Combinatorial Approaches for Mass Spectra Recalibration

Sebastian Böcker and Veli Mäkinen

***Abstract*— Mass spectrometry has become one of the most popular analysis techniques in Proteomics and Systems Biology. With the creation of larger datasets, the automated recalibration of mass spectra becomes important to ensure that every peak in the sample spectrum is correctly assigned to some peptide and protein. Algorithms for recalibrating mass spectra have to be robust with respect to wrongly assigned peaks, as well as efficient due to the amount of mass spectrometry data. The recalibration of mass spectra leads us to the problem of finding an optimal matching between mass spectra under measurement errors.**

We have developed two deterministic methods that allow robust computation of such a matching: The first approach uses a computational geometry interpretation of the problem, and tries to find two parallel lines with constant distance that stab a maximal number of points in the plane. The second approach is based on finding a maximal common approximate subsequence, and improves existing algorithms by one order of magnitude exploiting the sequential nature of the matching problem. We compare our results to a computational geometry algorithm using a topological line-sweep.

***Index Terms*— Biotechnology, mass spectrometry, combinatorial pattern matching, computational geometry.**

I. Introduction

Mass spectrometry is one of the most popular analysis techniques in the emerging field of Systems Biology: the analysis of Protein Mass Fingerprints [1] and tandem mass spectra for protein identification [2] and *de novo* sequencing [3] is performed daily in thousands of laboratories around the world. In addition, SELDI-TOF (surface enhanced laser desorption/ionization time-of-flight) mass spectrometry of protein mixtures is increasingly used for the identification of biomarkers [4]. Among the benefits of mass spectrometry is its unique accuracy: masses of sample molecules can be determined with an accuracy of parts of a neutron mass.

Mass spectra are usually externally calibrated, resulting in mass inaccuracies in the measured mass spectrum [5]. Such inaccuracies interfere with the interpretation of mass spectrometry data, because distinct peptides can have almost identical mass. This often leads to erroneous assignment of peaks in the (measured) sample spectrum, and can prevent a proper interpretation of the spectrum. For example, badly calibrated Protein Mass Fingerprints (PMF) spectra frequently do not allow an unambiguous identification of the protein. For SELDI-TOF mass spectrometry, subtle changes in protein intensities have to be detected: here, incorrect calibration of mass spectra is a severe problem for reproducibility of experiments [6], as well as for the comparability of data from different laboratories [7]. Calibration also shows a strong impact on correct sequence determination for *de novo* sequencing of peptides using tandem mass spectrometry [8].

In this paper, we study methods for robust recalibration of mass spectra. Here, one uses knowledge about the physics underlying the mass spectrometry measurement in combination with a hypothesis regarding proteins or peptides present in the sample, to increase the mass accuracy of the measurement. Assume we are given a PMF sample mass spectrum with inaccurate external calibration. If the simulated mass spectrum of a database protein shows reasonable similarity to the sample spectrum, then we can try to find a calibration of the sample spectrum that makes it "more similar" to the simulated spectrum and, at the same time, is in accordance with the physics underlying the measurement. Regarding peptide *de novo* sequencing using tandem mass spectrometry, almost all approaches generate a set of candidate sequences that are further evaluated including a recalibration of the sample spectrum [8]. For a set of SELDI-TOF mass spectra, usually an arbitrary spectrum from the set is used as the reference spectrum. Note that recalibration cannot *replace* external calibration of mass spectra because recalibration requires a decent initial calibration to start from. Instead, recalibration can improve the mass accuracy of externally calibrated mass spectra.

In the following, we assume that mass spectra are represented by a list of peak masses, plus potentially other peak attributes such as intensities. Modeling mass spectra as a continuous function is not beneficial for recalibration, because we want to concentrate on prominent features (i.e. intense peaks) of the spectrum rather than regions of low intensity that often represent biochemical and physical "noise". Let A and B be two sets of masses, where B corresponds to the reference spectrum and A to the measured spectrum.

We approach this problem in a two-step manner. First, we construct a linear transformation between mass spectra that is robust to outliers: search for the best linear transformation mapping a maximum number of points of A close to points of B. The detection of outliers is important because recalibration can easily be corrupted if we wrongly match two peaks in A, B that in fact stem from proteins or peptides with distinct masses. We review three combinatorial, deterministic methods for the efficient and robust identification of outliers using linear transformations. Our experiments demonstrate that linear transformations are sufficient to construct a peak mapping between spectra, even for more elaborate types of calibration functions [5]. Our simulations also show that weaker models of mass calibration, such as "shifting" one spectrum by adding a constant mass, lead to poor mass accuracies.

Second, we can use our knowledge about mass spectrometry physics to obtain a highly accurate recalibration of the mass spectra. If outliers are excluded and a peak matching between mass spectra is known, the recalibration problem can be efficiently

Manuscript received, revised, ...

First author contact information: Lehrstuhl für Bioinformatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany. boecker@minet.uni-jena.de

Second author contact information: Department of Computer Science, P.O. Box 68 (Gustaf Hällströmin katu 2b), FIN-00014 University of Helsinki, Finland. vmakinen@cs.helsinki.fi

solved using known techniques from approximation theory [9] or statistics, such as Ordinary Least Squares (OLS) regression.

Recalibration has been frequently proposed in the literature for the analysis of mass spectrometry data. Often, sample mass spectra are "shifted" without explicitly pointing out that this, in fact, recalibrates the sample mass spectrum. Previously existing methods usually ignore the fact that two peaks can be wrongly matched because corresponding peptides or proteins have almost identical mass. Wong, Cagney, and Cartwright [10] present a heuristic approach for recalibrating mass spectra but ignore calibration physics as well as outliers. Wolski et al. [11] assume that PMF sample mass spectra show significant overlap in peptides present, and use this similarity for iterative recalibration. For SELDI data the problem of wrong peak assignment is often less prevailing, and Jeffries [12] recalibrates SELDI mass spectra using splines as calibration functions. Matthiesen and co-workers [2], [13] recalibrate Tandem MS spectra to improve database search results, and heuristically remove outliers from this recalibration: see Fig. 2 and 4 in [2] for examples on the impact of recalibration on protein identification. Bern and Goldberg [8] notice that excluding outliers is mandatory for robust recalibration, and use Least Median of Squares regression to detect these outliers, see Sect. III-C below. Wool et al. [14] pre-calibrate mass spectra, based on the clustering of peptides around certain masses. See [15], [16] for the impact of accurate calibration for peptide identification.

II. Physics of Mass Spectrometry

A mass spectrometer cannot determine the masses of sample molecules directly but only measures a derived physical property, such as voltages U, V for quadrupole instruments,[1] or time-of-flight for TOF instruments. These physical properties are transformed into mass-to-charge ratios of sample molecules using a *calibration function*. The coefficients of this function are most often determined *externally* using a separately measured calibration mass spectrum that contains molecules of known mass only. The crux of this approach is that, in principle, subtle changes of instrument parameters make it necessary to determine a separate calibration function for every single mass spectrum. These subtle changes cause mass inaccuracies in the sample mass spectra.

