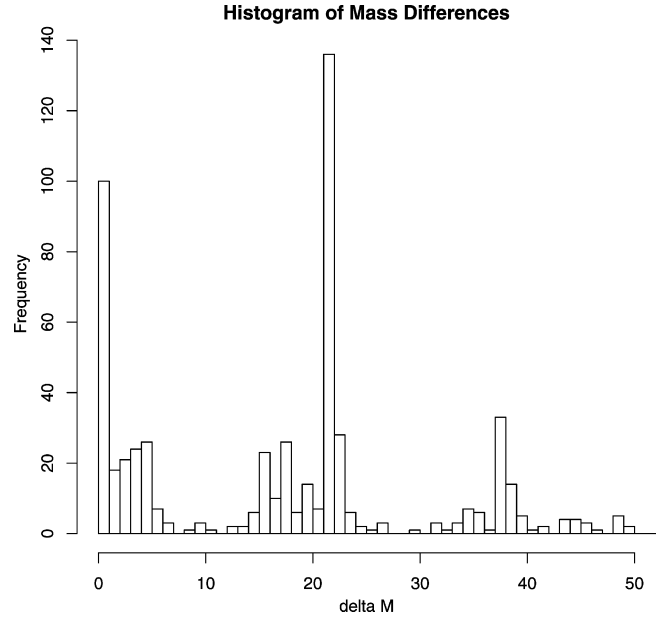

Figure 3. Histogram of pairwise mass differences, showing evidence for presence of sodium and potassium (obtained from the SPC data set used below).

Adduct frequencies can be estimated by looking at the histogram of feature pairwise mass differences (see Figure 3). At the masses of Na, K, and NH_4 (less a proton mass each) one can clearly observe clusters which indicate their presence (see Table 1 for adduct masses).

Constructing the graph in such a fashion obviously results in conflicting edge descriptions. Each pair of edges adjacent to a node might induce a constraint due to a conflicting charge/adduct combination. For example Figure 2 shows two edges adjacent to f_2, where the left edge assigns 4 protons and a positively charged potassium ion to f_2, while the right edge assigns 5 protons to f_2. Obviously, only one of these conflicting annotations can be fulfilled at a time. Hence, our goal is to choose a subset of the edges with an overall maximal weight which does not contain any pair of conflicting edges. We compute the optimal such subset by solving an ILP. In the following the method is specified in further detail.

Generating the Graph. Initially, our algorithm generates a table L of all feasible net adduct transitions, that is, the subset "lost" on one side and the subset "gained" on the other. Losing a proton and gaining a sodium adduct, for example, can serve as explanation for an edge between $[M + 2H]^{2+}$ and $[M + H + Na]^{2+}$ ions. The table contains net mass and net charge differences. An example using proton and sodium adducts can be found in Table 2. Note that we do not model redundant transitions, that is, elements that occur on both sides and would cancel each other out. Additionally, each adduct is assigned an a priori probability (e.g., from Figure 3), which allows to compute adducts transitions up to a probability threshold

Table 2. Example for Adduct Transition Table *L*

loss	gain	net charge	mass
Na^+	H^+	0	−21.9819
−	H^+	+1	1.0078
−	Na^+	+1	22.9892
H^+	$2Na^+$	+1	44.9712
Na^+	$3H^+$	+2	−19.9674
$2Na^+$	$4H^+$	+2	−41.9493
−	$2H^+$	+2	2.0146
−	HNa^+	+2	23.9965

which can be chosen generously but avoids adduct transitions that are unlikely to occur (e.g., all-sodium adducts in a charge 5 feature). Furthermore, the list is bound by the net charge value which cannot exceed q_{span}, which is the maximum number of charge states that can be bridged by edges in the graph. By default q_{span} is set to 4, which allows bridging q_3 (charge 3) and q_6 (charge 6), but would not join two nodes with q_3 and q_7. The size of *L* depends very much on the number of adducts allowed and q_{span}, but rarely exceeds 400 entries.

