Statistics and Its Interface Volume 5 (2012) 39–46

Spectral library searching for peptide identification in proteomics

Henry Lam

Spectral library searching is an emerging approach in peptide identification from tandem mass (MS/MS) spectra, a critical step in proteomic data analysis. Tandem mass spectrometry is the process by which peptides are fragmented by high energy in a mass spectrometer. The tandem mass spectra thus collected record the mass-to-charge ratios and abundance of the resulting fragments, and can be used to deduce the peptide sequence. Conceptually, spectral library searching is based on the premise that the fragmentation pattern of a peptide can be viewed as a reproducible fingerprint of that peptide, such that unknown spectra acquired under the same conditions can be identified by spectral matching. In practice, a spectral library is first meticulously compiled from a large collection of previously observed and identified MS/MS spectra, usually obtained from real proteomics experiments of complex mixtures. Then, a query spectrum is identified by spectral matching using recently-developed spectral search engines. A key component of this method is a similarity scoring function that numerically defines the similarity between two spectra. In addition to the similarity score, various methods exist to evaluate the statistical significance of the match, and hence the identification accuracy. This review aims to introduce statisticians, especially those unfamiliar with proteomics data analysis to this rapidly evolving field, and to provide a high-level description of the underlying algorithms and the outstanding challenges.

Keywords and phrases: Spectral libraries, Spectral searching, Proteomics, Mass spectrometry.

1. INTRODUCTION

Proteomics is the systematic study of the proteome, which is defined as the set of all proteins and their many isoforms in a biological system. A primary goal of proteomics, therefore, is the confident, high-throughput, and system-wide identification and quantification of all proteins in a biological sample. For the past 20 years, thanks to converging advances in genome sequencing, mass spectrometry, high-speed parallel computing, as well as in data analysis methodology, a variety of mass spectrometry-based proteomics technologies have become increasingly powerful and accessible to life science researchers. It is now possible to identify and quantify thousands of proteins, down to femtomolar concentrations, in moderate-scale proteomic experiments using the liquid chromatography–mass spectrometry (LC–MS) platform. This technology has empowered biologists to ask questions that cannot be addressed by traditional molecular biology techniques.

Among the many proposed experimental workflows, the most popular and well-developed method has been the "bottom–up" approach of shotgun proteomics. In shotgun proteomics, proteins are first enzymatically digested into shorter peptides, typically 5–30 amino acid residues in length, which are more amenable to LC–MS analysis. The peptide mixture is then optionally fractionated, before injection into a reverse-phase liquid chromatography column coupled to the mass spectrometer through an ion source. The ion source either applies a strong electric field to ionize the peptides in the case of electrospray ionization (ESI), or employs a laser to energize the peptides to the point of ionization, in the case of matrix-assisted laser desorption ionization (MALDI). In the mass spectrometer, peptide ions are first separated based on their mass-to-charge ratios. Then, selected ions are then isolated and fragmented to yield characteristic fragmentation patterns, a process termed tandem mass spectrometry (MS/MS). The fragmentation can be done by high-energy collision with inert gas molecules (collision induced dissociation, CID) or more recently by a gas-phase reaction triggered by electron transfer (electron transfer dissociation, ETD). The resulting fragments are detected and recorded in MS/MS spectra., which were then used to deduce the peptide sequence, usually by computational methods [1, 5, 22, 25].

To set the stage for our discussion of spectral library searching, it is perhaps useful to explain briefly the process of inferring sequence information from peptide fragmentation patterns. An illustrative example is shown in Figure 1, a typical CID spectrum from the peptide VEDALSATR (charge +2). In CID fragmentation, the amide bonds (between the carbonyl carbon and the amine nitrogen) along the backbone are the primary cleavage sites, producing b ions on the N-terminal side and y ions on the C-terminal side (Figure 1). The frequency of bond breaking events is however not evenly distributed along the length of the peptide, due to variation in bond strengths and other factors. For example, the y7 ion (the fragment DALSATR) is about 20

