Critical Analysis of the Computational Methods used to Discover Biomarkers to assist in the Early Detection of Disease

Vishnu Patankar
Biochem 218,
March 2005

1. Introduction

Using proteins as biomarkers has long been considered a promising clinical diagnostics approach for drug discovery and development. Some biomarkers, such as prostate-specific antigen, have been in use for many years. Many other potential biomarkers are being reported in the literature almost weekly, although few have been translated into the diagnostic arena. Progress has not been as rapid as we would like, especially given the advances in our understanding of the process by which disease develops and becomes lethal [45]. Bio-software has a pivotal role to play here - for e.g., to leverage knowledge gained from work with tumors where the biology is known and apply this to the complex proteomics of serum and other body fluids.

In this paper, we analyze the two most popular computer database search algorithms used in protein identification but will begin with a few preliminaries and motivation for the analysis.

1.1 Early Detection of Disease

What follows are study results from recent clinical and clinico-algorithmic studies relating to the value of biomarkers and the role played by computer algorithms in the early detection of disease.

1.1.1 Clinical study

Heart Disease: Levels of a specific protein biomarker in the blood could predict the risk of heart attack or death in those with coronary heart disease. The protein called placental growth factor (PGF) is known to trigger inflammation within hardened and narrowed coronary arteries. A recent study [43] suggests that PGF's presence could perhaps be used as a 'marker' for prognosis in heart disease. Levels of PGF were measured in a group of 547 patients with known heart disease. PGF was also measured in another group - of 626 patients presenting with acute chest pain in an emergency department. In those with heart disease, elevated PGF indicated an increased risk of heart attack or death within 30 days. In those with chest pain, raised PGF meant a three fold increased risk of heart attack or death. The study conclude that PGF is indeed a valuable biomarker for heart attack or heart death and that therapies targeting the inflammatory action of PGF would be a good approach for treating heart disease.

1.1.2 Clinico-Algorithmic studies

Prostrate cancer: The prostate-specific antigen test has been a major factor in increasing awareness and better patient management of prostate cancer (PCA), but its lack of specificity limits its use in diagnosis and makes for poor early detection of PCA. Identifying better biomarkers for early detection of PCA using protein profiling technologies can simultaneously resolve and analyze multiple proteins. Evaluating multiple proteins will be essential to establishing signature proteomic patterns that distinguish cancer from noncancer as well as identify all genetic subtypes of the cancer and their biological activity. One study [41] used a protein biochip surface enhanced laser desorption/ionization mass spectrometry approach coupled with an artificial intelligence learning

algorithm to differentiate PCA from noncancer cohorts. A blinded test set, separated from the training set by a stratified random sampling before the analysis, was used to determine the sensitivity and specificity of the classification system. A sensitivity of 83%, a specificity of 97%, and a positive predictive value of 96% for the study population and 91% for the general population were obtained when comparing the PCA versus noncancer (benign prostate hyperplasia/healthy men) groups.

Ovarian cancer: Another study [42] used proteomic patterns in serum that distinguish neoplastic from non-neoplastic disease within the ovary. A training set of spectra derived from analysis of serum from 50 unaffected women and 50 patients with ovarian cancer were analyzed by an iterative searching algorithm that identified a proteomic pattern that completely discriminated cancer from non-cancer. The discovered pattern was then used to classify an independent set of 116 masked serum samples: 50 from women with ovarian cancer, and 66 from unaffected women or those with non-malignant disorders. The algorithm identified a cluster pattern that, in the training set, completely segregated cancer from non-cancer. The discriminatory pattern correctly identified all 50 ovarian cancer cases in the masked set, including all 18 stage I cases. Of the 66 cases of non-malignant disease, 63 were recognized as not cancer. This result yielded a sensitivity of 100% (95% CI 93--100), specificity of 95% (87--99), and positive predictive value of 94% (84--99). These findings justify a prospective population-based assessment of proteomic pattern technology as a screening tool for all stages of ovarian cancer in high-risk and general populations.

1.2 Proteomics

Proteomics is the systematic study of the many and diverse properties of proteins in a parallel manner with the aim of providing detailed descriptions of the structure, function and control of biological systems in health and disease. Advances in methods and technologies have catalyzed an expansion of the scope of biological studies from the reductionist biochemical analysis of single proteins to proteome-wide measurements. Proteomics and other complementary analysis methods are essential components of the emerging 'systems biology' approach that seeks to comprehensively describe biological systems through integration of diverse types of data and, in the future, to ultimately allow computational simulations of complex biological systems.

1.3 Protein Sequencing in relation to DNA Sequencing

Forward Genetics a key element of reductionist research approaches in the 1980s attempted to move from an observed phenotype or function to the relevant genes and their products that caused that phenotype.

Reverse Genetics benefited from the advent of large-scale sequencing projects and their results [1] catalyzing the development of *reverse* approaches, which attempted to move from the gene sequence to function and phenotype. Such approaches included the observation of clusters of mRNA species showing coordinated expression patterns in different cellular states, either by expression arrays or by serial analysis of gene expression (SAGE [2]).

The rapid identification of proteins was limited only by our capacity to extract sequence information from proteins and peptides, and to correlate this information with the sequence databases. Mass spectrometry and database search algorithms fill this gap.

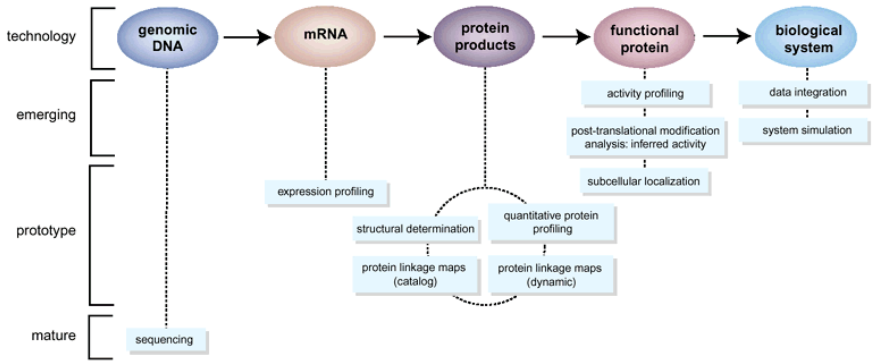

Figure 1: The current status of proteomic technologies.

The different data typically collected in proteomic research and the available technologies are listed. The relative maturity of the proteomic technologies and other key discovery science tools is apparent from the position of the respective technology on the graph.

2. Protein identification methodology

Broadly, two steps constitute the methodology used to identify proteins - Mass Spectrometry and Database Search. A protein mixture is digested, and the resulting peptides are analyzed by MS/MS to obtain experimental spectra. Search programs find database candidate sequences whose theoretical spectra are compared to the experimental spectrum. The best match (highest-scoring candidate sequence) defines the identified database peptide and the corresponding database protein. Validation software then determines whether the peptide and protein identifications are true or false.