We now construct the adduct-transition graph $G = (V, E)$, where *V* is a set of nodes n_i corresponding to features f_i from the set *F* and *E* is a set of undirected edges $e_j = \{n_k, n_l\}$. To generate edges between nodes, the algorithm enumerates all pairwise features within a small RT delta $delta_{RT}$, as charge ladders are a property of ESI and thus have similar RT. However, if method-specific RT shifts are known (e.g., in ICAT pairs), this can easily be accounted for by specifying an adduct's intrinsic RT shift. During the enumeration, mass differences are looked up in *L*, and for all matches an edge containing the putative charge and adduct of the left and right node is inserted as well as a score, which serves as an edge weight. All charges not explicitly explained by the adduct transition are implicitly modeled as H^+ and stored in the edge as well. Obviously, edges with adduct transitions requiring a feature to take up more charges than allocated are not realized. Edges are weighted by the product of probabilities induced by the adducts required to explain the mass difference (see Table 2). However, a more involved scoring scheme is easily implemented, which could account for, for example, mass and RT deltas, feature quality and violation of a feature's local charge prediction.

To reduce the number of false positives in highly complex maps, an additional filter can be used which reduces the number of edges in the graph. In the case that $\mathrm{ch}(e_k, n_i) = \mathrm{ch}(e_k, n_j)$, we add an edge $e_k = \{n_i, n_j\}$ only if $sign(\mathrm{int}(n_i) - \mathrm{int}(n_j)) = sign(\mathrm{pr}(e_k, n_i) - \mathrm{pr}(e_k, n_j))$, where n_i and n_j are the nodes connected by edge e_k; $\mathrm{int}(n)$, $\mathrm{pr}(e, n)$ and $\mathrm{ch}(e, n)$ are the intensity, probability, and charge of node *n* induced by edge *e*. In other words, we demand that features with lower probability also have a lower abundance. We only enforce this constraint for equal charge states, as it is very hard to predict ionization behavior across multiple charges.

As table *L* only contains nonredundant adduct-transitions, it is sometimes necessary to infer sibling edges to already existing ones, which contain explicit redundant adducts. An example is given in Figure 4. As edge e_3 induces purely protonated nodes n_2, n_4 it is in conflict with edges e_2 and e_4, each inducing a sodium adduct at n_2, n_4 respectively. To enable the final solution to contain a fully connected subgraph n_1, n_2, n_3, n_4, another edge e_5 needs to be created. These inferred edges are created between any two nodes n_i, n_j for any pair of

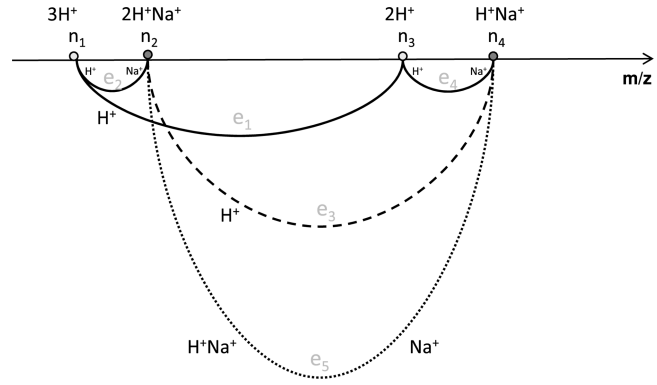

Figure 4. Example for edge inference. e_5 is inferred by using the adducts induced by e_2 and e_4. (Note that only edges important for edge inference are shown for clarity, e.g., $e = \{n_1, n_4\}$ is missing.)

edges e_k, e_l with either $e_k = \{n_i, n_m\}$ or $e_k = \{n_o, n_i\}$ and $e_l = \{n_j, n_p\}$ or $e_l = \{n_q, n_j\}$, using the adducts induced for n_i, n_j by e_k and e_l.

The graph-construction algorithm can be summarized in pseudo code like this, assuming the input *F* is sorted by RT:

for f_i in $F[start : end]$ **do**
  for f_j in $F[i+1 : index(rt(f_i) + delta_{RT})]$ **do**
    for (q_1, q_2) in Q $\times$ Q **do**
      adduct_candidates = massDeltaLookup$(f_i \cdot q_1, f_j \cdot q_2, mz_tol)$
      for ac_i in adduct_candidates **do**
        if (not intensityFilter(f_i, f_j, ac_i)) **then**
          continue
        end if
        insertEdge(E, ac_i, f_i, f_j)
      end for
    end for
  end for
end for
edgeInference(E)

The for-loop testing all feature charge combinations is optional and only used when the algorithm is in "discovery" mode, that is, it searches for edges without relying on the annotated charge of the feature but instead enumerates all possible values.