The concept of *recalibrating* mass spectra is to use hypothetical knowledge of the investigated sample to compute a more accurate calibration function. For example, assume that a sample contains an unknown protein. If we are given a database of proteins and have to decide what protein fits the measured mass spectrum best, then we can simulate a mass spectrum for every protein in the database, and use this predicted spectrum to calculate a new calibration function. To make this approach work, determination of the calibration function has to be robust and fast: only one of the proteins in the database corresponds correctly to the measured mass spectrum, but many simulated spectra can show some similarity, after counting for measurement errors. Computing a "wrong" calibration function in such cases will corrupt the subsequent analysis. Furthermore, the recalibration algorithm has to be fast, since recalibration must be performed for every simulated spectrum that shows at least some similarity to the measured mass spectrum.

[1]To be more precise, the physical property of the sample molecule is that it will pass through the quadrupole filter on a stable trajectory for some voltages U, V.

A. Time-of-Flight Mass Spectrometry

As an example, we next describe in more detail one of the predominant mass spectrometry techniques for analyzing biomolecules. Using Time-of-Flight (TOF) mass spectrometry, ionized sample molecules are accelerated in an electric field, and light molecules reach a higher final velocity than heavy molecules. Accordingly, one can use flight times of sample molecules to infer their masses. For the sake of brevity, we will talk about mass m instead of mass-to-charge ratio m/z of sample molecules, because TOF mass spectrometry is usually combined with pulsed ionization techniques (MALDI or SELDI) that produce predominantly single-charged molecules.

Using TOF mass spectrometry, we measure the time t that a molecule needs to get from the source to the detector. Ionized molecules are *accelerated* in an electric field with constant force F, then *drift* for some time through the field-free flight tube before they hit the detector. For a single charged molecule, the acceleration solely depends on its mass, $a = \frac{F}{m}$. Let s_A denote the distance of acceleration and s_D the distance of drift, then the total time-of-flight of a molecule with mass m is

$$t = t_\mathrm{A} + t_\mathrm{D} = \frac{1}{\sqrt{a}} \left(\frac{s_\mathrm{D} + 2s_\mathrm{A}}{\sqrt{2s_\mathrm{A}}} \right) \tag{1}$$

where $a = \frac{F}{m}$. Solving for m yields

$$m = F \frac{2s_\mathrm{A}}{(s_\mathrm{D} + 2s_\mathrm{A})^2} \cdot t^2 \tag{2}$$

and, hence masses have a quadratic dependence on time-of-flight values. We ignore initial velocities of molecules, as well as other "perturbations" in (1) and (2).

Deviating calibration results from multiple causes, such as matrix crystals of diverse heights, leading to a change of acceleration distance s_A from sample to sample, or deposits at the electrodes over time. Gobom et al. [5] show how calibration of mass spectra changes in a matter of hours, and also depends on the position of a sample on the target plate.

B. Recalibrating Mass Spectra

Calculating a calibration polynomial is possible for regular sample spectra as long as the exact masses of all sample molecules are known. In this case, we can use methods from approximation theory [9], [17] and statistics, such as Haar approximation or OLS, to compute the calibration function. But a sample spectrum may allow for wrong or ambiguous matching of detected and reference peaks. The above methods are not capable of detecting and excluding outliers from the fitting process.

So, we propose a two-step recalibration process. First, a *linear* mapping between sample spectrum peaks and reference masses is constructed. Here, the external calibration of the mass spectrum can be used. Restricting ourselves to linear mappings allows for very fast methods for this task. Second, a new calibration polynomial is calculated from these tuples using methods from approximation theory and statistics. Alternatively, we can find an optimal linear function for recalibration in the mass domain, independent of the underlying mass spectrometry physics: if the mass of a molecule depends linearly on its derived physical property (e.g. quadrupole instruments), then this will result in an optimal recalibration.

In some cases, one cannot robustly estimate all parameters of the calibration function, that is a spline or a high-order polynomial

[5]. In such cases, the linear recalibration of Sect. III can still correct the (absolute or relative) mass error of peaks: Gobom et al. [5] use calibration polynomials of order 15 and note that the relative mass error is proportional to the mass and, hence, can be corrected using a linear function.

A related approach for recalibration has successfully been applied to tandem mass spectra in [8], using (randomized) Least Median of Squares regression to exclude outliers from the analysis. Here, we concentrate on deterministic methods to facilitate reproducibility of the analysis.

III. Linear Recalibration of Mass Spectra

In the following, we describe three approaches for finding a linear recalibration of mass spectrometry data that can exclude outliers. All three algorithms are combinatorial and deterministic, but the third algorithm allows for a statistical interpretation. The first two algorithms have been developed by the authors, the third algorithm is based on topological line sweeping.

We formalize the calibration task as a *linear one-dimensional point set matching problem*: given two sets of real values, i.e. one-dimensional point sets $A, B \subseteq \mathbb{R}$, find a linear function $f : \mathbb{R} \to \mathbb{R}$ such that $\left|E_f\right|$ is maximum, where E_f is the edge set of a bipartite graph on A, B such that $\{a, b\} \in E_f$ if and only if $|f(a) - b| \leq \varepsilon$. Note that some $a \in A$ can be mapped into ε-distance of several $b \in B$ and vice versa. In fact, in most instances, there is a degenerate optimum solution mapping all points of A into ε-distance from one point of B. In our application such degenerated cases can be avoided by restricting the search space: the measurement technique gives some absolute limits for the maximum scale and translation values. Within that range of transformations, degenerated solutions a rare. A more rigorous way to define the problem, however, is to search for E_f that contains the largest one-to-one mapping; we call this the *linear one-dimensional one-to-one point set matching problem*.

In our application, A and B are the sets of mass values, and f is the recalibration polynomial of degree one. We detect outliers by allowing only matches satisfying the ε-limitation. A reasonable value for ε can be estimated depending on the measurement device and other conditions.

The rest of this section is as follows: Sect. III-A contains the basic matching algorithm for the linear one-dimensional point set matching problem with runtime $O((mn)^2 \log m)$ and $O(m + n)$ space, while the one-to-one variant is covered in Sect. III-D with runtime $O(m^3 n^2)$ and $O(m + n)$ space. In Sect. III-B and III-C we transform the problem into a geometric line stabbing problem where the set of points $S \subseteq A \times B$ contains all pairs that may potentially allow for recalibration; for our application, usually $N := |S|$ is linear in $m + n$. Sect. III-B presents a line sweep algorithm with runtime $O(N^2 \log N)$ and space $O(N)$. In Sect. III-C we adopt a topological line sweep algorithm with runtime $O(N^2)$ and space $O(N)$.