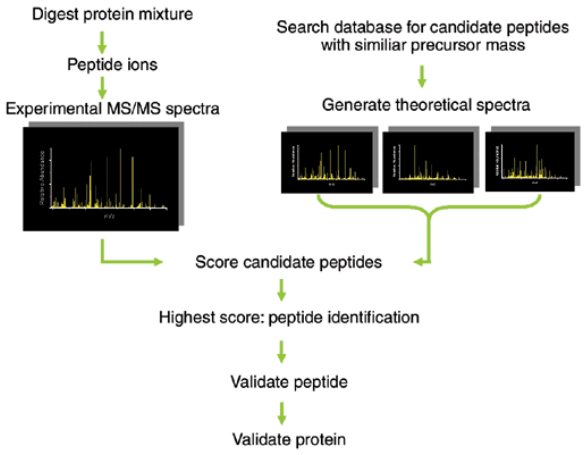

Figure 2. Overview of the protein identification process.

2.1 Mass spectrometry

A mass spectrometer measures the mass-to-charge ratio of charged species under vacuum and comprises an ionization source and a mass analyzer. In the late 1980s, two methods were developed that allowed the

'ionization' of peptides and proteins at high sensitivity and without excessive fragmentation. These breakthroughs were electrospray ionization (ESI [3]) and matrix-assisted laser desorption ionization (MALDI [4]), which had closely followed the development of laser desorption [5-6]. The success of these ionization methods in analytical protein chemistry led to the development of commercial mass spectrometers equipped with robust ESI or MALDI 'ion source' instruments, which rapidly penetrated the protein chemistry community.

The intrinsic mass of a eukaryotic protein is not a uniquely identifying feature. It was quickly recognized, however, that the masses of the various peptides generated by fragmentation of an isolated protein with an enzyme of known cleavage specificity could uniquely identify a protein. Because peptide ions fragment in a sequence-dependent manner, the MS/MS spectrum (two stage mass filter that results in fragment ion spectra) of a peptide, in principle, represents its amino acid sequence. Developments in instrument control software facilitated computer-controlled ion selection, such that MS/MS spectra could be generated from many peptide ions in a given sample without the need for operator intervention, effectively automating the process.

Hunt and co-workers [7] laid the groundwork for a gel-independent approach to proteomics by demonstrating the ability of LC-MS/MS systems to handle extremely complex peptide mixtures. Antigen-presenting lymphocytes continually digest proteins and present some of the resulting peptides bound to major histocompatibility complex (MHC) proteins for immune surveillance. Hunt used immunoprecipitation to isolate the peptide-MHC complexes, extracted the antigenic peptides and subjected the complex peptide mixtures to successive LC-MS/MS analyses. They also used a specific cytotoxic T cell response as a bioassay to confirm the presence of antigenic peptides in each fraction and correlated this functional data with the mass spectrometric data, thereby identifying the sequence of the antigenic peptides [8-9].

2.2 Database Search

In 1993, five independent reports were published that described the implementation of this insight in database search algorithms [7-11]. These algorithms, together with MALDI-TOF mass spectrometry peptide analysis, constituted a 'protein identification' method that is now known as peptide mass mapping (or peptide mass fingerprinting PMF). In this type of analysis, the collected 'MS spectra' are used to generate a list of proteolytic (peptide) fragment masses, which are then matched to the masses calculated from the same photolytic digestion of each entry in a sequence database, resulting in identification of the target protein. The success of this type of analysis is dependent on the specificity of the enzyme used (most frequently trypsin), the number of peptides identified from each protein species, and the mass accuracy of the mass spectrometer. Owing to its increasing sensitivity and ease of use, MALDI-TOF mass spectrometry has become the method of choice for protein identification by peptide mass mapping and is commonly used for identifying proteins separated by 2DE.

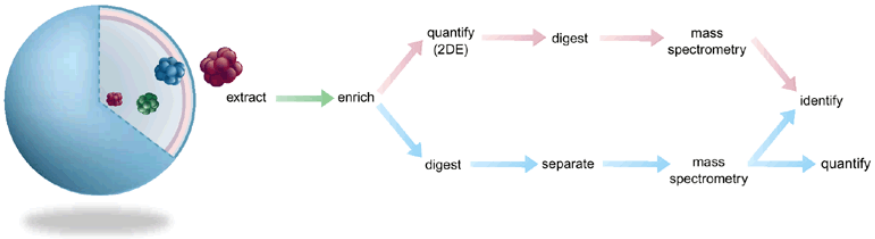

Figure 3. Quantitative protein analysis from the cell to the identified protein.

The two most common processes for quantitative proteome analysis are shown. In the first (top), 2DE is used to separate and to quantify proteins, and selected proteins are then isolated and identified by mass spectrometry. In the second (bottom), LC-MS/MS is used to analyze enzyme digests of unseparated protein mixtures, and accurate quantification is achieved by labeling the peptides with stable isotope. Both processes are compatible with protein fractionation or separation methods, such as subcellular fractionation, protein complex isolation and electrophoresis and chromatography, thereby providing additional biological context to the protein samples being analyzed.

The combination of LC-MS/MS and sequence database searching has been widely adopted for the analysis of

complex peptide mixtures generated from the proteolysis of samples containing several proteins. This approach is often referred to as 'shotgun' proteomics, and has the ability to catalog hundreds, or even thousands, of components contained in samples isolated from very different sources. A tryptic digest of the proteome of a typical human cell will therefore generate a peptide mixture containing at least hundreds of thousands of peptides. Even the most advanced LC-MS/MS systems cannot resolve and analyze such complexity in a reasonable amount of time.

For proteomic studies applying a forward (*function to sequence*) approach, determination of the sequence of the target proteins is usually a defined end point, because detailed functional analyses of the isolated species precede sequence analysis. For studies that apply reverse (*sequence to function*) approaches, knowing the sequence of the proteins in a sample is necessary but not sufficient. Reverse approaches, which are used in many proteomic studies, typically involve quantitative comparison of the protein profiles expressed by cells or tissues in different states.

The most valuable information on the system being studied is obtained from those proteins that are expressed differentially in a matrix of proteins of unchanged expression; therefore, proteomic technologies detecting differences in protein profiles need to be quantitative. Unfortunately, peptides analyzed in a mass spectrometer will produce different specific signal intensities depending on their chemical composition, on the matrix in which they are present and on other poorly understood variables. Thus, the intensity of a peptide ion signal does not accurately reflect the amount of peptide in a sample; in other words, mass spectrometry is inherently not a quantitative technique. However, two peptides of identical chemical structure that differ in mass because they differ in isotopic composition are expected, according to stable isotope dilution theory, to generate identical specific signals in a mass spectrometer.

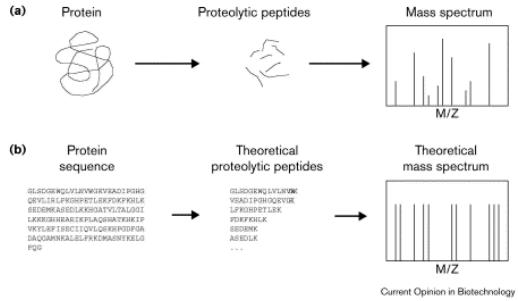

Figure 4. Protein identification using peptide mapping information.