Constructing the ILP. Having constructed the adduct-transition graph for our problem, it is straightforward to define the corresponding ILP. During this phase, all edges sharing one or more nodes are checked for consistency, that is, if any pair of edges induces an inconsistent adduct annotation for the shared feature. Consistency requires:

(1) identical charge

(2) identical adduct composition

An example for two inconsistent edges can be found in Figure 2. Feature f_2 is assigned adducts K^+ $4H^+$ by the left edge, whereas the right edge induces $5H^+$, both of which

cannot be true simultaneously. We introduce x_i to indicate the presence/absence of edge e_i from the solution and c_i as the score of edge e_i.

The ILP is defined as:

$$\max c^t x \quad (1)$$

$$s.t.\ x_i + x_j \leq 1, x_i, x_j \in 0, 1 \text{ for all pairs of inconsistent edges} \quad (2)$$

The ILP's output is a set of active edges, thus finding all connected components will automatically cluster nodes (features) into groups representing charge ladders with adducts and/or labeled pairs.

Postprocessing. During postprocessing clusters can be discarded using a filter which reduces spurious hits. The "backbone" filter will only allow clusters which have at least one feature whose charge can be explained by protons only, that is, is part of the backbone of a charge ladder. Otherwise (especially in complex maps when feature charge is estimated by the algorithm rather than using the feature's charge) wrong clusters might be found, for example, ($[3K]^{3+}$, $[5Na]^{5+}$).

Availability. The algorithm will be available in the next release version of the C++ software library OpenMS,[26] available for all major platforms at http://www.OpenMS.de, or can be obtained from the authors.

Results

The algorithm was applied to several real data sets. On all data sets analyzed here, the running times for our algorithm were below 5 s (2.26 GHz Core2Duo) and memory requirements did not exceed 500 Mb. Time and memory requirements can increase, however, if many adduct types are allowed and the feature finder charge is not fixed.

We will now show some practical cases, where decharging can help to increase data quality. We compare our approach to a commercially available tool (Xtract) and to an algorithm to identify pairs which is implemented in OpenMS. Comparison with other packages is difficult, since they are partially specialized for certain labeling methods like SILAC or ICAT. We plan to do further evaluations with other tools if they are applicable to the problems addressed in this work. All parameter settings, raw data, and result files are available for download at http://page.mi.fu-berlin.de/bielow/data/RSM2010/. Where not indicated otherwise, we used the OpenMS PeakPicker (v1.7) for centroiding raw data and the OpenMS FeatureFinder (v1.7) for generating feature maps.

Increasing Mass Precision. We applied our decharging algorithm to one of the SPC data sets (Mix1, LTQ-FT, 20060502data08).[27] The data set is a tryptic digest of 18 proteins measured in an LTQ-FT mass spectrometer and is available at http://regis-web.systemsbiology.net/PublicDatasets/. The interesting region of 500−4000 s and 400−1400 Da was excised and only every second scan was retained from the MS1 data, as only they contained the FT scans. The OpenMS PeakPicker and FeatureFinder were applied to the raw data, resulting in 1064 features. Subsequent internal calibration using high-confidence MS2 identifications was applied, to enable the calculation of a standard deviation between monoisotopic feature position in m/z and MS2 identifications. Decharging was applied to find clusters of corresponding features stemming from the same peptide with different adducts. We found evidence for adducts (see Figure 3), and thus allowed H^+, Na^+, K^+, and NH_4^+. The data set being a high-resolution measurement, the charges assigned by the FeatureFinder are mostly correct, although misassignments did occur (especially when the isotope pattern deviated strongly from the averagine model). Hence, we allowed the decharging algorithm to alter the FeatureFinder charge. The adduct-transition graph had 1064 nodes, 344 edges, inducing 167 constraints. About 35% of all features (371) were grouped into 155 clusters during decharging and their monoisotopic m/z position corrected using the average mass predicted by all members of the cluster. For all other features (693) no partner was identified. For 20 clusters an MS2 identification was available. This allowed to calculate the mass deviation between the predicted MS2 mass and the feature mass. The standard deviation between the features' monoisotopic m/z and the theoretical m/z position predicted by MS2 identifications prior to decharging was 1.044 ppm. In contrast, it was significantly reduced to 0.527 ppm after decharging, due to the fact that feature masses are averaged by our algorithm over all members of a cluster. The increase in mass precision by clustering obviously only applies to features which are members of a cluster and can thus benefit from decharging. Additionally, we examined those features whose charge (as assigned by the feature finder) was altered in the ILP solution. In total, the charge of 8 features was changed by our algorithm and except for one these reassignments were found to be correct by manual verification of the raw data.

