Combinatorial Approaches for Mass Spectra Recalibration

Sebastian Böcker and Veli Mäkinen*

AG Genominformatik, Technische Fakultät, Universität Bielefeld
PF 100 131, 33501 Bielefeld, Germany
Contact: boecker@CeBiTec.uni-bielefeld.de

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.

1 Introduction

Mass spectrometry is one of the most popular analysis techniques in the emerging field of Systems Biology: The analysis of Protein Mass Fingerprints and tandem mass spectra for protein identification and de novo sequencing 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. 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 [1]. 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.

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 [2]. For a set of SELDI-TOF mass spectra, usually an arbitrary spectrum from the set is used as the reference spectrum.

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

* Currently at Department of Computer Science, University of Helsinki, Finland.

Dagstuhl Seminar Proceedings 05471
Computational Proteomics
http://drops.dagstuhl.de/opus/volltexte/2006/545

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: We 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 describe three combinatorial, deterministic methods for the efficient and robust identification of outliers using linear transformations

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 solved using known techniques from approximation theory or statistics.

2 Physics of Mass Spectrometry and recalibration

A mass spectrometer cannot determine the masses of sample molecules directly but only measure a derived physical property, such as voltages U, V for quadrupole instruments, 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.

The concept of *recalibrating* mass spectra is to use hypothetical knowledge of the investigated sample, to compute a more accurate calibration function. For example, if we are given a database of proteins and have to decide what protein fits best the measured mass spectrum, 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: Many simulated spectra can show some similarity, if we take into account 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.

Calculating a calibration polynomial is possible as long as the exact masses of all sample molecules are known. Then we can use methods from approximation theory and statistics, such as Ordinary Least Squares, 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.

3 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 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 $|E_f|$ is maximum, where E_f is the edge set of a bipartite graph on A, B such that $\{a, b\} \in E_f$ iff $|f(a) - b| \leq \varepsilon$. Note that some $a \in A$ can be mapped into ε-distance of several $b \in B$ and vice versa: In fact, on 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. However, a more rigorous way to define the problem is to search for E_f that contains the largest one-to-one mapping, see Section 3.4.

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.

3.1 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, one observes 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, we construct all E_f for $f \in F$ incrementally, and choose the f that corresponds to the largest $|E_f|$: 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.

3.2 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 [3] we are given a set of N points in the plane, and we want to find a pair of parallel lines within distance ε from each other such that the number of input points in that intersect (stab) the area between the two lines, is maximized. Previously existing algorithms for this problem [4,5] have large space requirements of $O(N^2)$. In [3] 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,

$$S := \big\{(a,b) \,:\, a \in A,\, b \in B, \text{ and } b \in [s_0 a + t_0 - \varepsilon, s_1 a + t_1 + \varepsilon]\big\}. \tag{1}$$

Still, 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 $|a - a'| \ll \varepsilon$ and $|b - b'| \ll \varepsilon$, the optimal line may stab all four points (a,b), (a,b'), (a',b), and (a',b').

Our solution to the MLS Problem is based on the following idea: We are given a set $S \subseteq \mathbb{R}^2$ of points in the plane, and a distance ε. In the following, the distance between two parallel lines is not the Euclidean distance, but their distance at the y-axis. We ignore vertical line pairs that can be handled separately. We use the dual S' of the point set S by mapping each point $p = (p_x, p_y) \in S$ to a line $p^* : y = p_x x - p_y$. Here, 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, compute all ranges where a second line is in ε-distance, and finally sort the endpoints of these ranges. Then, we can scan through the endpoints keeping a counter how many ranges are active. See [3] for details where we consider the more complicated case that the two parallel lines have constant *Euclidean* distance. This algorithm solves the point set matching problem in time $O(|S|^2 \log |S|)$ and, for unrestricted scale and translation, in time $O\big((mn)^2(\log m + \log n)\big)$.

3.3 Topological Line-Sweep Algorithm

Consider the following variation of the line stabbing problem: We are given a set S of N points in the plane, and we search for to 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 a *line-sweep*. Souvaine and Steel [6] proposed such an algorithm that solves the problem in time $O(N^2 \log N)$ and space $O(N)$. In [4,7] the authors independently discovered a modification of the above algorithm that uses the *topological line-sweep* of [8]. 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. The above method computes the least median of squares (LMS) regression line for $k = 0.5N$ [6] . LMS regression is far more robust than other forms of regression such as Ordinary Least Square.

To apply the above method to our recalibration problem, we 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) in time $O(|S|^2)$ and, for unrestricted scale and translation, in time $O\big((mn)^2\big)$.

3.4 One-to-One Point Set Matching

The solution E_f obtained with the above algorithms usually does not define a (one-to-one) matching between A and B. A brute-force algorithm to solve the one-to-one mapping case 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. An existing approach [9] result in an algorithm with worst case runtime $O((mn)^3)$.