A. Point Set Matching Algorithm

To solve the matching problem, consider a set F of representative linear functions constructed as follows. Let $B(\varepsilon) = \{p - \varepsilon, p + \varepsilon \mid p \in B\}$. For each quadruple (a', a, b', b) such that $a', a \in A$ with $a' < a$ and $b', b \in B(\varepsilon)$ with $b' < b$, add function $f(x) = \frac{b-b'}{a-a'}(x - a') + b'$ to F. Each function in F defines a translation and scaling that maps two points of A into ε distance from some points of B. Conditions $a' < a$ and $b' < b$ prevent reflections. Using a simple shifting argument, we infer that this is the sufficient set of transformations to be examined. The size of this set is $O((mn)^2)$, where $m := |A|$ and $n := |B|$. To find the optimum transformation f, construct all E_f for $f \in F$ incrementally, and choose the f that corresponds to the largest $\left|E_f\right|$: for each representative translation $t = b' - a'$, where $a' \in A$ and $b' \in B(\varepsilon)$, construct the set of scale ranges $R(a', b') = \{[\frac{b-\varepsilon-b'}{a-a'}, \frac{b+\varepsilon-b'}{a-a'}] \mid a \in A, b \in B\}$. Sort the endpoints of ranges in $R(a', b')$ into increasing order, and scan through them incrementing and decrementing a counter to know at any point how many scale ranges are "active". The largest counter value is obtained at the optimum scale for the fixed translation. Repeating the process for all representative translations gives the overall optimum transformation. Noticing that the scale ranges corresponding to a fixed $a \in A$ can be obtained in sorted order by scanning through sorted B, the algorithm can be implemented to run in $O((mn)^2 \log m)$ time by merging the m sorted lists at each phase. As described here, the algorithm uses $O(mn)$ space. However, it is easy to reduce the space to $O(m+n)$ by using the *match matrix* interpretation given in Sect. III-D.

B. Maximum Line-Pair Stabbing Algorithm

We next use a geometrical interpretation of the problem to find the second efficient algorithm for mass spectra recalibration. In the *Maximum Line-Pair Stabbing* (MLS) problem, we are given a set of N points in the plane, and want to find a pair of parallel lines within distance ε from each other such that the number of input points that intersect (stab) the area between the two lines, is maximized. Previously existing algorithms for this problem [18], [19] have large space requirements of $O(N^2)$. In the following, we present an algorithm that solves MLSP in time $O(N^2 \log N)$ and space $O(N)$.

How do we transform the problem of mass spectra recalibration to an instance of MLSP? Recall that A, B denote the sets of mass values. We define a set of points in the plane $S := \{(a, b) : a \in A, b \in B\}$ and try to find a line-pair that stabs a maximum number of points in S. By this, we construct a point set matching that allows many-to-many mappings of A to B. To exclude degenerate cases, we assume that scale $s \in [s_0, s_1]$ and translation $t \in [t_0, t_1]$ are bounded by some intervals. Then, we can restrict our set of points in the plane,

$$S := \big\{(a, b) : a \in A,\, b \in B,\, b \in [s_0 a + t_0 - \varepsilon, s_1 a + t_1 + \varepsilon]\big\}. \tag{3}$$

Nonetheless, the solution will in general not define a one-to-one mapping between A and B: for distinct $a, a' \in A$ and $b, b' \in B$ with $\left|a - a'\right| \ll \varepsilon$ and $\left|b - b'\right| \ll \varepsilon$, the optimal line-pair may stab all four points (a, b), (a, b'), (a', b), and (a', b').

Our solution is based on the *duality transform* of a set of points in the plane introduced by Brown [20]. The *dual* of a point $p = (p_x, p_y)$ in the plane is the line $p^* : y = p_x x - p_y$, while the dual of a line $q : y = q_x x + q_y$ is the point $q^* = (q_x, -q_y)$. The vertical distance between a point p and a line q equals the vertical distance between the line p^* and the point q^*. Furthermore, the dual transform maintains the above/below relationship between a point and a line. See e.g. [21, Chapter 8.2] for more details.

We now describe our solution to the MLS Problem. We are given a distance ε and a set $S \subseteq \mathbb{R}^2$ of points in the plane. In the following, the distance between two parallel lines is not

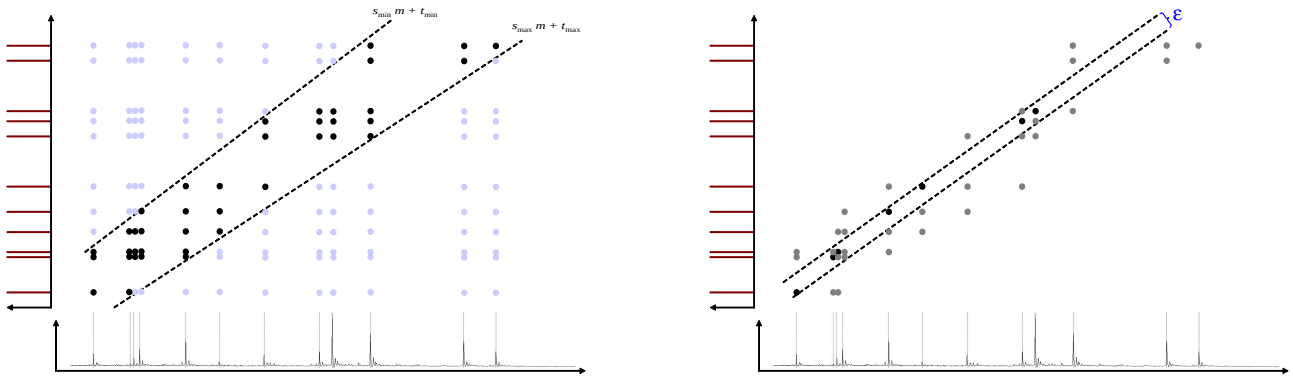

Fig. 1. On the geometric interpretation of the problem.

the Euclidean distance, but their vertical distance. Let us ignore vertical line pairs that can be handled separately. Given two parallel lines $q : y = q_x x + q_y$ and $q' : y = q_x x + q_y + \varepsilon$ then every line between q and q' must be parallel to q,q'. These lines, including q,q', are mapped to the line segment $p_x \times [-p_y - \varepsilon, -p_y]$ in the dual. Finding a line-pair that stabs a maximal number of points in S, is equivalent to finding a line segment $x \times [-y - \varepsilon, -y]$ such that the number of intersected lines in S' is maximal, over all choices of x and y. Note that the optimal line segment intersects the lines in S^* in some order, so there exists a first and a last line stabbed.

We iterate over all lines $p^* \in S^*$, and assume that p^* is the first line stabbed. Every other line $q^* : y = q_x x - q_y$ partitions p^* into a constant number of ranges as follows: only in the range between the intersection of p^* and q^*, and the intersection of q^* with the line parallel to p^* and at distance ε, can this line contribute to a line segment that stabs p^* first (see Fig. 2). Projection to the x-axis leads to the interval bounded by points $x = (p_y - q_y)/(p_x - q_x)$ and $x' = (p_y - q_y + \varepsilon)/(p_x - q_x)$. Attaching $+1$ or -1 to the endpoints depending on whether the endpoint is start or end of a range and then sorting these endpoints, one can scan through the endpoints keeping a counter how many ranges are active. The optimal choice for $(x, -y)$ corresponds to the overall largest count, and the line-pair $l : p_y = x p_x + y$ together with the parallel line l' at ε distance stabs a maximal number of points in S.