Finding Pairs in Labeled Data. We applied our algorithm to a centroided data set of MHC peptides[28] which contains 4117 scans and 3083 features. Nicotinic acid labeling was used to tag two samples with either a light or heavy label (in which four hydrogen atoms were replaced by deuterium) prior to mixing and LC−MS analysis. Our algorithm generally supports any kind of labeling as long as the mass difference can be formulated as an empirical formula (see below).

As the charge estimation using the OpenMS FeatureFinder was very reliable on this data set, feature charges were not altered. The set of possible adducts was set to H^+ and D4 − 4H, the former being simply protonation, the latter being an uncharged adduct describing the net mass gain of 4 Da for the heavy analogue due to deuterium exchange. We allowed up to two uncharged adducts for the computation of L. The adduct-transition graph had 3083 nodes, 653 edges, inducing 349 constraints.

To compare our results we tested the labeled pair finder of OpenMS which features pair finding using an arbitrary list of allowed mass and RT shifts. We found 293 pairs using this standard approach. Using the decharging algorithm we found 307 pairs, 16 of which have a partner pair with another charge state (see Figure 5 for an example). These 16 pairs can be condensed into 8, because they represent the same peptide. Additionally, it allows to compute the average of two intensity ratios (see Figure 6).

Sixty-four clusters of size three were also found, 11 of which all contained features of the same charge state, thus indicating a potential conflict in uniquely identifying the light and heavy pair: when ordered by m/z, it would be possible that either feature 1 and 2 or feature 2 and 3 represent the light and heavy peptide. Other pair-finding algorithms will most likely just pick one greedily. Even the common precaution in standard pair finding algorithms, requiring that any third feature must lie x Da further away from a pair would not be beneficial here as x would need to be larger than the pair mass difference to avoid

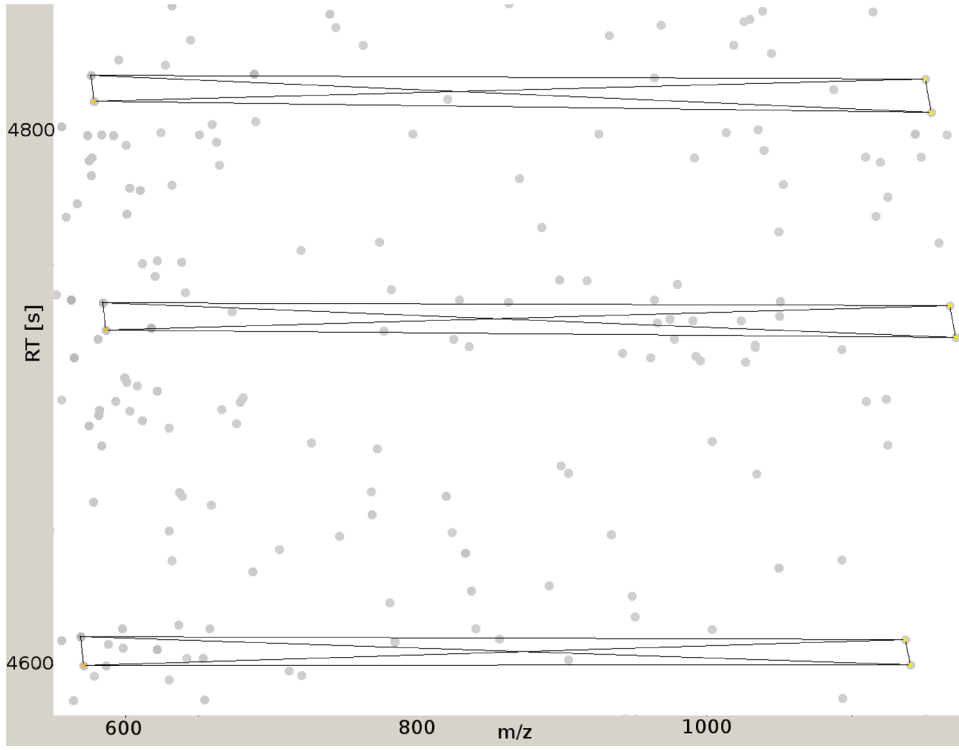

Figure 5. Example of charge ladders spanning two charge states (including light and heavy partners). Edges connect all features of the cluster as found during graph construction.