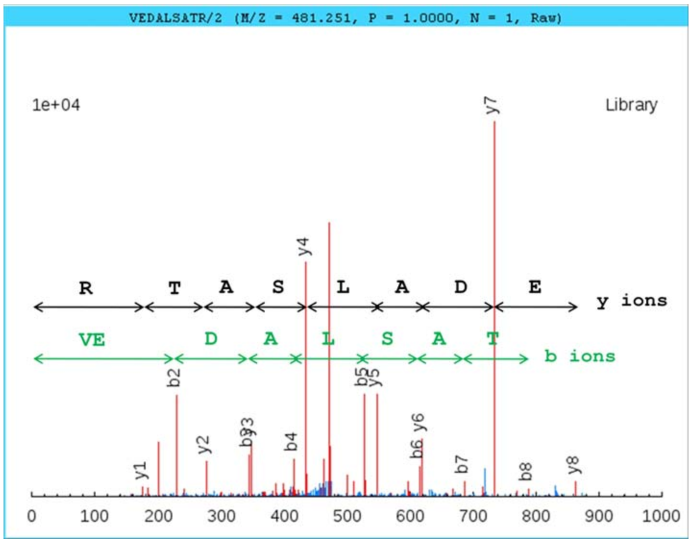

Figure 1. Assigning a tandem mass spectrum to its peptide sequence. The spectrum from the peptide VEDALSATR (charge 2) is shown. Two fragment ladders, the y ions (the fragments that contain the C-terminus, black solid arrows), and the b ions (the fragments that contain the N-terminus, green dashed arrows (color online)) are apparent; the corresponding fragment ion peaks are labeled (y1, b7, etc.). The distance between two peaks in a ladder should correspond to the mass of an amino acid. Note the variety of peak intensities and the presence of non-b,y ions (those without a peak label).

times more abundant than the y8 ion (the fragment EDALSATR). In addition, other fragmentation events, such as loss of neutral molecules like water and ammonia from the amino acid side-chains, as well as many secondary fragmentation events (i.e., fragmentation of fragments) will also occur to some extent, also in a somewhat sequence-specific manner. Therefore, the resulting MS/MS spectrum can contain hundreds of fragment ion peaks, of intensity spanning orders of magnitude, even for a short peptide. Fortunately, in most cases the canonical fragment ions (b and y ions) dominate the spectrum, as in Figure 1, and can be enriched to a large extent by simple intensity-based filters. The locations of these peaks in the spectrum can then be compared against theoretically calculated m/z values of these fragments for candidate peptide sequences, whether manually or computationally, to find the best match.

This last step of computationally assigning MS/MS spectra to their peptide identifications is often the rate-limiting step of the whole proteomics experiment, and has received well-deserved attention in the past decade. In the early days of proteomics, when data volume is much smaller, it was feasible to interpret tandem mass spectra, or at least to verify identification made by the computer, manually. But since then, rapid advances in instrumentation have necessitated the development of computational methods for this purpose. The most commonly used method is called sequence (database) searching [2, 8, 10, 17, 21], so named because a protein sequence database is used to define the "search space," namely, all putative peptide candidates. Nowadays, protein sequence databases of many model organisms are compiled from genome sequences by application of gene prediction methods, and as such roughly define all possible peptides that can be derived from that organism. Given this information, the sequence search engine then attempts to predict the theoretical fragmentation pattern of each candidate peptide, and compare the experimental observed spectra to these theoretical spectra for the best match. Apart from the m/z values of the canonical fragment ions (b and y ions), however, few details of the MS/MS spectrum are readily predictable from the sequence in practice. So sequence search engines typically assume that all canonical fragments are present at equal intensity in the theoretical spectra. As explained above, this fails to capture the observed richness of empirical spectra, in which peak intensities often differ by orders of magnitude, and uncommon peaks due to secondary or side-chain cleavages abound. Therefore peptide-to-spectrum similarity scoring in sequence searching is not optimal, contributing to diminished sensitivity and an inability to identify lower-quality spectra. Moreover, because of the vast search space, sequence searching is also painfully slow and often requires expensive computational infrastructure [20].

Recently, an alternative approach, spectral (library) searching, which promises to address these shortcomings of sequence searching, has drawn increasing interest. Spectral searching, however, is not a novel concept. For decades, analytical chemists have gathered reference mass spectra of small molecules, compiled them into a searchable library, and used spectral matching as a means to identify mass spectra from unknown analytes [4, 19]. In 1998, it was first suggested the same approach can be used in proteomics to identify peptide MS/MS spectra [27]. Faced with the difficulty of obtaining enough data to compile spectral libraries, the idea failed to catch on until recently, when several technological advances converge to produce an explosion of proteomic data. The development of LC/MS platforms that can handle complex protein samples, the advent of the modern mass spectrometer of much improved throughput, the maturation of proteomic data analysis methods, the standardization of data formats and the emergence of public data repositories all contribute to making it feasible to compile spectral libraries of peptides. In fact, the National Institute of Standards and Technology of the United States began in 2006 to extend their mass spectral library, previously consisting of small molecules, to include peptides. Parallel to this development is the advent of spectral search engines designed to utilize spectral libraries for peptide identification [3, 9, 15]. Table 1 lists some useful websites for obtaining spectral libraries and spectral search engines developed in the past few years.