(a) Proteins are digested with an enzyme and the masses of the proteolytic peptides are measured with mass spectrometry. (b) In the database search, each protein sequence in the database is digested according to the specificity of the enzyme. The masses of the resulting peptides are calculated and a theoretical mass spectrum is constructed. The measured mass spectrum is compared with the theoretical mass spectrum and a score qualifying the comparison is calculated. The protein sequences in the database are sorted according to the score and the protein sequence with the best score is selected.

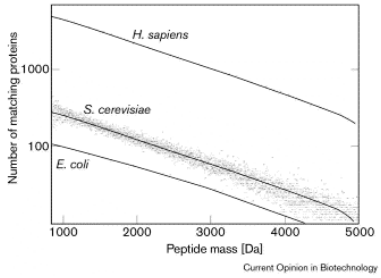

Figure 5. Information content in the mass of a single tryptic peptide.

For Escherichia coli (~4000 open reading frames [ORFs]), Saccharomyces cerevisiae (~6000 ORFs), and Homo sapiens (~100,000 ORFs), at a mass accuracy of 0.5 Da. For S. cerevisiae, the number of proteins at every mass unit is shown together with a smooth curve fitted to the data. For E. coli and H. sapiens, only the smooth fits are shown for clarity.

The success of protein identification by peptide mapping is a result of certain characteristics of proteins, including the limited number of proteins for each organism, the large differences in amino acid sequence, and the large mass difference between different amino acids. The figure above shows the number of proteins in different organisms that match the mass of a single tryptic peptide, indicating that a measurement of a few tryptic peptides is sufficient for identification of a protein when the genome sequence is available. Recent improvements in instrumentation have made it possible to determine peptide masses with a higher mass accuracy, which has improved the success rate for protein identification by peptide mapping. Other information that can be used to improve the quality of identifications includes amino acid composition, number of exchangeable hydrogens and partial amino acid sequence. The searches are usually restricted with additional information, such as species or taxonomic category, protein mass, and protein isoelectric point. Although peptide mapping is usually applied to pure proteins, the constituents of simple protein mixtures can also be identified by peptide mapping.

Peptide mapping has a high success rate for identifying simple protein mixtures from microorganisms with fully sequenced genomes; however, when studying mammals the success rate is presently considerably lower. The success rate of peptide mapping will increase in the near future when the human and, soon after, the mouse genomes will be completed. In the cases where peptide mapping does not provide sufficient information for confident identification, it is necessary to obtain more information. The most common method is to isolate ions corresponding to a proteolytic peptide in the mass spectrometer, fragment them by collisional excitation, and measure the masses of the fragment ions to obtain partial sequence information. The measured fragment mass spectrum is compared to theoretical mass spectra calculated from the protein sequences in the database [22-23].

3. Complexity comparison: Proteomics vs. Genomics

In comparison to its nucleic acid–based counterpart, genomics, the experimental complexity of proteomics is far greater. The technology is also not as mature and, owing to the lack of amplification schemes akin to PCR, only proteins isolated from a natural source can be analyzed. Proteomic analyses are therefore generally limited by substrate. The complexities of the proteome arise because most proteins seem to be processed and modified in complex ways and can be the products of differential splicing; in addition, protein abundance spans a range estimated at five to six orders of magnitude for yeast cells [12] and more than ten orders of magnitude for human blood serum—for example, from interleukin-6 at 2 pg/ml [13] to albumin at 50 mg/ml [14]. Thus, the relatively low number of human genes predicted from the genome sequence [15-16] has the potential to generate a proteome of enormous and as yet undetermined complexity.

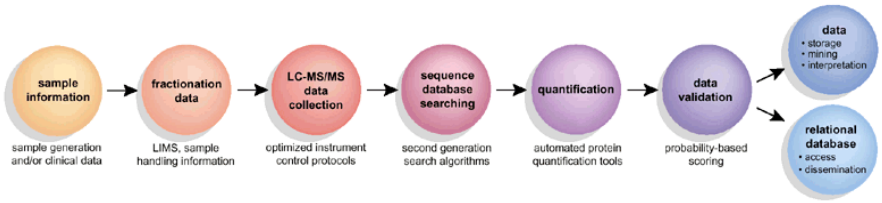

Figure 6. Quantitative proteomics and informatics.

Brief descriptions of the informatics requirements for each of the processes of biological analysis are listed. Handling these data requires significant computational infrastructure if it is to be carried out repeatedly on a large scale. Many of the algorithms used in the process are still not mature.

4. Database Search

Of the two steps used in Protein Identification, in this paper as mentioned in Section 2, we will critically analyze the components of the second step - Database Search.

An unintended consequence of whole-genome sequencing has been the birth of large-scale proteomics. What drives proteomics is the ability to use mass spectrometry data of peptides as an 'address' or 'zip code' to locate proteins in sequence databases. Two mass spectrometry methods are used to identify proteins by database search methods. The first method uses a molecular weight fingerprint measured from a protein digested with a site-specific protease [1-5]. A second method uses tandem mass spectra derived from individual peptides of a digested protein [6-7]. Because each tandem mass spectrum represents an independent and verifiable piece of data, this approach to database searching has the ability to identify proteins in mixtures, enabling a rapid and comprehensive approach for the analysis of protein complexes and other complicated mixtures of proteins [6,8-12]. New biology has been discovered based on fast and accurate protein identification [13-18]. As tandem mass spectral protein identification has proliferated, it has become increasingly important to understand the rationale of individual database search algorithms, their relative strengths and weaknesses, and the mathematics used to match sequence to spectrum.

4.1 Database Search Query

Experimental mass spectra of peptides are the main input to the database query. They are of two types - Peptide Spectra and Peptide Fragmentation Spectra.

4.1.1 Peptide mapping spectra

The simplest and most obvious scoring method for peptide mapping is to count the number of measured peptide masses that correspond to calculated peptide masses in the theoretical mass spectrum of each protein in the database. Several software tools are available on the Internet that use this method of ranking the proteins in the database according to the number of matching peptides, for example, PepSea [24], PeptIdent/MultiIdent [25 and 26], and MS-Fit [27]. This simple scoring method works well for high-quality experimental data, but has the disadvantage that it usually gives higher scores to larger proteins because the probability of random matching is higher. More sophisticated methods for identifying proteins are all based on counting the number of measured peptide masses that correspond to calculated peptide masses but they attempt to make better use of the mass spectrometric information compensating, for example, for effects of protein size [28, 29, 30, 31 and 32]. This usually leads to methods that are more selective and sensitive.

MOWSE [28] (http://srs.hgmp.mrc.ac.uk/cgi-bin/mowse and also implemented in MS-Fit at http://prospector.ucsf.edu/ucsfhtml3.2/msfit.htm) uses average properties of the proteins in the database to improve the sensitivity and selectivity of the identification. It takes into account the relative abundance of the peptides in the database when calculating the score, that is, the chance of getting a random match to a larger peptide is lower and therefore it will contribute to a higher degree to the score. Also the protein size effect is compensated for.