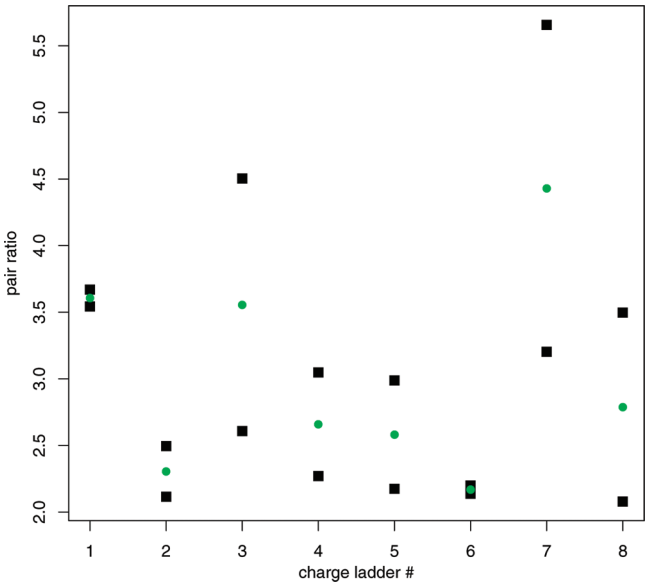

Figure 6. Intensity ratios (light vs heavy feature) from nicNHS pairs of different charge state. Ratios are depicted as squares, the average as circles. Some pairs (e.g., #6) show very similar intensity ratios in different charge states, whereas other charge ladders (e.g., #7) show a 2-fold difference. These differences can aid in determining the confidence in the observed ratios.

ambiguous pairing. Choosing x like this would result in many pairs not being discovered in the data set. Our approach will find the ambiguities, and allows discarding/marking those clusters. Another constellation for clusters of size three occurred 53 times: one light/heavy pair is identified, and additionally, a third feature (either light or heavy) of a different charge. The missing fourth feature was either not discovered during feature finding or is simply not detected by the instrument. An example of the former case is given in Figure 7. Without reference of another charge pair or MS2 identification it is difficult to infer which of the two partners is present, or if the identified feature even has a partner. Manual inspection of the data set suggests that about 60% of the 53 clusters indeed have a fourth feature which was simply not detected by feature finding. Reiterating the feature finding step using the 53 seeded positions suggested by our algorithm yielded 29 new features (10% increase in pair count).

Calculating Intact Protein Masses. We analyzed a hemoglobin HPLC/ESI−MS raw data set consisting of 10 scans containing *Hbα*1 and *Hbβ* measured on an LTQ Orbitrap XL mass spectrometer with a resolution of 100,000. We compared the results of our algorithm with the Xtract module of Thermo's Proteome Discoverer 1.0. This module allows decharging at the raw data level, and operates scan-wise. The maximum charge was set to 30 and we enabled the reporting of monoisotopic masses only. Xtract finished after 237 s of CPU time (2.26 GHz Core2Duo). Note that Xtract reports the monoisotopic protein mass as singly charged.

We used Hardkloer (v1.22)[29] to identify features scanwise and the postprocessing tool Kroenik (v1.3) to summarize features occurring in multiple scans. Minimum and maximum charge were set to 4 and 30, union and intersection mode were enabled, and S/N was set to 1. Decharging was set to consider sodium and potassium adducts and to correct for monoisotopic shifts of up to one position to the left or right. Alteration of charge values was disabled. The adduct-transition graph had 104 nodes, 315 edges, inducing 2590 constraints.

CPU time from the raw data to features took 15 s, subsequent decharging one second (2.26 GHz Core2Duo). Our algorithm found 68 distinct masses (clusters). The two largest-sized clusters represented the hemoglobin subunits—cluster *A* for *HBα* (size 14 ranging from charge 8−18 with 11 proton-only