Table 1. Useful websites for spectral library building and searching tools, adapted from Ref. [13]

NISTMS Search	Software and library download, instructions • http://peptide.nist.gov/
X!Hunter	Software download • ftp://ftp.thegpm.org/projects/xhunter/binaries Library download • ftp://ftp.thegpm.org/projects/xhunter/libs Web client to X!Hunter on remote server • http://xhunter.thegpm.org/
Bibliospec	Software download • http://depts.washington.edu/uwc4c/express-licenses/assets/bibliospec/ Library download and instructions • http://proteome.gs.washington.edu/software/bibliospec/documentation/
SpectraST	Software download • http://sourceforge.net/projects/sashimi/files/ (SpectraST is part of Trans Proteomic Pipeline) Library download • http://www.peptideatlas/speclib/ • http://peptide.nist.gov/ Instructions • http://tools.proteomecenter.org/wiki/index.php?title=SpectraST Web client to SpectraST on remote server • http://www.peptideatlas.org/spectrast/

Spectral library searching differs from sequence database searching in several ways. The most important of these are: (1) the use of experimental, as opposed to theoretical, spectra to match query spectra, and (2) a much reduced search space focused on known segments of the proteome. Both factors are cited as reasons for the improved performance of spectral searching. Experimental evidence of this improvement was amply provided in the literature and will not be repeated here [3, 9, 15, 28].

Spectral searching compares experimental spectra to experimental spectra, whereas sequence searching compares experimental spectra to theoretical spectra. As discussed above, the theoretical spectra considered in sequence searching are simplistic and do not resemble the experimental spectra that they are supposed to match. In contrast, armed with previously observed experimental spectra, spectral searching can take full advantage of all spectral features, including actual peak intensities, neutral losses from fragments, and various uncommon or even uncharacterized fragments, to determine the best match (Figure 2). As a result, the similarity scoring of spectral searching is more precise, and in principle should provide better discrimination between good and bad matches. A recent publication attempted to isolate this effect, and showed that the use of real reference spectrum for matching seemed to play a major role in the improved sensitivity of spectral searching, and that the peak intensities and non-canonical ions, information that is ignored by sequence searching, are both important contributors. It further demonstrated that spectral searching outperforms sequence searching to a greater extent when the query spectra are of lower quality, lending credence to the belief that spectral searching should be more effective in identifying biologically interesting low-abundance peptides, whose acquired spectra should have lower signal-to-noise ratios [28].

Spectral searching also benefits from a much reduced search space; it has fewer candidates to consider. By definition, spectral libraries only consist of previously observed and identified peptide ions of a proteome, whereas a sequence search engine considers all putative peptide sequences derivable from the corresponding sequence database. It is well known that most of these putative peptides are never observed in practice. Therefore, with typical search parameters, the search space of spectral searching can be several orders of magnitude smaller than sequence searching. This leads to a considerable saving in the running time required per query. In addition, one also expects the reduction of search space should also ameliorate the so-called "distraction effect," leading to improved sensitivity, although the extent of this improvement remains controversial and poorly understood.

In terms of limitations, given the aforementioned narrowing of search space, it is obvious that spectral searching can only be applied to situations where discovery of novel peptides or proteins is not the goal. Fortunately, more and more opportunities of scientific discovery lie in understanding how the known segments of the proteome change with time and circumstances, and how they interact to produce the biological function. Accordingly, in proteomics, there is a shifting emphasis from discovery-oriented endeavors to targeted and quantitative proteomics in which one is merely interested in studying known and previously observed peptides [6, 12]. Spectral searching is well suited to this type of workflows, especially when working in tandem with traditional

