Team 4: Identifying Sugars

P.Du[*], A.Glazyrina[†], O.Gutiérrez-Navarro[‡],
A.Howse[§], A.Moorthy[¶], I.Vukićević[‖],
Mentor: A.Kearsley[**]

June 29, 2012

1 Introduction

Gas chromatography-mass spectrometry is an analytical technique used to represent a given chemical compound by a spectrum of mass-to-charge ratios of varying intensities. The spectrum is represented by plotting the mass-to-charge ratios along the x-axis and the intensities along the y-axis. Such spectra are used to determine the identiy of unknown substances by comparisons to a library of known spectra.

Due to the possible chemical or mechanical errors which may occur when calculating a mass spectrum, this matching process can be very difficult. In particular, a mass spectrometer used to test unknown substances may produce a spectrum with a much lower resolution compared to the documented spectra in the library. This allows error to occur along both the x-axis and the y-axis, that is, in both the mass-to-charge readings and the recorded intensities.

The matching process is especially difficult in the case of sugars as their spectra share many common features, as illustrated in Figure 1. Based on past observations, we assume the worst case scenario for error when measuring a sugar is the observed mass-to-charge ratio being shifted by at most three positions in either direction and the intensity varying by as much as $\pm 20\%$.

This paper will analyse the existing techniques for comparing sugars and give suggestions of alternative methods that should improve upon the current methods in use.

1.1 Current Techniques

There is a wide range of literature published on the problem of compound identification using mass spectrometry. There are two main ways to match an experimental spectrum to one in the library database: reverse library search, which seeks evidence of a specific compound in the spectrum of the unknown, and forward search, which compares the unknown spectrum to the reference library in order to find the best match. The former, considered the slower approach, is utilized in probability-based matching systems [2, 6, 8] and will

[*]University of Alberta
[†]University of Alberta
[‡]Universidad Autonoma de San Luis Potosi
[§]Memorial University of Newfoundland
[¶]University of Guelph
[‖]Columbia University
[**]National Institute of Standards and Technology

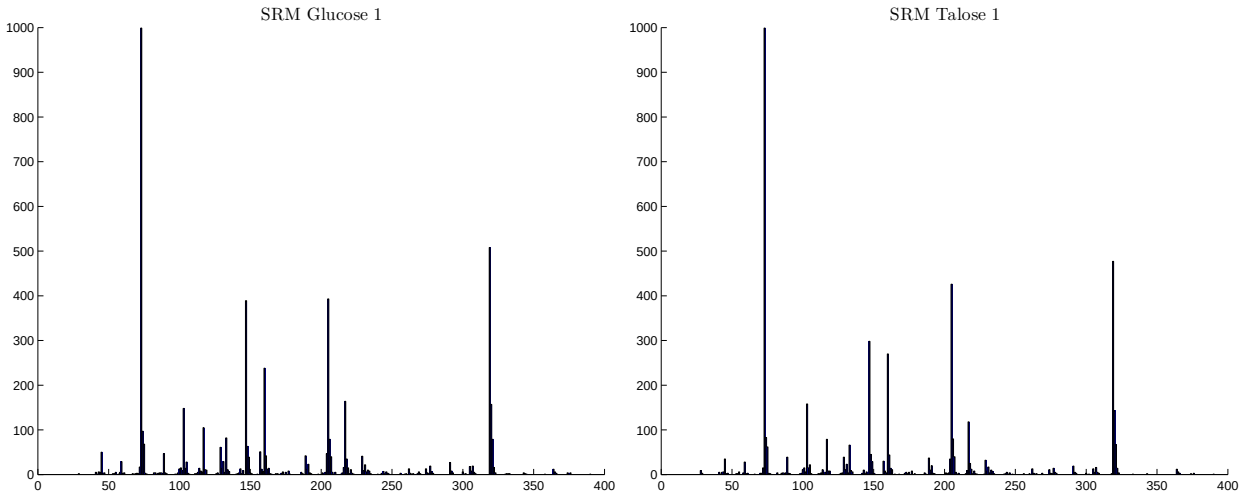

Figure 1: Mass spectra of a glucose and a talose standard reference material sample provided by NIST. Note that the structures are very similar and that the maximum occurs at $x = 73$ for both.

not be discussed in this paper. The latter uses a particular set of rules to compare the unknown to each compound in the library. Matching is assessed by computing various similarity measures between the spectra. These similarity measures are the subject of much of the research in this field.