The above algorithm solves the point set matching problem in time $O(|S|^2 \log |S|)$ and, for unrestricted scale and translation, in time $O\big((mn)^2(\log m + \log n)\big)$. An interested reader may refer to Appendix for a solution to a more generic statement of the stabbing problem, that we summarize here:

Theorem 1 *Given a set C of variable size circles in the plane, we can find the line going through maximum number of them in time $O(|C|^2 \log |C|)$ and space $O(|C|)$. The maximum line-pair stabbing problem is a special case, and can be solved within the same time and space bounds.*

C. Topological Line-Sweep Algorithm

Consider the following variation of the line stabbing problem: let S be a set of N points in the plane. Identify two parallel lines with *minimal* distance that stab at least k points in S, for some fixed k (say, $k = 0.5N$). Based on the dual interpretation presented in the previous section, an algorithm to solve this problem can be based on *line-sweeping*: find a segment $x \times [y, y + \varepsilon]$ of *minimal length* that intersects k lines in the dual.[2] One endpoint of that segment has to be an intersection point of two lines in the dual plane: otherwise one can find a shorter segment slightly to the left or to the right. Sort the intersection points of all lines $p^* : y = p_x x - p_y$ in the dual plane, for $p \in S$. Now, we sweep the arrangement with a vertical line keeping an index array that represents the relative order of line segments. At every crossing point, we update in constant time the lengths of segments intersecting k lines, using this array. We also update the array in constant time. This algorithm solves the problem in time $O(N^2 \log N)$ and space $O(N)$ and was first proposed by Souvaine and Steele [22].

Edelsbrunner and Souvaine [23] and Chattopadhyay and Das [18] independently discovered a modification of the above algorithm that uses a *topological line-sweep* [24]. Here, the arrangement of lines is no longer swept with a straight line, but instead with a curve that intersects every line in exactly one point. This modification reduces the complexity of the algorithm to $O(N^2)$ time and $O(N)$ space. See also Rafalin et al. [25] for how to handle degenerated cases.

Souvaine and Steele [22] noted that the above method computes the Least Median of Squares (LMS) regression line for any k. LMS regression [26] is far more robust than other forms of regression such as Ordinary Least Squares. For the geometrical problem introduced in the previous section, the solution is also robust to outliers in the input. It is not clear in advance which of the two formulations is preferable, being more robust in application. Potentially, every one of the two methods can outperform the other on certain instances of the problem.

To apply the above method to our recalibration problem, transform the sets A, B into a set S of points in the plane as defined in the previous section. Then, this algorithm solves the modified point set matching problem (where we ask for a linear transformation that maps at least k points into minimal ε-distance)

[2]For a general overview of line-sweep based algorithms, see de Berg et al. [21, Chapter 2].

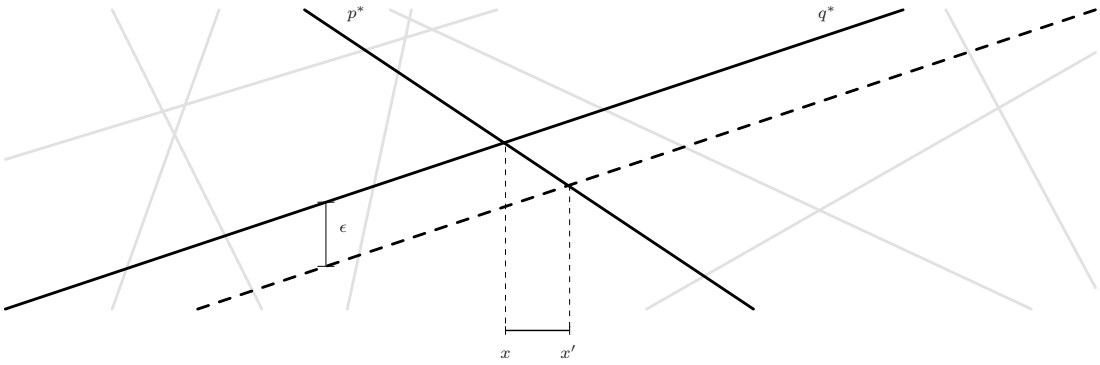

Fig. 2. Maximum line-pair stabbing algorithm: finding the interval where q^* can contribute to a segment starting in p^*.

in time $O(|S|^2)$ and, for unrestricted scale and translation, in time $O\big((mn)^2\big)$.

D. One-to-One Point Set Matching

All three algorithms presented above may produce a many-to-many matching between the original point sets. We opt, therefore, for another approach that considers only one-to-one matchings as feasible solutions.

Recall the voting algorithm in Sect. III-A: for each representative translation $t = b' - a'$, where $a' \in A$ and $b' \in B(\varepsilon)$, construct the set of scale ranges $R(a', b')$. Sort the endpoints of ranges in $R(a', b')$ into increasing order, and scan through them incrementing and decrementing a counter. The largest counter value is obtained at the optimum scale for the fixed translation, and repeating the process for all representative translations gives the overall optimum transformation f^*. The set E_{f^*} contains all $\{a, b\}$ with $|f^*(a) - b| \leq \varepsilon$.

The solution E_{f^*} obtained with that algorithm does not usually define a one-to-one matching between A and B. A brute-force algorithm to enforce a one-to-one mapping solution is as follows: at each phase of the previous algorithm that constructs sets E_f incrementally, let G_f be the bipartite graph with edge set E_f; solve the maximum matching problem on each G_f, and choose f corresponding to the overall largest maximum matching.

Notice that the graphs G_f change only by one edge at each incremental step (see Alt et al. [27], where this property is exploited in a more general setting). A property specific to our one-dimensional problem is that the maximum matchings on the graphs G_f can be computed greedily.[3] Let us introduce some notation to see why the greedy approach works here.

We represent our graph G_f using matrix notation analogous to edit matrices representing traces (see e.g. [30, Chapter 10]). Let $A = \{a_1, \ldots, a_m\}$ and $B = \{b_1, \ldots, b_n\}$ be the two point sets to be matched, and assume that the point sets are sorted in ascending order. Let us consider a fixed E_f such that $f(x) = s(x - a') + b'$, where $s = \frac{b - b'}{a - a'} \geq 0$. Consider a *match matrix* $M(1 \ldots m, 1 \ldots n)$ with $M(i, j) = \mathbf{1}$ if $\{a_i, b_j\} \in E_f$, and $M(i, j) = \mathbf{0}$ otherwise. It is easy to see (proof left for the reader) that the match matrix M is a *staircase matrix* (see Fig. 3):

[3]Compare to so-called *skis and skiers* folk theorem and to the linear time algorithm for minimum weight matching for points lying on a line [28]. See also [29] for more complete treatment of one-to-one, one-to-many, and other such types of matching without the added difficulty of geometric transformations.

(i) Each row of the matrix contains at most one *run* of **1**s, i.e. a maximal range of consecutive cells each containing value **1**.
(ii) Let i' and i, $i' < i$ be two rows containing a run of **1**s. Let the run at row i' cover indexes $c_{i'}, c_{i'} + 1, \ldots, d_{i'}$ and the run at row i cover indexes $c_i, c_i + 1, \ldots, d_i$. Then $c_{i'} \leq c_i$ and $d_{i'} \leq d_i$.

Notice that identical conditions on columns follow from (i) and (ii): M is a staircase matrix if and only if M^{T} is staircase matrix.