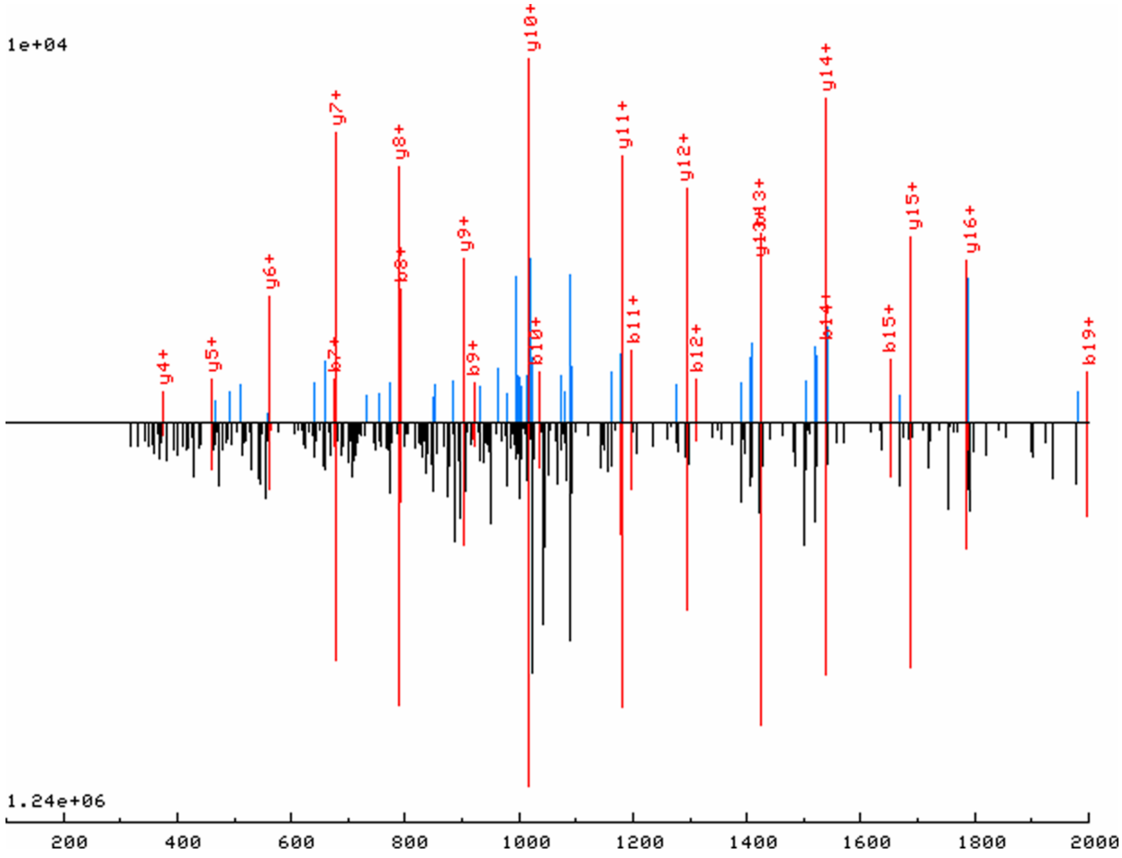

Figure 2. An example of peptide identification by spectral searching. The top spectrum is the consensus library spectrum of the peptide ion AVGSLTFDENYNLLDTSGVAK (+2). The bottom spectrum (upside down) is a query spectrum identified confidently by SpectraST. Note how spectral searching makes use of the reproducibility of peak intensities and non-b,y ions (those without peak labels) for a more global and precise similarity scoring, allowing it to tolerate occasional unmatched features. Adapted from Ref. [15].

discovery-based methods. For example, sequence searching can be employed first on a reference sample to construct a spectral library, followed by quick and sensitive spectral searching to find the same peptides in many parallel experiments. This would be ideal for large-scale quantitative proteomic experiments with many samples and replicates, such as a time series experiment, or a clinical study involving many subjects.

2. SPECTRAL LIBRARY SEARCHING ALGORITHMS

Several spectral search engines designed for proteomics applications have been developed in the past 5 years. Here we focus on the traditional, more well-established tools that perform straightforward spectral matching; newer methods that use libraries for peptide identification in some other ways are outside the scope of this review. For a more in-depth discussion on the usability and surrounding informatics support of these tools, the reader is referred to Ref. [13].

Conceptually, the premise of spectral library searching is very simple: that the fragmentation pattern of a molecule under some fixed conditions is a reproducible *fingerprint* of that molecule, such that unknown spectra acquired under the same conditions can be identified by spectral matching. In actual practice, spectra will inevitably contain experimental artifacts (e.g., random noise and signals from contaminants), or the fragmentation conditions might not be exactly the same. But very much like fingerprinting in forensic science, imperfect matches do not necessarily preclude correct identification, because the fingerprint typically contains far more information than is necessary to distinguish a significant match from a spurious one. The challenge therefore lies in developing spectral matching algorithms that uses that information properly to minimize false matches, but retains the robustness and flexibility to accommodate imperfect, but true matches.

A summarized work-flow of spectral library searching is illustrated in Figure 3. For any query spectrum, the search engine can first make use of the m/z value of the intact peptide ion (called the precursor), measured by the mass