An extensively used similarity measure is the cosine correlation between two sequences of characteristics representing the two mass spectra [3, 5, 9–12]. Most often, weighted peak intensities are taken as vectors of such characteristics. The drawback of this approach is that the choice of the weights is often determined by trial and error instead of theoretical justification, which leads to significant variations in the weights used by different researchers.

Also used within similarity measures are various distance metrics: Euclidean distance is commonly used in k-nearest neighbour methods [7]; normalized Euclidean distance and absolute value distance between weighted peak intensities [11] are also used extensively.

There are numerous variations of these methods. As an example, the approach currently taken by the National Institute of Standards and Technology (NIST) makes use of composite similarity measure using a modified dot product and a correction function which considers x-axis shifts (see Section 3.2.1 for more detail).

However, in spite of all the past work done to improve the performance of library search techniques which identify substances by their mass spectra, an optimal method has yet to be found.

2 Solution Approaches

For the purpose of algorithm development, we treat the spectral data as a single sparse vector representing the intensities and infer the mass-to-charge ratios by the coordinate indices. In what follows, we represent the unknown samples by $\mathbf{x_u}$ and the set of library spectra by $\mathbf{L} = \{\mathbf{l}_1, \mathbf{l}_2, ..., \mathbf{l}_n\}$, where n is the number of available library elements.

2.1 Bayesian Based Search Algorithm

One approach used to solve the issue of identifying unknown sugars through spectral library searching is described by the following probabilistic framework considering conditional probability and Bayesian-learning methods, similar to methods seen in many pattern recognition studies [1, 4].

The problem can be described as in (1), where the probability of a library entry, $\mathbf{l_i}$, being the match for the unknown vector, $\mathbf{x_u}$, is proportional to the product of the probability of likelihood, and the effects of prior knowledge.

$$\underbrace{P\left(\mathbf{l_i}|\mathbf{x_u}\right)}_{\text{Prob. of Occurrence}} \quad \propto \quad \overbrace{P\left(\mathbf{x_u}|\mathbf{l_i}\right)}^{\text{Likelihood}} \underbrace{P\left(\mathbf{l_i}\right)}_{\text{Prior Prob.}} \tag{1}$$

If a Gaussian distribution is assumed for the conditional likelihood density, its probability can be approximated by a monotonically decreasing exponential function as in equation (2), where D_1 is a distance metric considering both the library and unknown vector, and σ_{v_1} is an associated variance.

$$P\left(\mathbf{x_u}|\mathbf{l_i}\right) \sim \mathbf{exp}\left(-\frac{1}{2}\frac{D_1\left(\mathbf{l_i}, \mathbf{x_u}\right)}{\sigma_{v_1}^2}\right) \tag{2}$$

Similarly, the probability density associated with the prior knowledge may be assumed to be a normally-distributed monotonically decreasing function, and so can be approximated by equation (3), with the new distance measure, D_2, accepting descriptive features from the library entry, $\mathbf{q_L^i} \Subset \mathbf{l_i}$, and the unknown vector, $\mathbf{q_u} \Subset \mathbf{x_u}$, as representative of prior information.

$$P\left(\mathbf{l_i}\right) \sim \mathbf{exp}\left(\frac{-D_2\left(\mathbf{q_L^i}, \mathbf{q_u}\right)}{\sigma_{v_2}^2}\right) \tag{3}$$

By including the Gaussian approximations (2) and (3) in equation (1) and replacing the variance measures by parameters, a final representation can be developed as described by

$$P\left(\mathbf{l_i}|\mathbf{x_u}\right) \propto \mathbf{exp}\left(\alpha D_1\left(\mathbf{l_i}, \mathbf{x_u}\right) + \beta D_2\left(\mathbf{q_L^i}, \mathbf{q_u}\right)\right), \tag{4}$$

where $\alpha = \left(\frac{-1}{2\sigma_{v_1}^2}\right)$ and $\beta = \left(\frac{-1}{2\sigma_{v_2}^2}\right)$. Distance measures can be varied (see Section 3) and parameters α and β must be tuned to best represent the given distributions. The library vector with the greatest probability of occurrence is the best match as predicted by this method, hence it will maximize the exponential argument in (4). In other words,

$$P\left(\mathbf{l_i}|\mathbf{x_u}\right) \sim \alpha D_1\left(\mathbf{l_i}, \mathbf{x_u}\right) + \beta D_2\left(\mathbf{q_L^i}, \mathbf{q_u}\right).$$

2.2 Hierarchical Search Heuristic

As an alternative, a more heuristic approach was also developed to determine the identity of an unknown sugar using the spectral database. Similarity (δ_i) between sample and library vectors was compared using a hierarchical approach using key vector features to pare down the number of library entries prior to ranking.