Recall the incremental algorithm that updates the graph G_f by scanning scales from left to right for a fixed translation. We represent the graph G_f as a match matrix M. Deleting or inserting an edge in G_f corresponds to updating the value of a cell in M. Since M is a staircase matrix at each scale, each update extends or reduces a run of **1**s at some row: we maintain pointers to the start and end of the run in each row. Updating these m pointers for all mn scales takes overall runtime $O(m^2 n)$ for testing whether a pointer can be moved, and time $O(mn)$ for moving every pointer to the end of the matrix. A greedy search for a matching at each scale takes time $O(m)$ (pick for every row the first column not yet matched). We leave it for the reader to see why this greedy approach gives the maximum matching at each step. Consequently, we obtain the following result:

Theorem 2 *The linear one-dimensional one-to-one point set matching problem on two point sets A and B of sizes m and n, respectively, can be solved in $O(m^3 n^2)$ time.*

An example of a series of transitions from scale to scale in the above algorithm is given in Fig. 3.

For practical considerations, recall that we already know some maximum limit for translation and scale in the calibration setting. These limits can be taken into account in the matching algorithms: instead of examining the whole transformation space of size $O(m^2 n^2)$, we can restrict to a small subset of it, consisting of all translation-scale pairs from $S \times S$ for S defined in (3). Moreover, converting range r_i (r_j) to range r_{i+1} (r_{j+1}) can be done incrementally by deleting points from the beginning and adding new points to the end of the old range until the new limit is satisfied. Thus, the time complexity is proportional to the size of the restricted transformation space $|S|^2$, multiplied by the time requirement of each step ($\log n$ for the many-to-many case and m for the one-to-one case).

In the one-to-one case, it is easy to improve the greedy algorithm that we execute at each examined transformation. For example, one can maintain information on each diagonal such

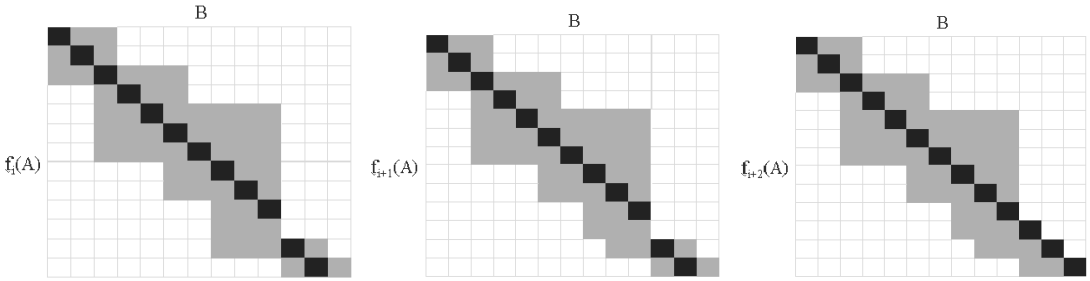

Fig. 3. The match matrix changes by one cell when moving from one scale to the next. The updates may influence the maximum matching found in the previous step, and hence the greedy algorithm is repeated in each step.

that the number of matches produced in a row at the current position in the diagonal, is known. This way the running time is proportional to the number of times diagonal are changed computing the greedy path. Still, in the worst case each step takes $O(m)$ time.

IV. Experiments

We implemented all algorithms in C++ for a restricted transformation space. For the topological line-sweep, we used the software library provided by Rafalin et al. [25], [31] that is capable of handling degenerate cases. We use three data sets to evaluate our approach:

1) SELDI mass spectrometry data from blood serum. This set consists of 20 mass spectra each containing about 20 mass peaks, picked by vendor software. We use one spectrum at the time as a reference, and compare it against all others. Limits for scale and translation are $s \in [0.999, 1.001]$, $t \in [-10, 10]$, and $\varepsilon = 2.5$ Da.
2) MALDI-TOF Protein Mass Fingerprint mass spectra for the sample organism *Corynebacterium glutamicum* using tryptic digestion. The protein database consists of 3501 protein sequences and reference spectra contain about 24 peaks. We use a set of 316 sample spectra each containing about 20 peaks. Limits are $s \in [0.999, 1.001]$, $t \in [-5, 5]$, and $\varepsilon = 0.75$ Da.
3) Two data sets of MALDI-TOF DNA mass spectra from RNAse A digest [32]. The two data sets contained a total of 208 reference spectra with about 64 peaks each, and 1511 sample spectra with 84 peaks each. In the following, we will treat these two data sets as one. Limits are $s \in [0.999, 1.001]$, $t \in [-5, 5]$, and $\varepsilon = 1.25$ Da.

Peaks are extracted from sample spectra using vendor software. No recalibration is executed for PMF and DNA mass spectra pairs where sample spectrum and predicted spectrum show five or less "common" peaks with mass inaccuracy as introduced above; that is, $|S| \geq 5$ must hold for S from (3). For example, about 10% or $130\,000$ PMF mass spectra pairs are recalibrated. Regarding the topological line-sweep, we search for line-pairs that stab $50\,\%$ of the points in S. For point set matching, the ε-values are reported above. For line-pair stabbing, we use a line-pair with fixed distance 2ε.

Note that the choice of parameters used for recalibration can be seen as a worst-case scenario: in fact, all data sets were of better quality. Choosing these parameters, we want to assure that all methods work fine for data of poor quality, even though a large number of wrong peak assignments is inevitable in this case.

We report runtimes of the three methods on a 900 MHz Ultra-Sparc III processor in Table I. There was no significant difference in calibration accuracy of the three approaches. The line-pair stabbing algorithm and the topological line-sweep algorithm show comparable performance, with slight advantages for the former. Recalibration using one-to-one point set matching leads to tenfold runtimes, but is fast enough for high throughput analysis of SELDI and PMF mass spectrometry data.

Let us concentrate on the SELDI test set with fixed ε distance. We first compute a peak matching between spectra using the one-to-one point set matching algorithm. We compute the empirical distribution (density) of mass differences before and after the linear mapping. Next, we use Chebyshev approximation to find the polynomial of degree two that minimizes the maximum distance of input pair values [33, min-max approximation], and compute the distribution after applying the new calibration polynomials. These three distributions are shown in Fig. 4. As can be seen, the distribution becomes significantly more focused after linear transformation, and even more focused after polynomial fitting, mostly because of avoiding the artificial effect of the fixed ε. This shows the validity of linear transformations for recalibration.

We test the weaker model of a zeroth order polynomial, to verify whether a constant shift can lead to a good mapping (Fig. 5), plotting the results of Chebyshev fitting with polynomials of degree $0, 1$ and 2 using the same input pairs as in the previous paragraph. Here, the polynomials of degrees zero and one map sample masses to reference masses and, hence, are not restricted to TOF mass spectra but can be applied to any type of mass spectrometry data. This shows that recalibration using a constant shift leads to unsatisfactory results.

We also tested the impact of the one-to-one mapping criterion compared to the other two algorithms. As we already restricted the search space to allow only small translations and scales, the many-to-many mapping case did not have degenerate solutions. In fact, the size of the many-to-many mapping was only $0.94\,\%$ greater than the size of the one-to-one mapping. Nonetheless, one-to-one mappings may help to avoid such degenerate cases for automated recalibration of high-throughput data.