ProFound [29] is an expert system for protein identification using Bayesian theory to rank the protein sequences in the database by their probability of occurrence. It takes into account detailed information about each individual protein sequence in the database and allows for incorporation of additional experimental information (e.g. amino acid composition or sequence information) when available. In addition, empirical information about patterns observed for the distribution of proteolytic peptides along the protein sequence is included in the algorithm. One advantage of the Bayesian approach is that different types of information can be included in a natural way and therefore it is possible to make optimal use of all available information and increase the sensitivity and selectivity of the algorithm. ProFound can also be used to identify simple protein mixtures. A two-step approach is used where first the proteins in the database are ranked according to how well they match the experimental data assuming a single protein is present. In the second step, the top ranking proteins are fused together pairwise, in groups of three, and so on. These fusion proteins are then ranked according to how well they match the experimental data.

Mascot [30] is based on the MOWSE algorithm but in addition it uses probability-based scoring. The probability that the observed match between experimental data and a protein sequence is a random event is approximately

calculated for each protein sequence in the database. The proteins are then ranked with decreasing probability of being a random match to the experimental data.

PeptIdent2 [32] is an algorithm that has been optimized using a genetic algorithm. PeptIdent2 is a generic algorithm with many coefficients and does not incorporate any knowledge about protein properties. The coefficients are optimized using a training set of protein mass spectra. This is a very different approach than that of ProFound, MOWSE, and Mascot, where the algorithms are based on either our knowledge of the properties of individual proteins or database averages.

4.1.2 Peptide fragmentation spectra

In contrast to mass spectra of peptide maps, which contain global information about a protein, peptide fragmentation mass spectra contain rich information on a small section of a protein. The information on the sequence of each peptide enables the identification of a protein from a single peptide. This allows searching of databases that contain incomplete gene information, for example, expressed sequence tags (ESTs). The use of peptide fragmentation mass spectra is also the method of choice for identifying complex protein mixtures. There are several approaches to using peptide fragment information for protein identification.

PepSea [24] uses information from fragmented proteolytic peptides. First, a peptide sequence tag has to be extracted. A peptide sequence tag is a short partial amino acid sequence of a proteolytic peptide together with information of the mass of the peptide and the masses of the parts of the peptide that have not been sequenced. This approach is very fast but requires extraction of the peptide sequence tag prior to searching.

SEQUEST [33-36] uses data from un-interpreted peptide fragment mass spectra (i.e. the information from the whole mass spectrum is used). A cross-correlation function is calculated between the measured fragment mass spectrum and the protein sequences in the database. The cross-correlation function is used to score the proteins in the database. SEQUEST supports the use of information from several fragment mass spectra in the database search. This approach does not require extraction of any information from the mass spectra but the searches are time consuming.

PepFrag [37] and MS-Tag [27] use peptide fragment mass information in combination with other mass spectrometric information, such as amino acid composition, to identify proteins.

Mascot [30] (http://www.matrixscience.com/cgi/search_form.pl?SEARCH=MIS) uses the same probability-based scoring algorithm for fragment information as for peptide maps. It also supports the use of information from several fragment mass spectra in the database search.

4.2 Database Search Algorithms

Four basic approaches have been developed to model matches to sequences [37]:

4.2.1 Descriptive models

Descriptive algorithms are based on a mechanistic prediction of how peptides fragment in a tandem mass spectrometer, which is then quantified to determine the quality of the match between the prediction and the experimental spectrum. Mathematical methods such as correlation analysis have been used to assess match quality. SEQUEST is based on one such model.

4.2.2 Interpretative models

Interpretative approaches are based on manual or automated interpretation of a partial sequence from a tandem mass spectrum and incorporation of that sequence into a database search. Matches between the sequence and the spectrum have been scored using probabilities or correlation methods.

4.2.3 Stochastic models

Stochastic models are based on probability models for the generation of tandem mass spectra and the fragmentation of peptides. Basic probabilities of fragment ion matches are obtained from training sets of spectra of known sequence identity. Stochastic models use statistical limits on the measurement and fragmentation process to create a likelihood that the match is correct.

4.2.4 Statistical and probability models

Statistical and probability models determine the relationship between the tandem mass spectrum and sequences. The probability of peptide identification and its significance are then derived from the model. MASCOT is based on one such model.

4.3 Search Engine Algorithm comparison: MASCOT vs. SEQUEST

The most commonly used algorithms for mass spectrometry based protein identification are MASCOT, MS-Fit, ProFound and SEQUEST. Due to space considerations, this paper will compare and contrast two of the four algorithms referring to a third algorithm ProFound as relevant. MASCOT and SEQUEST and chosen to allow broad sampling of models.

4.3.1 MASCOT

This group of methods uses models based on empirically generated fragment ion probabilities [45,48,51]. In these methods no a priori determined probabilities are used. They generate a model that relates the sequences to a spectrum and determine the peptide identification score from this model. Thus, in the simplest models the frequencies of matches of b- and y-ions are determined and used to calculate a probability of sequence identification determined by the product of probabilities of its fragment matches. Several variations of this approach have been implemented in database searching algorithms [43, 45, 48, 51]. Mascot [41] uses a model analogous to the one previously developed for identifying proteins from their peptide mass fingerprint. Mascot may also use some empirical observations about fragment intensities and ion series continuity. The actual description of the model is not available in peer-reviewed literature and therefore we are not able to describe this algorithm in detail, even though it is one of the most widely used database search programs. But MASCOT had its origins in the MOWSE algorithm which we will briefly detail.

The first stage of a Mowse search is to compare the calculated peptide masses for each entry in the sequence database with the set of experimental data. Each calculated value which falls within a given mass tolerance of an experimental value counts as a match. A molecular weight range for the intact protein can be used as a pre-filter. Rather than just counting the number of matching peptides, Mowse uses empirically determined factors to assign a statistical weight to each individual peptide match. The matrix of weighting factors is calculated during the database build stage, as follows:

A frequency factor matrix, F, is created, in which each row represents an interval of 100 Da in peptide mass, and each column an interval of 10 kDa in intact protein mass. As each sequence entry is processed, the appropriate matrix elements $f_{i,j}$ are incremented so as to accumulate statistics on the size distribution of peptide masses as a function of protein mass. The elements of F are then normalized by dividing the elements of each 10 kDa column by the largest value in that column to give the Mowse factor matrix M:

$$m_{i,j} = \frac{f_{i,j}}{\left| f_{i,j} \right|_{\max \text{ in column } j}}$$

After searching the experimental mass values against a calculated peptide mass database, the score for each entry is calculated according to:

$$\text{Score} = \frac{50{,}000}{M_{\text{Prot}} \times \prod_{n} m_{i,j}}$$

Where MProt is the molecular weight of the entry and the product term is calculated from the Mowse factor elements for each match between the experimental data and peptide masses calculated from the entry.