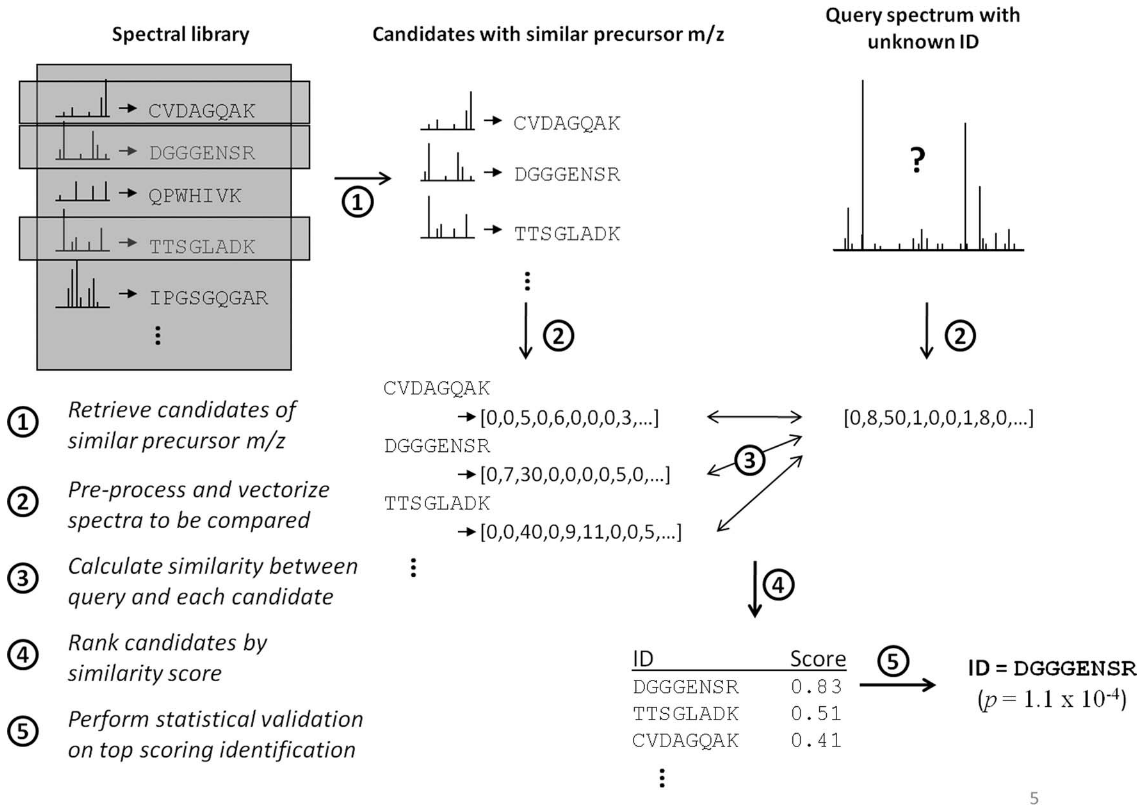
Spectral library
Candidates with similar precursor m/z
Query spectrum with unknown ID
CVDAGQAK
DGGGENSR
QPWHIVK
TTSGLADK
IPGSGQGAR
①
CVDAGQAK
DGGGENSR
TTSGLADK
?
②
②
CVDAGQAK
→[0,0,5,0,6,0,0,0,3,...]
DGGGENSR
→[0,7,30,0,0,0,0,5,0,...]
TTSGLADK
→[0,0,40,0,9,11,0,0,5,...]
[0,8,50,1,0,0,1,8,0,...]
③
④
① Retrieve candidates of similar precursor m/z
② Pre-process and vectorize spectra to be compared
③ Calculate similarity between query and each candidate
④ Rank candidates by similarity score
⑤ Perform statistical validation on top scoring identification
ID Score
DGGGENSR 0.83
TTSGLADK 0.51
CVDAGQAK 0.41
⑤
ID = DGGGENSR
($p = 1.1 \times 10^{-4}$)
5

Figure 3. Work-flow of spectral searching, from the retrieval of candidates from the spectral library to statistical validation of putative identifications.

spectrometer before tandem mass spectrometry, to select candidates from the spectral library. Then each candidate spectrum is compared to the query spectrum after some preprocessing, to yield a similarity score. The top scoring (most similar) candidate is taken to be the putative identification of the query spectrum. The same process is repeated for each query MS/MS spectrum acquired in the experiment.

With respect to the most important step of similarity scoring, the aforementioned engines generally share the same approach, but differ slightly in the details. The first step in determining the similarity of two spectra is to employ various heuristics to de-noise the spectra. While more sophisticated methods based on signal processing techniques are available, in practice they are too complex to implement and usually not worth the computational cost. Instead, arbitrary thresholding based on the absolute intensity or relative intensity, or limiting a spectrum to only a fixed number of most intense peaks (in the entire spectrum or within sliding m/z windows), are typical methods. The logic behind this approach is simple: one expects that the majority of information of a reference spectrum is captured in the handful of most intense peaks, and that there should be a limited number of prominent fragment ions for any given peptide under typical fragmentation conditions. However, it has also been shown that oversimplification of spectra will hurt the discrimination power [16], and that minor ions such as fragment neutral losses indeed carry information that helps boost the sensitivity of spectral searching [28].

The second step of similarity scoring involves the coarse-graining and vectorization of the spectra. Namely, the entire m/z range is subdivided into a predefined number of “bins,” and the peak lists are converted to a high-dimensional vector, with each of the elements being the summed intensity within one bin. The bin width can be chosen to reflect the mass resolution of the instrument; with typical ion trap instruments a bin width of 1 Da/e is customary. This process of “binning” converts peak lists of different lengths into equal-size vectors, so that they can be easily compared. In addition, various experimental artifacts such as imperfect mass accuracy and peak splitting are partially dealt with by this simple coarse-graining, at the expense of some loss in discrimination.