δ_1 - Maximum Peak Similarity:

The first feature considered was the index of the maximum peak. In spectral data, major peaks can be strong indicators of a compounds identity [6]. However, because of potential low resolution readings and machine noise associated with the spectra, assuming that maximum peaks occur in identical locations for

similar library and unknown vectors would be inappropriate. As such, a tolerance, ε_1, is used within (5) to allow only those library entries satisfying this restriction to be considered in subsequent tests. This tolerance can vary from 0, indicating a high importance in major peak location, to the entire length of the vector, for use with high noise spectrum.

$$\delta_1 \;=\; ||\text{index}\left(\max\left(\mathbf{l_i}\right)\right) - \text{index}\left(\max\left(\mathbf{x_U}\right)\right)|| \;\leq\; \varepsilon_1 \tag{5}$$

δ_2 - Auxiliary Peak Similarity:

Auxiliary peak similarity uses additional peaks sorted by intensity to a specified level to construct the next set of features considered in the hierarchy. Auxiliary peaks have weaker constraints than its maximum peak counterparts, as an error is permitted in both the mass-to-charge ratio (index of the vector) and the intensity value (vector element value). This elimination process can be described as follows:

$$\delta_2 \;=\; ||\mathrm{S}_1\left(\mathbf{l_i}, \omega\right) - \mathrm{S}_1\left(\left(\mathbf{x_U}, \omega\right)\right|| \;\leq\; \varepsilon_2$$

where the function S_1 sorts the vector in descending order and purges entry 1 (maximum peak) and entries with index greater than ω. The sorted library vectors with difference less than the second tolerance, ε_2, are passed through to the next component of the hierarchical search.

δ_3 - Distance:

The final step in the hierarchy will be ranking of filtered results based on distance measures, similar to the distance metrics (see Section 3) used in the Bayesian based approach, as well as alternative measures of similarity which do not satisfy the requirements of metrics (see Section 3.2). To provide increased confidence in final readings, these measures are used in parallel.

3 Distance Measures

3.1 Metrics

Distance metrics are used extensively in current literature to measure similarity between the spectral library and the unknown spectrum. For the sugar identification problem three main distance measures between the vectors of intensities are evaluated:

1. ℓ^1 (Manhattan) distance. This distance is computed by summing absolute differences between pairs of vector components. The formula is given by

$$d_{\text{Manhattan}}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} |p_i - q_i|.$$

2. ℓ^2 (Euclidean) distance. In an n-dimensional vector space, the ℓ^2 distance is defined as

$$d_{\text{Euclidean}}(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}.$$

3. Canberra distance. Canberra distance metric is a weighted variation of the ℓ^1 distance, given by

$$d_{\text{Canberra}}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|}.$$

3.2 Alternative Measures

In this section we consider other measuring techniques which are not necessarily metrics but still prove to be effective in comparing spectra.

3.2.1 NIST Measure

The method by which the official NIST software computes the similarity of two chemical spectra is described in paper [10] by Stein. We denote by $\boldsymbol{\ell}$ and $\boldsymbol{u}$ two spectra, $m_{\boldsymbol{\ell}}$ and $m_{\boldsymbol{u}}$ the respective maximum mass-to-charge ratios, and by $a_{\boldsymbol{\ell}}(m)$, $a_{\boldsymbol{u}}(m)$ the intensities observed at mass-to-charge values $m = 1, 2, \ldots$. Furthermore, if we define $N_{\boldsymbol{u}}$ to be the number of nonzero intensities in vector $\boldsymbol{u}$ and N_C the number of pairs of nonzero intensities common to both $\boldsymbol{\ell}$ and $\boldsymbol{u}$, we have the following composite similarity measure approximating that used by NIST:

$$M = \frac{1000}{N_{\boldsymbol{u}} + N_C} \left(N_{\boldsymbol{u}} F_1 + N_C F_2\right)$$

where

$$F_1 = \frac{\sum_{m=1}^{\max(m_{\boldsymbol{\ell}}, m_{\boldsymbol{u}})} m\sqrt{a_{\boldsymbol{\ell}}(m) a_{\boldsymbol{u}}(m)}}{\sqrt{\sum_{m=1}^{m_{\boldsymbol{\ell}}} m a_{\boldsymbol{\ell}}(m) \sum_{m=1}^{m_{\boldsymbol{u}}} m a_{\boldsymbol{u}}(m)}}$$

and

$$F_2 = \frac{1}{N_C} \sum_{c=2}^{N_C} \left[\frac{a_{\boldsymbol{\ell}}(m_c)}{a_{\boldsymbol{\ell}}(m_{c-1})} \frac{a_{\boldsymbol{u}}(m_{c-1})}{a_{\boldsymbol{u}}(m_c)}\right]^n,$$

with $n = 1$ or -1 if the enclosed term is less than or greater than one, respectively. Note that the intensities at the m_c values are taken to be nonzero for all c, i.e. a shift in the mass-to-charge ration is considered with only neighbouring pairs of nonzero values of the intensities. To conclude, F_1 is essentially a modified cosine similarity measure and F_2 takes into account shifts by one index in the x-variable.

3.2.2 Modified NIST Measure

The limitations of the F_2 function defined above motivate the study of a third function which considers shifts by two indices in the x-direction. Define $\mathbf{N}_D$ to be the set of pairs of mass-to-charge ratios which have nonzero intensities common to both $\boldsymbol{\ell}$ and $\boldsymbol{u}$ such that

$$\mathbf{N}_D = \{m_d \mid [a_{\boldsymbol{u}}(m_d) \neq 0, a_{\boldsymbol{\ell}}(m_d) \neq 0, a_{\boldsymbol{u}}(m_{d+2}) \neq 0, a_{\boldsymbol{\ell}}(m_{d+2}) \neq 0] \text{ and } [a_{\boldsymbol{u}}(m_{d+1}) = 0 \text{ or } a_{\boldsymbol{\ell}}(m_{d+1}) = 0]\}.$$

We let N_D be the number of elements in $\mathbf{N}_D$. We now define

$$F_3 = \frac{1}{N_D} \sum_{d=3}^{N_D} \left[\frac{a_{\boldsymbol{\ell}}(m_d)}{a_{\boldsymbol{\ell}}(m_{d-2})} \frac{a_{\boldsymbol{u}}(m_{d-2})}{a_{\boldsymbol{u}}(m_d)}\right]^n,$$

with $n = 1$ or -1 if the enclosed term is less than or greater than one, respectively. The modified NIST (mNIST) measure is then taken to be

$$M = \frac{1000}{N_{\boldsymbol{u}} + N_C + N_D} \left(N_{\boldsymbol{u}} F_1 + N_C F_2 + N_D F_3\right).$$

3.2.3 Earth Mover's Approximation

Another idea proposed was a dissimilarity measure inspired by the idea of the Earth Mover's Distance. There is no closed expression for this measure, and as such we describe it as an algorithm. We consider vectors $\boldsymbol{\ell}$ and $\boldsymbol{u}$ as rows of bins, which have $a_{\boldsymbol{\ell}}(m)$ and $a_{\boldsymbol{u}}(m)$ amounts of mass in bin m, respectively. We wish to determine how much mass must be moved to bring vector $\boldsymbol{u}$ to be equal to vector $\boldsymbol{\ell}$, with the restriction that that mass can only be moved left or right by one bin, and allowing mass to be added or removed to attain equality. We calculate this amount by calculating the total amount of mass moved, added, and removed by following the steps outlined in the following algorithm.

```
for i=1:length(L)-1
    if L(i) == U(i)
        % bins contain equal amounts of mass, proceed to left
        continue;
    end
    if     L(i) > U(i) && L(i+1) < U(i+1)
        % move mass from U(i+1) to U(i)
    elseif L(i) > U(i) && L(i+1) > U(i+1)
        % add mass to U(i)
    elseif L(i) < U(i) && L(i+1) > U(i+1)
        % move mass from U(i) to U(i+1), removing any excess which may remain in U(i)
    elseif L(i) < U(i) && L(i+1) < U(i+1)
        % remove mass from U(i)
    end
end
```

In general, the less mass required to be moved to transform $\boldsymbol{u}$ into $\boldsymbol{\ell}$, the more similar the spectra are.