Probability Based Mowse
Mascot incorporates a probability based implementation of the Mowse algorithm. The Mowse algorithm is an excellent starting point because it accurately models the behavior of a proteolytic enzyme. By casting the Mowse score into a probabilistic framework, there are a number of additional benefits:

- A simple rule can be used to judge whether a result is significant or not.
- Different types of matching (peptide masses and fragment ions) can be combined in a single search.
- Scores from different searches and on different databases can be compared.
- Search parameters can be optimized more readily by iteration.
- Matches using mass values (either peptide masses or MS/MS fragment ion masses) are always handled on a probabilistic basis. The total score is the absolute probability that the observed match is a random event. Reporting probabilities directly can be confusing. Partly because they encompass a very wide range of magnitudes, and also because a "high" score is a "low" probability, which can be ambiguous. For this reason, scores are reported as -10*LOG10(P), where P is the absolute probability. A probability of 10-20 thus becomes a score of 200.

Significance Level
Given an absolute probability that a match is random, and knowing the size of the sequence database being searched, it becomes possible to provide an objective measure of the significance of a result. A commonly accepted threshold is that an event is significant if it would be expected to occur at random with a frequency of less than 5%. This is the value which is reported on the master results page.

It is important to distinguish between a significant match and the best match. Ideally, the correct match is both the best match and a significant match. However, significance is a function of data quality. It may be that there are just not enough mass values or the mass measurement accuracy is not good enough to get a significant match. This doesn't mean that the best match isn't correct, it just means that you must study the result more critically.

The best match is still correct, but it is barely significant. If we did 20 such searches, we could expect to get this score by chance alone because there is such a huge number of entries in the sequence database. If the search is repeated once more, but with a mass tolerance of ±2.0 Da, the match is lost. None of the scores are significant and the correct match drops to third place. Fortunately, it is clear from the significance level that this is not a reliable match, and there is no danger of this result becoming a false positive.

Expectation Values
Each protein score in a peptide mass fingerprint, and each ions score in an MS/MS search, is accompanied by an expectation value. This is the number of matches with equal or better scores that are expected to occur by chance alone. It is directly equivalent to the E-value in a Blast search result. For a score that is exactly on the default significance threshold, ($p<0.05$), the expectation value is also 0.05. Increase the score by 10 and the expectation value drops to 0.005. The lower the expectation value, the more significant the score.

Mass Tolerances
The score in a peptide mass fingerprint is usually inversely related to the mass tolerance, as shown in the example above. This is not always the case for an MS/MS ions search, where increasing the peptide mass tolerance may have little effect on the score. This is because most of the discrimination comes from the MS/MS fragment ion matches. Opening up the peptide mass tolerance means that Mascot has to test many more peptides, (and so the search takes longer!), but the major contributions to the final score, the MS/MS fragment ion matches, are unchanged.

In fact, if the peptide mass tolerance is set too tightly, in an effort to improve discrimination, one or more of the peptide matches may be lost, which will dramatically reduce the overall score.

Limitations
Like any statistical approach, the Probability Based Mowse algorithm depends on assumptions and models.

One of these assumptions is that the entries in the sequence databases are random sequences. This is not

always a good assumption. Some of the most glaring examples involve extended repeats, such as AAC62527, porcine submaxillary apomucin. Although the molecular weight of this protein is 1.2 MDa, over 80% of the sequence is composed of an identical 7 kDa repeat. It is difficult to know how to treat such cases. If a single experimental peptide mass is allowed to match to multiple calculated masses, then a single experimental mass which matches within a repeat will give a huge and meaningless score. But, if duplicate matches are not permitted, it will be virtually impossible to get a match to such a protein because the number of measurable mass values is too small to give a statistically significant score.

Another assumption is that the experimental measurements are independent determinations. This will not be true if the data include multiple mass values for the same peptide, even if these are from ions with different charge states in an electrospray LC-MS run. Good peak detection and thresholding (in both mass and time domains for LC-MS) are essential for any scoring algorithm to give meaningful results.

Sequence Query Scoring

Amino acid sequence or composition information, if present, is treated as a rigorous filter on the candidate sequences. Ambiguous sequence or composition data can be used (in a manner similar to a regular expression search in computing) but it still functions as a filter, not a probabilistic match of the type found in a BLAST or FASTA search.

Recently, a group of database search algorithms have been implemented that use collective properties of database sequences to calculate the probability that a sequence match is a random event. Thus, we have proposed to divide all database fragment ions into two groups: matches and misses46. Then, we assume that a hypergeometric probability models the frequencies of database peptides based on the number of matches. According to this model a probability that a peptide match is a random event is predicted from the hypergeometric probability of choosing K1 matches (number of matches of a peptide) in N1 trials (the number of fragment ions of the peptide) from a pool of fragments consisting of N fragments (number of all database fragments) K of which are matches (number of matches of all fragment ions to a spectrum). The hypergeometric probability of this event is:

$$P_{K,N}(K_1,N_1)=\frac{C_K^{K_1}\times C_{N-K}^{N_1-K_1}}{C_N^{N_1}}$$

The probability of a peptide being a random match to the tandem mass spectrum is defined in the space that comprises all peptides whose mass match the mass of the precursor peptide. The significance of a peptide match is determined as a type I error of the null hypothesis—all fragment matches are random. OMSSA, a recently developed database search algorithm, uses a similar approach, where the peptide matches are modeled after the Poisson distribution. Database search algorithms based on the number of matches trend to spectral quality owing to the fact that a match to a background peak and a match to a sequence ion are not distinguishable. Statistical models produce a statistical confidence for a match between the spectrum and database sequences. This confidence is based on the frequency of fragment ions in the database itself, and the probability a spectrum is a random match rather then the closeness of fit to a fragment model.

4.3.2 SEQUEST

SEQUEST [38] is an example of a program that uses a descriptive model for peptide fragmentation and correlative matching to a tandem mass spectrum. It uses a two-tiered scoring scheme to assess the quality of the match between the spectrum and amino acid sequence from a database. The first score calculated, the preliminary score (*Sp*), is an empirically derived score that restricts the number of sequences analyzed in the correlation analysis. *Sp* sums the peak intensity of fragment ions matching the predicted sequence ions and accounts for the continuity of an ion series and the length of a peptide. The original *Sp* score is:

$$S_p=\left(\sum_k I_k\right) m(1+\beta)(1+\rho)/L$$

where the first term in the product is the sum of ion abundances of all matched peaks, m is the number of matches, is a 'reward' for each consecutive match of an ion series (for example, 0.075), is a 'reward' for the presence of an immonium ion (for example, 0.15) and L is the number of all theoretical ions of an amino acid sequence.