Our tests confirmed that excluding outliers is mandatory for accurate recalibration: of the mass pairs initially accepted for recalibration in S, only 10–$20\,\%$ are used in a linear recalibration. These results demonstrate that the data really contains outliers and one must not use algorithms that are sensitive to them. We are not aware of efficient Chebyshev fitting algorithms that could allow some percentage of outliers. We note that the SELDI data set contained some spectra of very low quality, and also that recalibration performance can probably be improved using more

	SELDI spectra	PMF spectra	DNA spectra
number of recalibrations	166	129408	156097
line-pair stabbing	0.225 ms	0.315 ms	2.237 ms
topological line-sweep	0.343 ms	0.397 ms	2.959 ms
1-1 point set matching	2.325 ms	4.204 ms	67.470 ms

TABLE I

RUNTIMES PER RECALIBRATION IN MILLISECONDS, MEASURED ON A 900 MHZ ULTRASPARC III PROCESSOR. SEE TEXT FOR DETAILS.

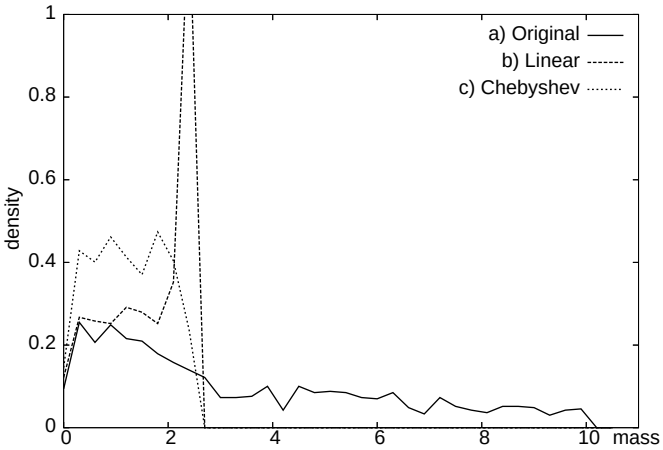

Fig. 4. Comparison of distributions: a) original differences, b) differences after initial linear mapping, c) differences after Chebyshev fitting. Absolute mass difference in Da on x-axis, empirical distribution on y-axis.

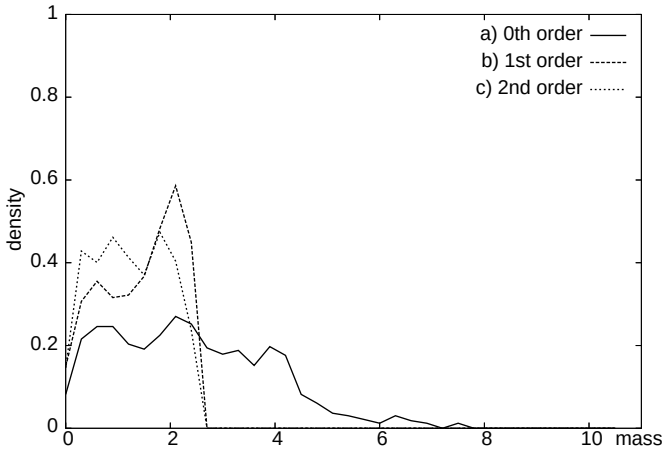

Fig. 5. Comparison of distributions using different order polynomials in Chebyshev fitting. Absolute mass difference in Da on x-axis, empirical distribution on y-axis.

than 20 peaks picked by vendor software. But the fundamental problems remain the same (outliers) or are even more pressing (wrongly assigned peak pairs) for larger peak lists.

The comparison of many-to-many and one-to-one mappings suggest that many-to-many mapping usually suffices for recalibration, despite the heuristic choice of restrictions for translation and scale, and reduces the runtime of the algorithms significantly. After obtaining the mapping using linear transformations, it makes sense to find the best first or second order polynomial using, say, Chebyshev fitting, to make the final mapping as accurate as possible. This operation is possible because the outliers are detected in the first phase.

V. CONCLUSION

We studied the problem of recalibrating mass spectra and proposed a two-step procedure for this task. The first step uses a linear function to compute a mapping between masses; we described efficient combinatorial algorithms for executing this step that are robust to outliers. The second step uses known methods for polynomial fitting given the input pairs that contain no outliers. The two-step procedure is motivated by the fact that the mass errors are "almost linear," and a robust and fast linear fitting insensitive to outliers can work as a good estimate. Our experiments provide evidence that this observation is valid in practice. One can easily fine-tune these methods by taking into account non-unit weights of different masses to focus recalibration on prominent peaks in the sample spectrum, or use mass deviations $\varepsilon(m)$ that depend on the mass m to be recalibrated.

We also studied the recalibration of TOF mass spectra and showed how "second-order" polynomials can be used for this task. In principle, an optimal solution can be found by constructing a polynomial of degree two that maps time-of-flight values in the sample spectrum to mass values in the reference spectrum. This construction must be robust to outliers. Iterative algorithms for computing such outlier-sensitive mappings exist, but have high running time.

The recalibration procedures described herein are currently integrated into an analysis pipeline for the identification of proteins using Protein Mass Fingerprint (to be reported elsewhere). We plan to make the procedures available for database searching and *de novo* sequencing approaches using using tandem mass spectrometry. We also want to use our methods for the recalibration of photoionization mass spectra of flames [34].

As future work we want to develop faster combinatorial outlier-sensitive polynomial-fitting algorithms, applicable to data analysis problems in all fields of science. Another interesting future objective is to improve the simple greedy algorithm in Sect. III-D as the algorithm does not take any advantage of the fact that the staircase matrix changes in a well-structured manner at each incremental step.

APPENDIX
EXTENSIONS OF LINE-PAIR STABBING

There are three dual ways to describe the *Maximum Line-Pair Stabbing Problem* (see Fig. 6):

- Given a set of points, find a pair of parallel lines within vertical distance ε from each others such that the number of points in the closed slab between the lines is maximum. For short, the resulting line-pair is said to stab the maximum number of points.
- Given a set of circles with diameter ε, find a line that that stabs the maximum number of them.
- Given a set of lines, find a vertical line segment starting at point (x, y) of length $\delta(x)$ that crosses maximum number of lines. (Function δ will be defined later.)

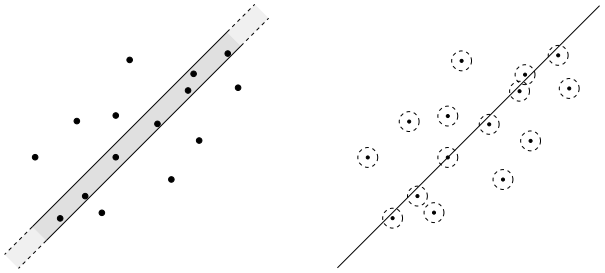

Fig. 6. Stabbing points with a line-pair (left) and stabbing circles with a line (right).

Let us consider the generalization of the maximum line-pair stabbing problem to variable size circles. This variation does not have an interpretation as a line-pair stabbing problem but is a natural extension of the second dual problem mentioned above. It can also be interpreted as the third dual problem, which is fortunate, since that allows us to extend our solution given in Sect III-B. Note that this generalization is of interest in the setting of mass spectra recalibration: using circles of different size, one can use different levels of inaccuracies for different masses, such as a mass difference that is linear in mass, or any other function.