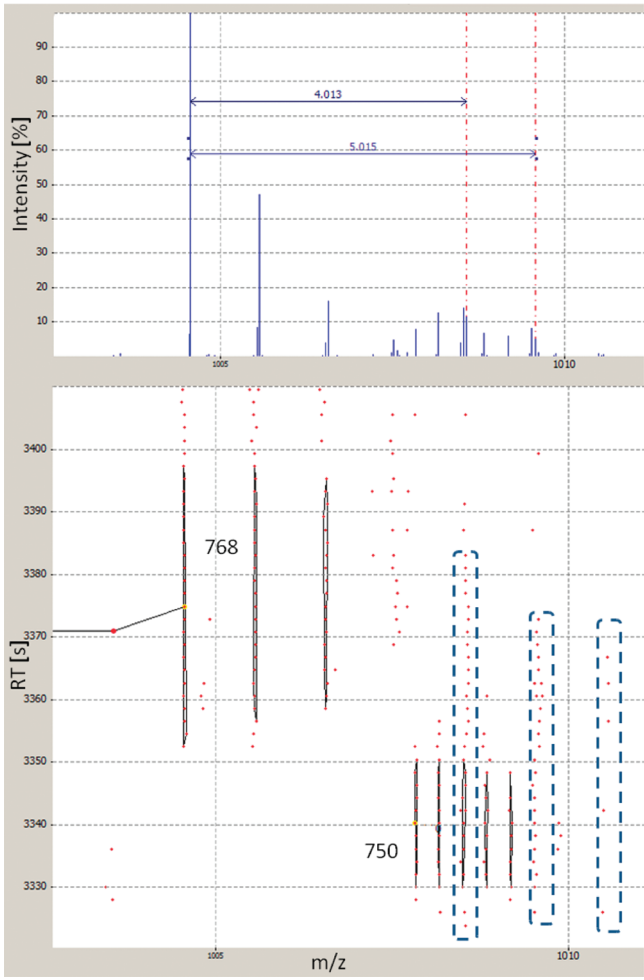

Figure 7. By evidence from a triple (2× q_2, 1× q_1) we can infer the presence of the heavy partner of the q_1 feature #768. The top section shows the projection of the *m*/*z* dimension. One can clearly see signals at 4 and 5 Da from the monoisotopic mass trace of feature #768. The missing feature's mass traces are indicated by dashed boxes in the map view (lower section). The reason for not identifying this feature is probably the presence of feature #750.

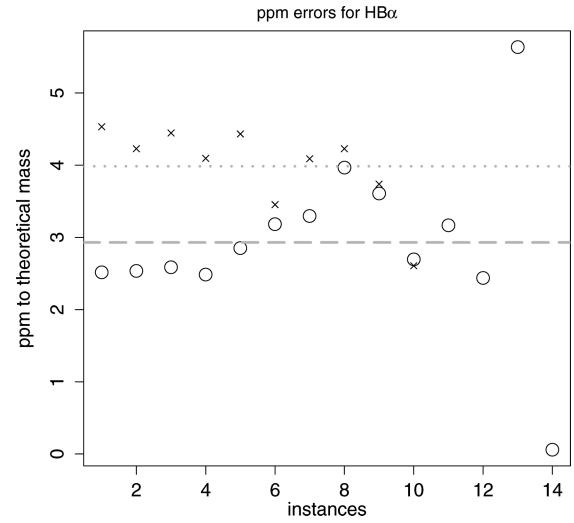

Figure 8. Deviation of observed masses for *HB*α (in ppm). Circles represent charged features clustered into *A*, crosses are the mass estimate errors by Xtract from the 10 scans. Dashed (our) and dotted (Xtract) lines are the mean values.

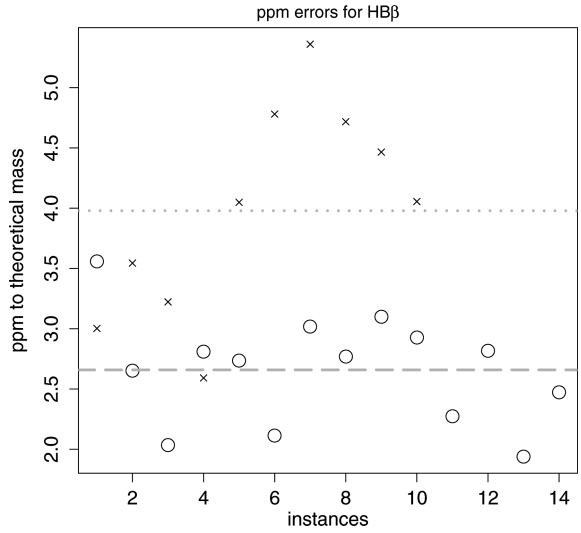

Figure 9. Deviation of observed masses for *HBβ* (in ppm). Circles represent charged features clustered into *B*, crosses are the mass estimate errors by Xtract from the 10 scans. Dashed (our) and dotted (Xtract) lines are the mean values.

features, 2 potassium and 1 sodium adducted features, average measured monoisotopic mass was 15116.92939 Da, molecular mass calculated from the sequence was 15116.88510 Da), cluster *B* for *HBβ* (size 14 ranging from charge 9−17 with 10 proton-only features, charge 11 occurring split into two proton-only features with different RT, 2 sodium and potassium adducted features each, average measured monoisotopic mass was 15857.29186 Da, molecular mass calculated from the sequence was 15857.24969 Da).