The second score is a cross-correlation of the experimental and theoretical spectra. This score is referred to as *XCorr*. The theoretical spectrum is generated from the predicted fragment ions, the b- and y-ions for each of the sequences. In the theoretical spectrum the main ion series products are assigned an abundance of 50, a window of 1 atomic mass unit around the main fragment ions is assigned intensity 25, and water and ammonia losses are assigned intensity of 10. The theoretical and normalized experimental spectra are cross-correlated to obtain similarities between the spectra. First, a cross-correlation of the two discrete data sets, experimental data (E) and theoretical spectrum (T), is taken:

$$Corr(E,T)=\sum_{I=0}^{N-1} x_i y_{i+\tau}$$

The correlation is processed and averaged to remove the periodic noise in the interval of (-75 to 75). In addition to the preliminary and cross-correlations scores, SEQUEST produces another important quantity, normalized difference of Xcorr values between the best sequence and each of the other sequences. This value, *Cn*, is important in distinguishing the best match from other lower-scoring matches. That is, Cn is useful to determine the uniqueness of the match. If a match is reasonably unique to a sequence, the *Cn* value will be large (>0.1). *XCorr* is independent of database size and reflects the quality of the match between spectrum and sequence, whereas *Cn* is database dependent and reflects the quality of the match relative to near misses.

The cross-correlation score is a sensitive measure. However, like other measures based on additive features, it is dependent on peptide mass, charge state and spectral quality. Thus it has been observed that larger peptides score higher than similar-quality smaller peptides. Very dense (potentially noisy) spectra can have high cross-correlation scores. To address these issues, a few modifications have been made to the cross-correlation score. To normalize XCorr for spectral noise and peptide size, the XCorr value is divided by auto-correlation of the experimental spectrum or by the square root of the products of auto-correlations of experimental and theoretical spectra. A statistical confidence can then be readily derived from the normalized cross-correlation scores. SEQUEST has been shown to have good sensitivity and flexibility and is applicable to data generated by different types of mass spectrometers.

4.4 Comparison: MASCOT vs. SEQUEST

In prior sections, we have looked in detail at the qualitative comparison between MASCOT and SEQUEST. We will now focus on objective and experimental results obtained in a study that help to evaluate MASCOT and SEQUEST head-to-head.

4.4.1 Evaluation Methodology and Criteria

The software tools for protein identification using mass spectrometric information will give a top-ranking candidate even if all the matching peptides are random matches. It is important to determine the quality of the identification, that is, what the probability is that the identified protein is a false positive.

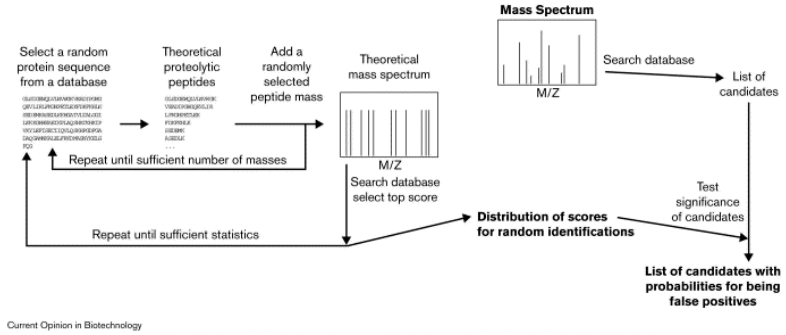

Figure 7. Simulations provide a method for determining the quality of the search results [41].

One method for assessing this is by using simulations [36]. In the simulations, protein sequences were randomly selected from a protein sequence database, digested according to the specificity of an enzyme, a single peptide was randomly chosen, and its mass calculated and stored. This procedure was repeated and a theoretical mass spectrum was constructed. This theoretical mass spectrum was then used in a database search and the top score was saved. The protein sequence with the highest score was in nearly all cases a false positive, that is, the peptide matches were random. These searches were repeated with different theoretical mass spectra and a distribution of scores for random identification was obtained. Subsequently, the distribution of scores for random identification can be used to assess the quality of the results when experimental data is used in a database search, that is, each protein candidate in the list can be associated with a probability for it being a false positive. Other methods are attempts at directly calculating the probability that the masses observed in a mass spectrum would correspond to proteolytic peptides from a protein sequence. The direct calculations are, however, less reliable than the simulation because it is necessary to make approximations because of the complexity of the process. Objective methods for assessing the quality of search results have become more important as high-throughput proteome analysis is becoming more widespread [36, 42, 43 and 44].

The software tools for protein identification have matured and the algorithms have been refined to give higher selectivity and sensitivity. High-throughput analysis has become increasingly common in proteome projects and requires automatic analysis of the mass spectrometric data. An important part of automation is quality control and therefore development of methods to determine the quality of the search results has become a focus.

In [39] commonly used algorithms for mass spectrometry based protein identification, Mascot, MS-Fit, ProFound and SEQUEST, were studied in respect to the selectivity and sensitivity of their searches. The influence of various search parameters were also investigated. Approximately 6600 searches were performed using different search engines with several search parameters to establish a statistical basis. The applied mass spectrometric data set was chosen from a current proteome study. As a side effect, they present a software solution for fully automated triggering of several peptide mass fingerprinting (PMF) and peptide fragmentation fingerprinting (PFF) algorithms. The development of this high-throughput method made an intensive evaluation based on data acquired in a typical proteome project possible. Previous evaluations of PMF and PFF algorithms were mainly based on simulations. The system setup was a classic 3 tier - Web Server, Middle-Tier Application Server and a Back-End SQL Server.

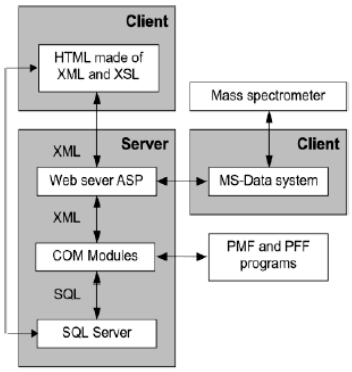

Figure 8. Multi-tier Evaluation setup.

Mouse brain samples were obtained from a European study for mapping genes on chromosomes. MALDI-TOF MS analyses were used. The data set consisted of 89 different PSD spectra. In all searches, the nondedundant NCBI database consisting about 600,000 entries were used. Most parameters were constant between PMF and PFF search engines.

4.4.2 Experimental Results

The results were assessed automatically and evaluation was based on all matches that were reported on rank one by the PMF and PFF search programs. All other ranks were ignored. PMF search results were considered correct of the sequences of the matched peptides corresponded to the previously identified reference protein sequence. PFF search results were considered correct of the sequence of the found peptide or one of its isobaric derivations was contained in the reference protein.

Generally, an important criteria for judging performance of PMF programs is their ability to report true positives. Plots of scoring distributions are shown below. MASCOT (53.0 % or 89 proteins identified correctly) and ProFound (53.6 % or 90 proteins identified correctly)clearly separated true positives by their search scores and besides a few exceptions, MASCOT and ProFound identified the same set of proteins. Both provide score values that correspond to a significance level of 5%. Dotted lines show correct identification and solid lines show incorrect identification.