3.2.4 Tanimoto Index

The Tanimoto index is a method of describing the similarity between two chemical samples by equating them with binary vectors. A method to measure the similarity between chemical structures using the Tanimoto index previously appeared in [3, p. 78], where the authors considered a set of chemical substructures, each with a corresponding vector entry which would be set to 1 if said substructure was present, and 0 if absent. The Tanimoto index between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is thus calculated by

$$T(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_i x_i \wedge y_i}{\sum_i x_i \vee y_i}$$

where $\wedge$ is the logical AND operator and $\vee$ is the logical OR. As such, $T(\boldsymbol{x}, \boldsymbol{y})$ assumes values between 0 and 1, with $T(\boldsymbol{x}, \boldsymbol{y}) = 1$ if $\boldsymbol{x} = \boldsymbol{y}$.

Instead of comparing chemical substructures, we instead consider whether a given chemical spectra has a nonzero intensity at a given m/z value, assigning a value of 1 if so, and 0 otherwise. As such, the Tanimoto index in this case will measure the similarity of two chemical spectra without considering the magnitudes of the intensities observed.

4 Methodology

Our experimental library consisted of 41 mass spectra: 21 disaccharides and 20 monosaccharides. We tested our methods with two different experimental set-ups coded in MATLAB.

The first test involved introducing error in one of the library spectra. This was done so that it simulates actual experimental error as best as possible. We specify the percentage of change in the intensities with an associated probability of occurance, increasing and decreasing being equally likely. We also allow for changes in the mass-to-charge ratio by a specified amount by specifying the probabilities of shifting by 1, 2, or 3 positions, with both directions equally likely.

The second test was on a set of 24 unknown samples, provided by NIST, which we try to match to the elements in the library and compare results to those computed by the NIST algorithm. The next section summarizes our results for the second test using various similarity measures.

5 Numerical Results

The tests were performed by applying the hierarchical and Bayesian based methods as well as the various distance measures discussed above. We do not include the results from the first test where we induced the noise because the correct match was always found. A comprehensive summary of results is given in Table 1 which indicates the number of times the correct compound was found in the top three hits.

	Library size	**Number of test spectra**	**Distance measures**			**EM**	**Hier**	**Tan**	**NIST**	**mNIST**	**Bayes**
			Manh	**Eucl**	**Canb**						
Di	21	11	5	3	4	6	6	4	11	10	10
Mono	20	13	11	10	9	10	11	7	13	13	13
Total	**41**	**24**	**16**	**13**	**13**	**16**	**17**	**11**	**24**	**23**	**23**

Table 1: Comparative performance of different methods. Definitions: “Di” - disaccharides, “Mono” - monosaccharides, “Manh” - Manhattan, “Eucl” - Euclidean, “Canb” - Canberra, “EM” - Earth Mover, “Hier” - Hierarchical approach, “Tan” - Tanimoto index, “mNIST” - Modified NIST, “Bayes” - Bayesian-based method

It is observed that the Manhattan, or ℓ^1 distance, performs the best among the three considered distance metrics. Better results are obtained for monosaccharides since correct identification of disaccharides is complicated by extensive noise in the test spectra.

EM measure performance is comparable with the ℓ^1 distance metric results. Both of them are slightly outperformed by hierarchical method. The poorest results are obtained by Tanimoto index measure - less than a half of correct hits. Of the new methods, the Bayesian and the mNIST measures give the best results.

If we instead consider the ideal case where success is defined by getting the correct match as the top hit, the results of this case are tabulated in Table 2.

The mNIST measure performs the best in this scenario, followed by the Bayesian method. The original NIST formula has the lowest success rate. Note that the Hierarchical method will have the correct top hit 82% of the time if the correct compound made it through to the final stage of the algorithm.