As Xtract reports several masses (one per scan) for each hemoglobin subunit, we extracted the relevant regions to obtain an overall of 10 molecular mass values for each subunit. By averaging these 10 mass estimates we obtained masses 15116.95 and 15857.31 Da for *HB*α and *HBβ*. Figure 8 and Figure 9 show the relative mass deviations for both methods, the horizontal lines indicating the relative mass deviation from the theoretical mass for our approach and Xtract. Note that our approach is closer to the predicted theoretical mass (at 0 ppm) for both subunits. However, the instrument seems not to be optimally calibrated, as both method's standard deviation (0.4987 and 0.3858 ppm for our method with 2 and 1 outliers removed, 0.3510 and 0.8808 ppm for Xtract with 1 and 0 outliers removed) is lower than the gap to the theoretical mass of the protein. Outliers were removed using *z*-scores and a *p*-value threshold of 0.95. Hence, our decharging algorithm can also be utilized to recalibrate the mass spectrometer with signals of multiply charged ion species. Our approach can additionally group all sodium and potassium peaks into the main cluster (if desired by the user), which further disentangles the results. Furthermore, our approach also works on single spectra, thus allowing us to estimate the mass error from multiple charges - an information not provided by Xtract.

Conclusion

We demonstrated that decharging is useful for many applications in quantitative proteomics. The algorithm is not restricted to a specific instrument or resolution, and although it is intended for ESI data, it should also be applicable to MALDI

data when multiply charged ions are observed (e.g., for whole protein measurements[15]). We obtained promising results, when testing the algorithm on nonpublic, poly peptide CE/MS data[22] containing high charge states up to 16 and several adduct types. In future work we plan on multiple improvements: First, enabling the scoring function parameters to be estimated from the data, and second, reducing the memory footprint, which can grow to several gigabytes if the search for edges is exhaustive. This can be achieved by solving disjunct subgraphs of G instead of using G directly to derive the ILP.

Acknowledgment. We thank Sandro Andreotti (Free University of Berlin) and Clemens Gröpl (University of Greifswald) for fruitful discussions and Andreas Bertsch (University of Tübingen) for providing the nicotinic acid data set. C.B. and S.R. are supported by the European Commissions's 7th Framework Program (GA202222).