The third and perhaps most important step is similarity scoring. The scoring function then takes these vectors of spectra as input, and computes a quantity that reflects how similar the two vectors are. There are numerous ways to define the similarity of two vectors, but historically, for the purpose of mass spectrum comparison, two types of measures are commonly used [24]. On one extreme is the shared peak count, i.e. the number of peaks that are found in both of the spectra to be compared, divided by some normalization factor, and sometimes further adjusted to properly account for a calculated probability of a random peak match. This function, however, does not take into account of the peak intensity at all, which somewhat defeats the purpose of spectral library searching. In fact, this scoring function is commonly used in sequence search engines, for which the theoretical spectra do not have well-predicted peak intensities. On the other extreme is the dot product, which simply measures the cosine of the angle subtended by the two vectors in high-dimensional space. (A dot product of one indicates overlapping vectors, whereas a dot product of zero indicates orthogonal vectors.) A matching peak that is twice as intense will contribute 4 times as much to the dot product, as the intensities from either spectrum are multiplied. As such, the dot product weighs the peak intensity heavily, and can be prone to error when there are dominating peaks or when the peak intensities have low reproducibility for some reason. Existing spectral search engines therefore adopt a sensible approach that strikes a balance between these two extremes, to accommodate different spectral shapes and to anticipate some noisy fluctuations in the peak intensities. There was also a proposal to quantify the variability of individual peaks and use such information to weigh matched intensity, in a hidden-Markov model-based matching algorithm [26].

It is worth noting that the spectral matching algorithms discussed above are still evolving and may not be optimal. In fact, the best solution may be different for different types of query data and for libraries constructed in different manners. Despite the empirical success of these algorithms, there remains a lack of a theoretical framework for systematically studying aspects of these algorithms, due partly in our poor understanding of the fragmentation patterns of peptides and the nature of noise in tandem mass spectra. This will likely remain fertile ground of research in the near future.

3. STATISTICAL VALIDATION OF SPECTRAL SEARCH RESULTS

As with the traditional method of sequence database searching, the spectral search engine considers all candidates within a certain precursor m/z window for each query spectrum, and returns the top scoring (most similar) match among the candidates, along with some numerical measure of how good the match is. A particular problem in proteomics is that many query spectra are not identifiable in the first place, due to a myriad of reasons. Some spectra simply originate from pure noise or non-peptides. More frequently, the correct answer is not in the sequence database or spectral library searched for, and thus it is impossible for any search engine to reach the correct answer. Therefore, the top scoring match for any query spectrum is not necessarily the correct answer, even assuming that search engine is perfect. Put differently, one must entertain the possibility that none of the candidates is actually the correct answer. In proteomics, it is therefore essential that a follow-up step of statistical validation be undertaken to decide whether to accept each of the identifications returned by the search engine. To guide this decision, the false discovery rate among the accepted identifications needs to be estimated. While manual validation, which involves actually examining the spectra in light of the purported identification to see if the peaks are well explained, is possible and sometimes necessary for a small number of identifications deemed critical for the biological questions asked, it is impractical on a large scale. Therefore automatic statistical validation is usually practiced, using a variety of approaches and software tools [11, 18].

While this topic of statistical validation will be covered in details elsewhere in this special issue, there are several points pertaining to statistical validation that are unique in the case of spectral searching. The first issue is a lack of a reliable model for the null score distribution, i.e., that of random (and hence incorrect) matches In both sequence and spectral searching, the search engine will typically provide not only the measure of similarity, such as the dot product, and but also additional scores that help establish the statistical significance of the match. For instance, some search engines report the difference in similarity score between the top match and the runner-up, while some report a p-value-like score that quantifies the probability that the top match is a random event, calculated by assuming a certain parametric score distribution of the incorrect matches. However, in the case of spectral searching, because of the complexity of real spectra, it is difficult to formulate the peak matching process in combinatoric terms. (This is possible in the case of sequence searching, which largely ignores the intensity dimension and assumes simplistic fragmentation patterns.) Therefore an empirical approach is perhaps more suitable for spectral searching at this point, although with accumulating data and knowledge about peptide fragmentation, a theoretical model may yet be feasible. However, even if one adopts an empirical approach, due to its reduced search space, each query spectrum may only be matched against dozens of candidates, and occasionally insufficient sampling of the background score distribution becomes an issue [13]. A usable but imperfect alternative is to adopt a parametric mixture modeling approach, exemplified by the statistical validation tool PeptideProphet. In this method, limited training data is used to determine the shapes of the score distributions of incorrect and correct identifications, in the form of common statistical distributions. The parameters of these distributions are obtained by fitting to the observed score histogram by the expectation-maximization algorithm. In the early days of SpectraST, for example, it was assumed that the incorrect score distribution conforms to a gamma distribution and the correct score distribution is approximated by a normal distribution, for the purpose of PeptideProphet modeling.