We let $C = \{(c_x, c_y, c_r) \mid c_x, c_y, c_r \in \mathbb{R}, c_r > 0\}$ be a set of *circles*, where (c_x, c_y) gives the center point and c_r the radius of the circle $c \in C$. Our goal is to find a line $\ell : y = mx + b$ that stabs a maximum number of circles in C. Notice that there exists a line ℓ that is *tangential* to at least two circles, say $c, d \in C$: ℓ is at distance c_r from (c_x, c_y) and at distance d_r from (d_x, d_y). Each pair of circles defines exactly four such tangential lines (this is not the case if one circle is inside the other, but this case will be handled automatically anyway). Let T be the set consisting of the tangential lines over all pairs of circles. Comparing each line $\ell \in T$ to each circle in C gives a naive algorithm with time complexity $O(|C|^3)$. We show next how to do the same in $O(|C|^2 \log |C|)$ time.[4]

Consider the third dual interpretation of the problem. The image of circle $c \in C$ is the following: its center (c_x, c_y) is mapped into a line $c^* : y = c_x x - c_y$, each of its tangentials with angle in $(-\pi/2, \pi/2)$ is mapped into a point $(x, c_x x - c_y + \delta_{c_r}(x))$, and each of its tangentials with angle in $(\pi/2, 3\pi/2)$ is mapped into a point $(x, c_x x - c_y - \delta_{c_r}(x))$. Functions δ_{c_r} are defined as $\delta_{c_r}(m) = c_r\sqrt{1^2 + m^2}$, i.e. the fixed radius ε is replaced by the variable radius c_r.

Let C^* denote the set of lines that are the dual images of the circle centers in C. We associate to each $d^* : y = d_x x - d_y$ in C^* an *area*: the area $U(d^*)$ of d^* is the area in between the functions $d_x x - d_y + \delta_{d_r}(x)$ and $d_x x - d_y - \delta_{d_r}(x)$. Our objective in the dual interpretation is the following. Search for a point p such that $p = (m, c_x m - c_y + \delta_{c_r}(m))$ or $p = (m, c_x m - c_y - \delta_{c_r}(m))$ for some line $c^* : y = c_x x - c_y$ in C^*, and that intersects maximum number of areas $U(d^*)$ for $d^* \in C^*$. Such $p = (m, -b)$ represents the optimal (tangential) line $\ell : y = mx + b$ in the primal problem.

After the above dual interpretation, the computation is almost identical to the previously described. Consider two lines $c^* : y = c_x x - c_y$ and $d^* : y = d_x x - d_y$ in C^*, and the case where $(m, -b)$ is chosen so that $-b = c_x m - c_y$. This time we need to solve

$$c_x m - c_y \pm \delta_{c_r}(m) = d_x m - d_y \pm \delta_{d_r}(m), \tag{4}$$

for all combinations of fixing $\pm$. This will give us the ranges $R^+ \subseteq \mathbb{R}$ such that $m \in R^+$ if and only if point $(m, c_x m - c_y + \delta_{c_r}(m))$ is included in the area $U(d^*)$. Similar ranges $R^- \subseteq \mathbb{R}$ can be computed for point $(m, c_x m - c_y - \delta_{c_r}(m))$. Now, let $\mathbf{R}^+(c^*)$ and $\mathbf{R}^-(c^*)$ be the multisets of ranges for line c^* formed by repeating the above process for each $d^* \in C^*$. We process the sets independently. After sorting the set, we again find in linear time the points m^+ and m^- intersecting maximum number of ranges in $\mathbf{R}^+(c^*)$ and $\mathbf{R}^-(c^*)$, respectively. One of the points, $(m^+, c_x m^+ - c_y + \delta_{c_r}(m^+))$ or $(m^-, c_x m^- - c_y - \delta_{c_r}(m^-))$, represents the optimal tangential line in the primal problem for the chosen $c \in C$. Repeating the process on each $c \in C$ gives the optimum overall solution. The running time and space usage are $O(|C|^2 \log |C|)$ and $O(|C|)$, respectively.

Notice that there are some special cases when the algorithm does not work properly. In the case that the solution is a vertical line, one can proceed naively testing all vertical lines crossing a circle boundary and counting how many other circles they intersect. If a circle is totally inside another one, they do not have common tangentials. This is no problem if there is another circle outside, since any tangential to the inner-most circle will go through also those circles that contain it. If there is no such outside circle, one can choose any line going through the inner-most. This concludes the proof of Theorem 1.

[4]To the best of our knowledge, the topological line sweep of Souvaine and Steel [22] for Least Median of Squares cannot be generalized to variable circle sizes without introducing an additional $\log |C|$ factor to the runtime for doing binary search. So, the overall runtime of this method would be $O(|C|^2 \log |C|)$, too.

Acknowledgment

We wish to thank Tobias Marschall and Marcel Martin for implementing the algorithms and running the experiments. We are also grateful to the anonymous reviewers of their devotedness to improving the readability and overall coherency. The authors were supported by "Deutsche Forschungsgemeinschaft" (BO 1910/1-1) within the Computer Science Action Program. Second author was also partially supported by the Academy of Finland.

References

[1] W. J. Henzel, C. Watanabe, and J. T. Stults, "Protein idendification: The origins of peptide mass fingerprints," *J. Am. Soc. Mass Spectr.*, vol. 14, pp. 931–942, 2003.

[2] R. Matthiesen, M. B. Trelle, P. Hojrup, J. Bunkenborg, and O. N. Jensen, "VEMS 3.0: algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins." *J Proteome Res*, vol. 4, no. 6, pp. 2338–2347, 2005. [Online]. Available: http://dx.doi.org/10.1021/pr050264q

[3] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie, "PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry," *Rapid Commun. Mass Spectrom.*, vol. 17, no. 20, pp. 2337–2342, 2003.

[4] B.-L. Adam, Y. Qu, J. W. Davis, M. D. Ward, M. A. Clements, L. H. Cazares, O. J. Semmes, P. F. Schellhammer, Y. Yasui, Z. Feng, and G. L. Wright, jr., "Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men," *Cancer Research*, vol. 62, pp. 3609–3614, 2002.

[5] J. Gobom, M. Mueller, V. Egelhofer, D. Theiss, H. Lehrach, and E. Nordhoff, "A calibration method that simplifies and improves accurate determination of peptide molecular masses by MALDI-TOF MS," *Anal. Chem.*, vol. 74, no. 15, pp. 3915–3923, 2002.

[6] K. A. Baggerly, J. S. Morris, and K. R. Coombes, "Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments," *Bioinformatics*, vol. 20, no. 5, pp. 777–785, 2004.

[7] O. J. Semmes, Z. Feng, B.-L. Adam, L. L. Banez, W. L. Bigbee, D. Campos, L. H. Cazares, D. W. Chan, W. E. Grizzle, E. Izbicka, J. Kagan, G. Malik, D. McLerran, J. W. Moul, A. Partin, P. Prasanna, J. Rosenzweig, L. J. Sokoll, S. Srivastava, S. Srivastava, I. Thompson, M. J. Welsh, N. White, M. Winget, Y. Yasui, Z. Zhang, and L. Zhu, "Evaluation of serum protein profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry for the detection of prostate cancer: I. assessment of platform reproducibility," *Clin. Chem.*, vol. 51, pp. 102–112, 2005.