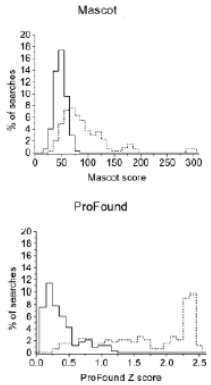

Figure 9. Comparison of PMF algorithm performance.

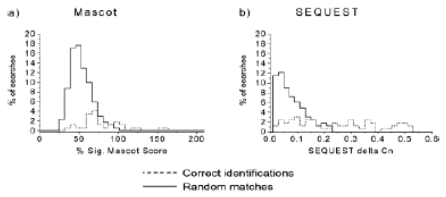

Figure 10. Comparison of PFF algorithm performance.

In the PFF category, 89 different MALDI-PSD data sets were used. Identical search parameters were used for both programs. MASCOT identified 16 proteins (17.9%). The above plots show the comparison of true positives and false positives. MASCOT reported a significance level of 5%. Only 3 true positives had a score above the 5% significance level. SEQUEST search results are ranked by cross-correlation of virtually reconstructed mass spectra of protein database entries and the acquired mass spectral data. The differences between the normalized cross-correlation parameter of the first and second ranked amino acid sequences (delta *Cn*) are said to be a trend to cutoff true positives from false positives. SEQUEST identified 33 proteins (37%) and showed a more obvious separation of true and false positives as compared to MASCOT. Of note is the fact that MASCOT identified one protein that SEQUEST did not. However, with SEQUEST there is no significance level or probability reported to attach value to a particular identification.

4.4.3 Effect of Search Parameters

Preselection of proteins/Species Filtering
Both algorithms allow preselection of proteins from the sequence database either by choice of taxonomy ('all', 'animals', 'mammals', 'mouse') or by restrictions to the protein's pI and Mr range. MASCOT is sensitive to the number of preselected sequences. Though phylogenetic trees are used to divide the FASTA sequence databases, poorly standardized conventional names and locations of species information causes MASCOT and ProFound to have different effective databases for the algorithms to work on. Restriction of the number of database entries input to the algorithms is key to speed up and find true matches for both algorithms. SEQUEST's correlation values delta Cn change significantly with restrictions but the normalized correlation values remain relatively the same.

Mass accuracy parameter
Both algorithms allow for tolerances in mass accuracy 0.7 - 1.6 Da. MASCOT is sensitive to mass accuracies worse than 50 ppm i.e. there is a spike in false positives reported with this level of mass accuracy. Changing the mass tolerance parameter for fragment ions in SEQUEST had little effect on the search result because this parameter only influences the preliminary scoring function which selected peptides for the cross-correlation.

Parent ion mass tolerance
Peptide mass tolerance has a contained effect on scoring in an MS/MS ion search because most of the discrimination is due to fragment ion matches. Increasing peptide mass tolerance simply increases the number of peptides that have to be tested. However, search speed and the risk of obtaining high scoring false positives

decreases. Both algorithms performed best at 0.3 Da. And in the case of good quality spectra, small changes to parent ion mass tolerance are usually not crucial. On the other hand, if the peptide can be modified, searching with no restriction on the parent ion tolerance is the only chance to jump out of a local optimum.

Variable modifications
Both algorithms allow for variable amino acid modifications. Both algorithms reported more false positives when variable modifications were allowed (MASCOT to a greater extent) and this is attributable to the geometric increase in the number of virtually digested peptides. SEQUEST is more robust to variable modifications. However, this strategy allows for getting out of local optimums and reporting edge true positive cases.

Allowed missed cleavage sites
MASCOT performed better when one missed cleavage site was allowed. The noise generated when two missed cleavage sites was allowed reported too many false positives. Although, this strategy allows for greater sequence coverage. If search speed is not crucial, allowing two missed cleavage sites is a good starting point.

4.4.4 Conclusion and Algorithmic Research Challenges

SEQUEST performed better overall although a combination of various search algorithms via voting might be a sophisticated approach towards higher confidence in a database search result.

Automated analysis of tandem mass spectra is a critical process for new analytical strategies such as 'shotgun proteomics'. As tandem mass spectrometers have improved, the acquisition of hundreds of thousands of spectra has become not uncommon, and thus, accurate approaches to identify and validate sequence matches will make this method all the more powerful. Although a variety of algorithms have been demonstrated to provide accurate matches between tandem mass spectra and sequences, all suffer from an inability to provide verifiable matches to poor-quality spectra. Reliable and sensitive methods to assess spectral quality and assign quality indices to spectra will be critical for decreasing computational load and lowering false-positive rates. Most algorithms are very accurate for peptides that follow general rules of fragmentation, but a subset of amino acid sequences and more highly charged peptide ions deviate from these rules; thus, a better understanding of relationships between peptide sequences and fragment ion intensity will assist in designing better models for matching spectra to sequences. Additional studies to better understand the strengths and weaknesses of the various algorithms will help to design algorithms with better sensitivity and selectivity.

- False positives are a perpetual concern in database searching. They can arise for several reasons. Data-dependent algorithms for large-scale acquisition of tandem mass spectra do not discriminate between peptide ions and other types of ions that may be present. Thus, search algorithms are often confronted with a collection of spectra that could be single peptide ions, chemical noise, nonpeptide molecules or mixtures of correlating isobaric peptides, which are then matched to amino acid sequences. Good data preprocessing or a search of a library of contaminants can help remove nonpeptide spectra prior to a search.

- Peptides are often present at a wide range of concentrations in a sample, and peptides present at the limit of detection can produce poor quality fragmentation. The issue of sensitivity is more difficult to correct as it is heavily dependent on the limit of detection of a mass spectrometer. The effects can range from incomplete dissociation to poor ion statistics for fragment ions, making them indistinguishable from noise. In these cases incomplete fragmentation patterns or poor signal-to-noise ratios may lead to a solution that is not unique or correct.

- The chemistry of peptide fragmentation is also not completely understood, and thus, fragmentation models used in database searching may not accommodate aberrant fragmentation processes and result in false positives. Several statistical studies of peptide fragmentation have been performed to better understand the contributions of specific amino acids to fragmentation processes. In time, improved models will account for more of the aberrant fragmentation processes.

- Sequence conservation can lead to confusing results. If the same peptide sequence exists in multiple

proteins, all of the proteins will be identified. Without additional peptide data it would be impossible to determine which protein produced the peptide that generated the tandem mass spectrum. Identifying this situation is straightforward, as most algorithms track all proteins that a spectrum matches.

- A final possibility, and perhaps of more concern, are amino acid sequences that do not produce a unique fragmentation pattern but share enough of the same fragment ions to be indistinguishable from one another. In these cases a unique amino acid sequence can not be determined directly from the fragmentation pattern and other means are required to determine the absolute identity of the peptide. In particular, small peptides, less than eight amino acids in length, may not produce a fragmentation pattern that achieves a unique result.

6. Future Trends

Four main challenges to be addressed in order for proteomics to have a substantial impact on eukaryotic biology within the systems biology model.