	NIST	mNIST	Bayesian	Hierarchy
Disaccharides	6	9	6	5
Monosaccharides	7	8	9	9
Total	**13**	**17**	**15**	**14**

Table 2: Number of correct top hits

6 Observations and Conclusions

We found that the NIST search algorithm was the most consistent in always having the correct match as part of the top three hits, which is the comparison considered in Table 1. However, by looking at Table 2 we see that it gave the poorest performance when considering only the first hit. The mNIST, Bayesian, and Hierarchical approaches all outperformed the NIST algorithm in the ideal measurement. We note that the methods which produced the best results were obtained using modifications of the NIST algorithm: the mNIST method considers an addition to the NIST algorithm and the Bayesian based approach uses the NIST algorithm as its likelihood measure. All of these points imply that there is room for improvement with the current search method employed by NIST.

We believe that the hierarchy method can be improved substantially just by overcoming a single obstacle: the problem is that since with each hierarchical step, the algorithm discards a set of spectra from the library and then does a final similarity measure on the remaining elements. Therefore, if the real compound has been discarded there is no possibility of making an accurate match. This may be improved by considering, for example, a binning method where we still consider the discarded spectra but give a preference to the spectra measured in the final steps. We can further improve this method by considering different measures in the last step of the algorithm. One possibility would be to apply the Bayesian based method to compare the remaining library elements to the unknown sample.

The Bayesian based method itself proved to be promising as well, but also has many possibilities for improvement. The measures used as the prior and likelihood functions can be optimized, as well as the parameters α and β.

The Modified NIST method worked the best when compared to the other approaches considered. It would be fruitful to consider theoretically a sum of sums of the form F_2 and F_3 to see if it can be shown that this has a converging limit and relate this limit to the actual solution.

All of the methods considered yielded good results for particular subsets of noise patterns. If a correct union of these methods could be found much improvement could be made to the current NIST search algorithm.

Acknowledgements

We would like to thank the IMA for organizing the Math Modeling in Industry Workshop which inspired this work. We are grateful to PIMS and NSF for the financial support which allowed us to attend and get a glimpse of what a career in industry is comprised of. Finally, the team would like to thank Anthony Kearsley, our mentor, for providing us with an interesting problem and for guiding us through the research process.

References

[1] E.R. Arce-Santana, J.M. Luna-Rivera, D.U. Campos-Delgado, and O. Gutierrez-Navarro. An object-tracking algorithm based on bayesian-learning. In *Industrial Electronics, 2007. ISIE 2007. IEEE International Symposium on*, pages 1681 –1686, june 2007.

[2] B.L. Atwater, D.B. Stauffer, F.W. McLafferty, and D.W. Peterson. Reliability ranking and scaling improvements to the probability based matching system for unknown mass spectra. *Analytical Chemistry*, 57(4):899–903, 1985.

[3] W. Demuth, M. Karlovits, and K. Varmuza. Spectral similarity versus structural similarity: mass spectrometry. *Analytica Chimica Acta*, 516(1-2):75–85, 2004.

[4] R.O. Duda, P.E. Hart, and D.G. Stork. Pattern classification and scene analysis 2nd ed. 1995.

[5] A. Henderson, J.D. Moore, and J.C. Vickerman. Sims informatics. *Surface and Interface Analysis*, 2012.

[6] F.W. McLafferty, R.H. Hertel, and R.D. Villwock. Probability based matching of mass spectra. rapid identification of specific compounds in mixtures. *Organic Mass Spectrometry*, 9(7):690–702, 1974.

[7] G. Ritter, H. Woodruff, S. Lowry, and T. Isenhour. An algorithm for a selective nearest neighbor decision rule (corresp.). *Information Theory, IEEE Transactions on*, 21(6):665–669, 1975.

[8] E.S.C. Shih and M.J. Hwang. Protein structure comparison by probability-based matching of secondary structure elements. *Bioinformatics*, 19(6):735–741, 2003.

[9] S. Sokolow, J. Karnofsky, and P. Gustafson. The finnigan library search program. *Finnigan Application Report*, 2:1–45, 1978.

[10] S.E. Stein. Chemical substructure identification by mass spectral library searching. *Journal of the American Society for Mass Spectrometry*, 6(8):644–655, August 1995.

[11] S.E. Stein and D.R. Scott. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, 5(9):859–866, 1994.

[12] D.L. Tabb, M.J. MacCoss, C.C. Wu, S.D. Anderson, and J.R. Yates III. Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Analytical chemistry*, 75(10):2470–2477, 2003.