[8] M. W. Bern and D. Goldberg, "EigenMS: De novo analysis of peptide tandem mass spectra by spectral graph partitioning," in *Proc. of RECOMB 2005*, ser. Lect. Notes Comput. Sc., vol. 3500. Springer, 2005, pp. 357–372.

[9] E. Cheney, *An Introduction to Approximation Theory*, 2nd ed. Amer. Mathematical Society, 2000, reprint of 1982 edition.

[10] J. W. Wong, G. Cagney, and H. M. Cartwright, "SpecAlign—processing and alignment of mass spectra datasets," *Bioinformatics*, vol. 21, no. 9, pp. 2088–2090, 2005.

[11] W. E. Wolski, M. Lalowski, P. Jungblut, and K. Reinert, "Calibration of mass spectrometric peptide mass fingerprint data without specific external or internal calibrants," *BMC Bioinformatics*, vol. 6, p. 203, 2005.

[12] N. Jeffries, "Algorithms for alignment of mass spectrometry proteomic data," *Bioinformatics*, vol. 21, no. 14, pp. 3066–3073, 2005.

[13] R. Matthiesen, J. Bunkenborg, A. Stensballe, O. N. Jensen, K. G. Welinder, and G. Bauw, "Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V2.0." *Proteomics*, vol. 4, no. 9, pp. 2583–2593, Sep 2004. [Online]. Available: http://dx.doi.org/10.1002/pmic.200300792

[14] A. Wool and Z. Smilansky, "Precalibration of matrix-assisted laser desorption/ionization-time of flight spectra for peptide mass fingerprinting." *Proteomics*, vol. 2, no. 10, pp. 1365–1373, 2002. [Online]. Available: http://dx.doi.org/3.0.CO;2-9

[15] K. R. Clauser, P. Baker, and A. L. Burlingame, "Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching." *Anal Chem*, vol. 71, no. 14, pp. 2871–2882, Jul 1999.

[16] V. Egelhofer, K. Bssow, C. Luebbert, H. Lehrach, and E. Nordhoff, "Improvements in protein identification by MALDI-TOF-MS peptide mapping." *Anal Chem*, vol. 72, no. 13, pp. 2741–2750, Jul 2000.

[17] T. J. Rivlin, *An Introduction to the Approximation of Functions*. New York: Dover, 1981, reprint of 1969 edition.

[18] S. Chattopadhyay and P. Das, "The K-dense corridor problems," *Pattern Recogn. Lett.*, vol. 11, no. 7, pp. 463–469, 1990.

[19] F. Y. Chin, C. A. Wang, and F. L. Wang, "Maximum stabbing line in 2D plane," in *Proc. of COCOON 1999*, ser. Lect. Notes Comput. Sc., vol. 1627. Springer, 1999, pp. 379–388.

[20] K. Q. Brown, "Geometric transforms for fast geometric algorithms," Report CMUCS-80-101, Dept. Comput. Sci., Carnegie-Mellon Univ., Pittsburgh, USA, 1980.

[21] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf, *Computational Geometry: Algorithms and Applications*, 2nd ed. Springer, 2000.

[22] D. L. Souvaine and J. M. Steele, "Time- and space-efficient algorithms for least median of squares regression," *J. Am. Stat. Assoc.*, vol. 82, no. 399, pp. 794–801, 1987.

[23] H. Edelsbrunner and D. L. Souvaine, "Computing least median of squares regression lines and guided topological sweep," *J. Am. Stat. Assoc.*, vol. 85, no. 409, pp. 115–119, 1990.

[24] H. Edelsbrunner and L. J. Guibas, "Topologically sweeping an arrangement." *J. Comput. Syst. Sci.*, vol. 38, no. 1, pp. 165–194, 1989.

[25] E. Rafalin, S. Souvaine, and I. Streinu, "Topological sweep in degenerate cases," in *Proc. 4th Workshop on Algorithm Engineering and Experiments (ALENEX 2002)*, ser. LNCS, vol. 2409, 2002, pp. 577–588.

[26] P. J. Rousseeuw, "Least median of squares regression," *J. Am. Stat. Assoc.*, 1984.

[27] H. Alt, K. Mehlhorn, H. Wagener, and E. Welzl, "Congruence, similarity and symmetries of geometric objects," *Discrete Comput. Geom.*, vol. 3, no. 3, pp. 237–256, 1988.

[28] R. Karp and S. Li, "Two special cases of the assignment problem," *Discrete Mathematics*, vol. 13, no. 2, pp. 129–142, 1975.

[29] J. Colannino, M. Damian, F. Hurtado, J. Iacono, H. Meijer, S. Ramaswami, and G. Toussaint, "An $O(n \log n)$-time algorithm for the restriction scaffold assignment problem," *Journal of Computational Biology*, vol. 13, no. 4, pp. 979–989, 2006. [Online]. Available: http://www.liebertonline.com/doi/abs/10.1089/cmb.2006.13.979

[30] D. Sankoff and J. B. Kruskal, Eds., *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Reading, Mass.: Addison-Wesley, 1983.

[31] E. Rafalin, "LMS regression using guided topological sweep in degenerate cases," Software library available at http://www.cs.tufts.edu/research/geometry/lms/, 2002.

[32] M. Lefmann, C. Honisch, S. Boecker, N. Storm, F. von Wintzingerode, C. Schloetelburg, A. Moter, D. van den Boom, and U. B. Goebel, "A novel mass spectrometry based tool for genotypic identification of mycobacteria," *J. Clin. Microbiol.*, vol. 42, no. 1, pp. 339–346, 2004.

[33] T. J. Rivlin, *Chebyshev Polynomials: From Approximation Theory to Algebra and Number Theory*. New York: Wiley-Interscience, 1990.

[34] C. A. Taatjes, N. Hansen, A. McIlroy, J. A. Miller, J. P. Senosiain, S. J. Klippenstein, F. Qi, L. Sheng, Y. Zhang, T. A. Cool, J. Wang, P. R. Westmoreland, M. E. Law, T. Kasper, and K. Kohse-Höinghaus, "Enols are common intermediates in hydrocarbon oxidation," *Science*, vol. 308, no. 5730, pp. 1887–1889, 2005.

Sebastian Böcker did his PhD studies in mathematics at Bielefeld University, and received his PhD degree in 1999 on a topic related to computational phylogeny. He then joined the company SEQUENOM and worked in industry for three years, both in the Hamburg subsidiary and the San Diego headquarters. In 2003, he became head of the DFG-funded research group Computational Mass Spectrometry at Bielefeld University. Since 2006, Dr. Böcker holds the chair of bioinformatics at Friedrich-Schiller-University Jena.

Veli Mäkinen did his PhD studies in computer science at University of Helsinki, and received his PhD degree in 2003 on a topic related to combinatorial pattern matching. During 2004-2005 he worked as a postdoctoral researcher in the Computational Mass Spectrometry research group at Bielefeld University. Supported by a grant from Academy of Finland, he is now working as a postdoctoral researcher in the From Data to Knowledge (FDK) research unit at University of Helsinki. Since 2006, Dr. Mäkinen holds an Adjunct Professor position at University of Helsinki.