This assumption however is entirely empirically driven and based upon the observation of the behavior of a few small training datasets, and since then it was found to cause an overestimation of the error rates when more generally applied [14].

Second, the popular non-parametric approach of decoy counting is not as easily applied to spectral searching than to sequence searching. This empirical approach involves introducing known wrong answers to the search space, and uses the number of matches to these "decoys" to estimate the frequency of errors made by search engines, under the assumption that incorrect matches are equally likely to hit real and decoy candidates. In sequence searching, decoys are generated by reversing or shuffling real protein sequences [7]. This works because sequence searching makes no assumption about the finer details of peptide fragmentation, but rather relies only on the m/z values of canonical fragments, which are accurately determined from the sequence alone. In spectral searching, however, decoys must take the form of spectra that should necessarily generate wrong answers when matched, but are realistic enough to mimic the features of real spectra. For instance, spectra of randomized peaks will not act as effective decoys because they have a much lower chance of matching real spectra due to unrealistic peak-to-peak distances and intensity profiles, violating the assumption of decoy counting. At present, two solutions have been in use: one might use the spectral library of a different organism as decoys, or more generally, one can create artificial decoy spectra for this purpose using the spectral search engine SpectraST. SpectraST attempts to retain peptide-like features in a spectrum by using a real spectrum as a template and re-positioning explainable peaks to match a decoy (e.g. shuffled) sequence. The unexplained peaks in the template are also kept so as to mimic the noise level in real reference spectra. This method was shown to satisfy the aforementioned assumption for decoy counting [14].

Another complication in terms of statistical validation of spectral search results is the potential for error propagation and non-specific matches. In proteomics, the library spectra are generated from real data of complex mixtures, and identified by imperfect computational methods, such as sequence searching. The library spectra cannot be viewed as true gold standards themselves, and will be associated with their own error rates, however low those might be. In other words a small fraction of library spectra may be tagged with incorrect peptide identifications, and any spectral match to these spectra will always yield a wrong identification even when the similarity score is extremely high. Another related problem is that some library spectra, while correctly identified, might contain strong signals from a contaminating species. It is possible that a high-scoring spectral match can be found chiefly due to matching these contaminant peaks, rather than peaks that originate from the peptide with which the library spectrum is identified [16]. For many technical reasons, it is difficult for the library builder to eliminate completely these two types of questionable spectra from the library. Therefore an effective statistical validation approach must deal with this type of errors, which are unique to spectral searching. One approach that has been taken involves associating each library spectrum with some probability of identification accuracy, which will be multiplied by the probability of finding the true spectral match to yield the final identification probability. In other words, one assumes that the correct identification of the library spectrum, and the correct matching of this spectrum to a query spectrum are independent events [15].

Finally, at least in their present nascent state, spectral libraries are still very far from covering all of the proteome expected to be seen by mass spectrometry. Sequence databases, on the other hand, are much closer to complete proteome coverage, at least if one considers only unmodified peptides, thanks in large part to our ability to sequence the whole genome of an organism. Therefore for any given query spectrum, there is a much higher chance that the correct answer is not in the spectral library, than that the correct answer is not in a sequence database, although it must be said that incomplete proteome coverage also applies to sequence searching due to unconsidered post-translational modifications. However it remains unclear how the search space consideration should be factored into the statistical validation of search results. In the case of spectral searching, to obtain more insight into this issue, it might make sense to separate the identification probability explicitly into two terms, one that estimates the likelihood that the correct answer is in the library, and one that estimates the likelihood that the match is correct, given that the correct answer is in the library. This has been attempted for spectral searching of small molecules [23], but is much more difficult technically to apply in proteomics.

4. CONCLUDING REMARKS

The purpose of this review is to introduce the readers to the promise and pressing challenges in the area of spectral library searching in proteomics, in particular in topics that should be of interest to statisticians. It is worth stressing that spectral searching in proteomics is merely 5 years old and many important problems remain unsolved, many of which are statistical in nature. Existing methods are largely developed through trial-and-error, proven to work – often brilliantly – by experiments, but were often not systematically studied or optimized. It is therefore no accident that this review describes more open questions than known answers. To summarize, there is now an urgent need to formulate the processes of spectral matching and spectral library searching in statistical terms, so as to establish a theoretical framework for the rational design of better algorithms. For the problem of spectral matching, much work remains to be done to understand the variability of peptide fragmentation patterns, and to extract information more effectively from the MS/MS spectrum. A model for the noise that is inherent in the process is perhaps needed. With the accumulation of proteomics data in the public sphere, there is no shortage of training data for model development. For the problem of statistical validation of search results, the primary need is in an accurate and robust model for incorrect spectral matches. The hope is that, if nothing else, this review will serve to pique the interest of more accomplished statisticians and to invite them to tackle these interesting problems.