Our problem has an extra property that allows a more efficient way to find the maximum matchings. This property is basically a consequence of the sequential nature of the data; there is always an optimal non-crossing matching. To obtain a faster algorithm we observe that a simple greedy algorithm that matches points in sequential order to the first available one is sufficient. The greedy algorithm runs in $O(m)$ time. We omit the further details of this approach and refer the reader to [3].

For practical considerations, recall that in the calibration setting we already know some maximum limit for translation and scale. These limits can be taken into account in the point-set matching algorithms: Instead of examining the whole transformation space we can restrict to a small subset of it. Then, the time complexity is proportional to the size of the restricted transformation space, multiplied by the time requirement of each matching step ($\log n$ for the many-to-one case and m for the one-to-one case).

4 Experiments

We implemented the algorithms in C++ with restricted transformation space. For the topological line-sweep, we used the software library [10] that is capable of handling degenerate cases. We use three data sets to evaluate our approach: (i) SELDI mass spectrometry data from blood serum. This set consists of 20 mass spectra each containing about 20 mass peaks. (ii) 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. (iii) Two data sets of MALDI-TOF DNA mass spectra from RNAse A digest. The two data sets contained a total of 208 reference spectra with about 64 peaks each, and 1511 sample spectra with 84 peaks each.

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 (1). 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 2.5, 0.75, 1.25 are used for the three data sets. For line-pair stabbing, we use a line-pair with fixed distance 2ε.

	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 1. Runtimes per recalibration in milliseconds, measured on a 900 MHz UltraSparc III processor.

We report runtimes of the three methods on a 900 MHz UltraSparc III processor in Table 1. There was no significant difference in calibration accuracy of the three approaches. As one can see, 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.

We also evaluated the validity of linear transformations for recalibration: We found that the mass difference distribution becomes significantly more focused after linear transformation, and even more focused after fitting with polynomials of degree two. In comparison, using the weaker model of a zeroth order polynomial we found that a constant shift cannot lead to a good mapping.

We found 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. This demonstrates that the data really contains outliers and one cannot use algorithms that are sensitive to them. The comparison of many-to-one and one-to-one mappings suggest that computing a many-to-one mapping usually is enough for recalibration.

5 Conclusion

We studied the problem of recalibrating mass spectra, and proposed a two-step procedure for this task: In the first step, we use a linear function to compute a mapping between masses. We described efficient combinatorial algorithms for executing this step that are robust to outliers. In the second step, we use 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 give evidence that this observation is valid in practice. We also studied the recalibration of TOF mass spectra and found that second order polynomials can be used for this task.

Acknowledgment

We wish to thank Tobias Marschall and Marcel Martin for implementing the algorithms and running the experiments. The authors were supported by "Deutsche Forschungsgemeinschaft" (BO 1910/1-1) within the Computer Science Action Program.

References

1. Gobom, J., Mueller, M., Egelhofer, V., Theiss, D., Lehrach, H., Nordhoff, E.: A calibration method that simplifies and improves accurate determination of peptide molecular masses by MALDI-TOF MS. Anal. Chem. **74** (2002) 3915–3923
2. Bern, M.W., Goldberg, D.: EigenMS: De novo analysis of peptide tandem mass spectra by spectral graph partitioning. In: Proc. of RECOMB 2005. Volume 3500 of Lect. Notes Comput. Sc., Springer (2005) 357–372
3. Böcker, S., Mäkinen, V.: Maximum line-pair stabbing problem and its variations. In: Proc. of European Workshop on Computational Geometry (EWCG 2005), Eindhoven, Netherlands (2005) 183–186
4. Chattopadhyay, S., Das, P.: The K-dense corridor problems. Pattern Recogn. Lett. **11** (1990) 463–469
5. Chin, F.Y., Wang, C.A., Wang, F.L.: Maximum stabbing line in 2D plane. In: Proc. of COCOON 1999. Volume 1627 of Lect. Notes Comput. Sc., Springer (1999) 379–388
6. Souvaine, D.L., Steele, J.M.: Time- and space-efficient algorithms for least median of squares regression. J. Am. Stat. Assoc. **82** (1987) 794–801
7. Edelsbrunner, H., Souvaine, D.L.: Computing least median of squares regression lines and guided topological sweep. J. Am. Stat. Assoc. **85** (1990) 115–119
8. Edelsbrunner, H., Guibas, L.J.: Topologically sweeping an arrangement. J. Comput. Syst. Sci. **38** (1989) 165–194
9. Alt, H., Mehlhorn, K., Wagener, H., Welzl, E.: Congruence, similarity and symmetries of geometric objects. Discrete Comput. Geom. **3** (1988) 237–256
10. Rafalin, E.: LMS regression using guided topological sweep in degenerate cases. Software library available at http://www.cs.tufts.edu/research/geometry/lms/ (2002)