References

(1) Fenn, J.; Mann, M.; Meng, C.; Wong, S.; Whitehouse, C. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **1989**, *246*, 64–71.
(2) Nesatyy, V. J.; Groh, K.; Nestler, H.; Suter, M. J.-F. On the acquisition of +1 charge states during high-throughput proteomics: Implications on reproducibility, number and confidence of protein identifications. *J. Proteomics* **2009**, *72*, 761–770.
(3) Prakash, H.; Mazumdar, S. Direct correlation of the crystal structure of proteins with the maximum positive and negative charge states of gaseous protein ions produced by electrospray ionization. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 1409–1421.
(4) Svoboda, M.; Meister, W.; Kitas, E. A.; Vetter, W. The influence of strongly acidic groups on the protonation of peptides in electrospray MS. *J. Mass Spectrom.* **1997**, *32*, 1117–1123.
(5) Krusemark, C. J.; Frey, B. L.; Belshaw, P. J.; Smith, L. M. Modifying the Charge State Distribution of Proteins in Electrospray Ionization Mass Spectrometry by Chemical Derivatization. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 1617–1625.
(6) Iavarone, A. T.; Jurchen, J. C.; Williams, E. R. Supercharged Protein and Peptide Ions Formed by Electrospray Ionization. *Anal. Chem.* **2001**, *73*, 1455–1460.
(7) Catalina, M. I.; van Den Heuvel, R. H. H.; van Duijn, E.; Heck, A. J. R. Decharging of globular proteins and protein complexes in electrospray. *Chem.—Eur. J.* **2005**, *11*, 960–968.
(8) Benesch, J. L. P.; Robinson, C. V. Mass spectrometry of macromolecular assemblies: preservation and dissociation. *Curr. Opin. Struct. Biol.* **2006**, *16*, 245–251.
(9) Sterling, H. J.; Williams, E. R. Origin of supercharging in electrospray ionization of noncovalent complexes from aqueous solution. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 1933–1943.
(10) Kaltashov, I. A.; Abzalimov, R. R. Do ionic charges in ESI MS provide useful information on macromolecular structure. *J. Am. Soc. Mass Spectrom.* **2008**, *19*, 1239–1246.
(11) Du, P.; Angeletti, R. H. Automatic Deconvolution of Isotope-Resolved Mass Spectra Using Variable Selection and Quantized Peptide Mass Distribution. *Anal. Chem.* **2006**, *78*, 3385–3392.
(12) Horn, D. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 320–332.
(13) Mann, M.; Meng, C. K.; Fenn, J. B. Interpreting mass spectra of multiply charged ions. *Anal. Chem.* **1989**, *61*, 1702–1708.
(14) Reinhold, B. B.; Reinhold, V. N. Electrospray ionization mass spectrometry: Deconvolution by an entropy-based algorithm. *J. Am. Soc. Mass Spectrom.* **1992**, *3*, 207–215.
(15) Malyarenko, D. I.; Cooke, W. E.; Bunai, C. L.; Manos, D. M. Automated assignment of ionization states in broad-mass matrix-assisted laser desorption/ionization spectra of protein mixtures. *Rapid Commun. Mass Spectrom.* **2009**, *24*, 138–146.
(16) Zhang, Z.; Marshall, A. A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 225–233.
(17) Zheng, H.; Ojha, P. C.; McClean, S.; Black, N. D.; Hughes, J. G.; Shaw, C. Heuristic charge assignment for deconvolution of electrospray ionization mass spectra. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 429–36.
(18) Wehofsky, M.; Hoffmann, R. Automated deconvolution and deisotoping of electrospray mass spectra. *J. Mass Spectrom.* **2002**, *37*, 223–9.
(19) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367–1372.
(20) Li, X.-J.; Zhang, H.; Ranish, J. A.; Aebersold, R. Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Anal. Chem.* **2003**, *75*, 6648–6657.
(21) Monroe, M. E.; Tolić, N.; Jaitly, N.; Shaw, J. L.; Adkins, J. N.; Smith, R. D. VIPER: an advanced software package to support high-throughput LC-MS peptide identification. *Bioinformatics (Oxford, England)* **2007**, *23*, 2021–203.
(22) Wittke, S.; Kaiser, T.; Mischak, H. Differential polypeptide display: the search for the elusive target. *J. Chromatogr., B: Anal. Technol. Biomed. Life Sci.* **2003**, *803*, 17–26.
(23) Draper, J.; Enot, D. P.; Parker, D.; Beckmann, M.; Snowdon, S.; Lin, W.; Zubair, H. Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour 'rules'. *BMC Bioinf.* **2009**, *10*, 227.
(24) Tautenhahn, R. Bioinformatics Research and Development. *Annotation of LC/ESI-MS Mass Signals*; Berlin: Heidelberg, 2007; pp 371−380.
(25) Keller, B. O.; Sui, J.; Young, A. B.; Whittal, R. M. Interferences and contaminants encountered in modern mass spectrometry. *Anal. Chim. Acta* **2008**, *627*, 71–81.
(26) Sturm, M.; Bertsch, A.; Gröpl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; Kohlbacher, O. OpenMS -an open-source software framework for mass spectrometry. *BMC Bioinf.* **2008**, *9*, 163.
(27) Klimek, J.; Eddes, J. S.; Hohmann, L.; Jackson, J.; Peterson, A.; Letarte, S.; Gafken, P. R.; Katz, J. E.; Mallick, P.; Lee, H.; Schmidt, A.; Ossola, R.; Eng, J. K.; Aebersold, R.; Martin, D. B. The standard protein mix database: a diverse data set to assist in the production of improved Peptide and protein identification software tools. *J. Proteome Res.* **2008**, *7*, 96–103.
(28) Lemmel, C.; Weik, S.; Eberle, U.; Dengjel, J.; Kratt, T.; Becker, H.-D.; Rammensee, H.-G.; Stevanovic, S. Differential quantitative analysis of MHC ligands by mass spectrometry using stable isotope labeling. *Nat. Biotechnol.* **2004**, *22*, 450–454.
(29) Hoopmann, M. R.; Finney, G. L.; MacCoss, M. J. High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. *Anal. Chem.* **2007**, *79*, 5620–5632.

PR100177K