ACKNOWLEDGMENT

H. L. is financially supported by the Research Grant Council of the Hong Kong Special Administrative Region, China (Project No. HKUST 601909).

Received 31 March 2011

REFERENCES

[1] Aebersold, R. and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature* **422**(6928) 198–207.

[2] Craig, R. and Beavis, R. C. (2004). TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics* **20** 1466–1467.

[3] Craig, R., Cortens, J. C., Fenyo, D. and Beavis, R. C. (2006). Using annotated peptide mass spectrum libraries for protein identification. *J. Proteome Res.* **5** 1843–1849.

[4] Domokos, L., Hennberg, D. and Weimann, B. (1984). Computer-aided identification of compounds by comparison of mass spectra. *Anal. Chim. Acta.* **165** 61–74.

[5] Domon, B. and Aebersold, R. (2006). Mass spectrometry and protein analysis. *Science* **312**(5771) 212–217.

[6] Domon, B. and Aebersold, R. (2010). Options and considerations when selecting a quantitative proteomics strategy. *Nat. Biotechnol.* **28**(7) 710–21.

[7] Elias, J. E. and Gygi, S. P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4** 207–214.

[8] Eng, J. K., McCormack, A. L. and Yates, J. R. III (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5** 976–989.

[9] Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S. and MacCoss, M. J. (2006). Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal. Chem.* **78** 5678–5684.

[10] Geer, L. Y., et al. (2004). Open mass spectrometry search algorithm. *J. Proteome Res.* **3** 958–64.

[11] Käll, L., Storey, J. D., MacCoss, M. J. and Noble, W. S. (2008). Posterior error probabilities and false discovery rates: two sides of the same coin. *J. Proteome Res.* **7** 40–44.

[12] Kuster, B., Schirle, M., Mallick, P. and Aebersold, R. (2005). Scoring proteomes with proteotypic peptide probes. *Nat. Rev. Mol. Cell Biol.* **6** 577–583.

[13] Lam, H. and Aebersold, R. (2011). Building and searching tandem mass (MS/MS) spectral libraries for peptide identification in proteomics. *Methods*, PMID:21277371.

[14] Lam, H., Deutsch, E. W. and Aebersold, R. (2010). Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. *J. Proteomics Res.* **9** 605–610.

[15] Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., King, N., Stein, S. E. and Aebersold, R. (2007). Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7** 655–667.

[16] Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., Stein, S. E. and Aebersold, R. (2008). Building consensus spectral libraries for peptide identification in proteomics. *Nat. Methods* **5** 873–875.

[17] MacCoss, M. J. (2005). Computational analysis of shotgun proteomics data. *Current Opinion in Chemical Biology* **9**(1) 88–94.

[18] Nesvizhskii, A. I., Vitek, O. and Aebersold, R. (2007). Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **4** 787–797.

[19] Owens, K. G. (1992). Application of correlation analysis techniques to mass spectral data. *Appl. Spectrosc. Rev.* **27** 1–49.

[20] Patterson, S. D. (2003). Data analysis – The Achilles heel of proteomics. *Nat. Biotechnol.* **21** 221–222.

[21] Perkins, D. N., Pappin, D. J. C., Creasy, D. M. and Cottrell, J. S. (1999). Probability-based protein identification by searching sequence database using mass spectrometry data. *Electrophoresis* **20** 3551–3567.

[22] Steen, H. and Mann, M. (2004). The ABC's (and XYZ's) of peptide sequencing. *Nature Reviews Molecular Cell Biology* **5**(9) 699–711.

[23] Stein, S. (1994). Estimating probabilities of correct identification from results of mass spectral library searches. *J. Am. Soc. Mass Spectrom.* **5** 316–323.

[24] Stein, S. E. and Scott, D. R. (1994). Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **5** 859–866.

[25] Washburn, M. P., Wolters, D. and Yates, J. R. III (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19** 242–247.

[26] Wu, X., Tseng, C.-W. and Edwards, N. (2007). HMMatch: Peptide identification by spectral matching of tandem mass spectra using hidden Markov models. *J. Comput. Biol.* **14** 1025–1043.

[27] Yates, J. R. III, Morgan, S. F., Gatlin, C. L., Griffin, P. R. and Eng, J. K. (1998). Method to compare collision-induced dissociation spectra of peptides: Potential for library searching and subtractive analysis. *Anal. Chem.* **70**(17) 3557–3565.

[28] Zhang, X., Li, Y., Shao, W. and Lam, H. (2011). Understanding the improved sensitivity of spectral library searching over sequence database searching in proteomics data analysis. *Proteomics*, doi:10.1002/pmic.201000492.

Henry Lam
Department of Chemical and Biomolecular Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
China
E-mail address: kehlam@ust.hk