- The first challenge is the enormous complexity of the proteome. For some proteins, in excess of 1,000 variants (splice and translation isoforms, differentially modified and processed species) have been described. The detection, and particularly the molecular analysis of this complexity, remains an unmatched task.

- The second challenge is the need for a general technology for the targeted manipulation of gene expression in eukaryotic cells. An approach that has proved successful for the systematic analysis of biological systems relies on iterative cycles of targeted perturbations of the system under study and the systematic analysis of the consequences of each perturbation. Although recent advances in using RNA interference in higher eukaryotic cells open up exciting possibilities, the general targeted manipulation of biological systems in these species remains unsolved.

- The third challenge is the limited throughput of today's proteomic platforms: iterative, systematic measurements on differentially perturbed systems demand a sample throughput that is not matched by current proteomic platforms.

- The fourth challenge is the lack of a general technique for the absolute quantification of proteins. The ability to quantify proteins absolutely, thereby eliminating the need for a reference sample, would have far-reaching implications for proteomics—from the determination of the stoichiometry of protein complexes to the design of clinical studies aimed at discovering diagnostic markers.

Fortunately, proteomics will have an impact on clinical and biological research well before these challenges are met. It is expected that precise clinical diagnosis based on highly discriminating patterns of proteins in easily accessible samples, particularly body fluids, may be the area in which proteomics will make its first significant contribution [41, 42].

References

[1] Adams, M.D. et al. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature 377, 3-174 (1995).
[2] Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Serial analysis of gene expression. Science 270, 484-487 (1995).
[3] Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F. & Whitehouse, C.M. Electrospray ionization for mass spectrometry of large biomolecules. Science 246, 64-71 (1989).
[4] Karas, M. & Hillenkamp, F. Laser desorption ionization of proteins with molecular masses exceeding 10000 daltons. Anal. Chem. 60, 2299-2301 (1988).
[5] Tanaka, K., Ido, Y., Akita, S., Yoshida, Y. & Yoshida, T. Detection of high mass molecules by laser desorption time-of-flight mass spectrometry. In Proc. 2nd Japan-China Joint Symp. Mass Spectrom. (eds. Matsuda, H. &

Xiao-tian, L.) 185-188 (Osaka, Japan, 1987).
[6] Tanaka, K. et al. Protein and polymer analyses up to m/z 100,000 by laser ionization TOF-MS. Rapid Commun. Mass Spectrom. 2, 151-153 (1988).
[7] Henzel, W.J. et al. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. Proc. Natl. Acad. Sci. USA 90, 5011-5015 (1993).
[8] Mann, M., Hojrup, P. & Roepstorff, P. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biol. Mass Spectrom. 22, 338-345 (1993).
[9] Pappin, D.J.C., Hojrup, P. & Bleasby, A.J. Rapid identification of proteins by peptide-mass fingerprinting. Curr. Biol. 3, 327-332 (1993).
[10] James, P., Quadroni, M., Carafoli, E. & Gonnet, G. Protein identification by mass profile fingerprinting. Biochem. Biophys. Res. Commun. 195, 58-64 (1993).
[11] Yates, J.R., III, Speicher, S., Griffin, P.R. & Hunkapiller, T. Peptide mass maps: a highly informative approach to protein identification. Anal. Biochem. 214, 397-408
[12] Gygi, S.P., Rochon, Y., Franza, B.R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. Mol. Cell. Biol. 19, 1720-1730 (1999).
[13] Lai, R. et al. Prognostic value of plasma interleukin-6 levels in patients with chronic lymphocytic leukemia. Cancer 95, 1071-1075 (2002).
[14] Ritchie, R.F., Palomaki, G.E., Neveux, L.M. & Navolotskaia, O. Reference distributions for the negative acute-phase proteins, albumin, transferrin, and transthyretin: a comparison of a large cohort to the world's literature. J. Clin. Lab. Anal. 13, 280-286 (1999).
[15] Venter, J.C. et al. The sequence of the human genome. Science 291, 1304-1351 (2001).
[16] Lander, E.S. et al. Initial sequencing and analysis of the human genome. Nature 409, 860-921 (2001).
[17] Appella, E., Padlan, E.A. & Hunt, D.F. Analysis of the structure of naturally processed peptides bound by class I and class II major histocompatibility complex molecules. EXS 73, 105-119 (1995).
[18] Hunt, D.F. et al. Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. Science 255, 1261-1263 (1992).
[19] Henderson, R.A. et al. HLA-A2.1-associated peptides from a mutant cell line: a second pathway of antigen presentation. Science 255, 1264-1266 (1992).
[20] Hunt, D.F. et al. Peptides presented to the immune system by the murine class II major histocompatibility complex molecule I-Ad. Science 256, 1817-1820 (1992).
[21] Adam, B.L. et al. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. Cancer Res. 62, 3609-3614 (2002).
[22] M. Mann and M. Wilm, Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem 66 (1994), pp. 4390–4399.
[23] J.K. Eng, A.L. McCormack and J.R. Yates, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spec 5 (1994), p. 976.
[24] M. Mann, P. Hojrup and P. Roepstorff, Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biol Mass Spectrom 22 (1993), pp. 338–345.
[28] D.D.J. Pappin, P. Højrup and A.J. Bleasby, Rapid identification of proteins by peptide-mass finger printing. Curr Biol 3 (1993), pp. 327–332.
[29] W. Zhang and B.T. Chait, ProFound — an expert system for protein identification using mass spectrometric peptide mapping information. Anal Chem 72 (2000), pp. 2482–2489.
[30] D.N. Perkins, D.J. Pappin, D.M. Creasy and J.S. Cottrell, Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20 (1999), pp. 3551–3567.
[31] P. Berndt, U. Hobohm and H. Langen, Reliable automatic protein identification from matrix-assisted laser desorption/ionization mass spectrometric peptide fingerprints. Electrophoresis 20 (1999), pp. 3521–3526.
[32] P.R. Gras, M. Muller, E. Gasteiger, S. Gay, P.A. Binz, W. Bienvenut, C. Hoogland, J.C. Sanchez, A. Bairoch, D.F. Hochstrasser et al., Improving protein identification from peptide mass fingerprinting through a parametized multi-level scoring algorithm and an optimized peak detection. Electrophoresis 20 (1999), pp. 3535–3550.
[33] P.R. Griffin, M.J. MacCoss, J.K. Eng, R.A. Blevins, J.S. Aaronson and J.R. Yates, III, Direct database searching with MALDI-PSD spectra of peptides. Rapid Commun Mass Spectrom 9 (1995), pp. 1546–1551.
[34] J.R. Yates, III, J.K. Eng and A.L. McCormack, Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. Anal Chem 67 (1995), pp. 3202–3210.
[35] J.Rd. Yates, J.K. Eng, A.L. McCormack and D. Schieltz, Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal Chem 67 (1995), pp. 1426–1436.
[36] Eriksson, B.T. Chait and D. Fenyö, A statistical basis for testing the significance of mass spectrometric

protein identification results. Anal Chem 72 (2000), pp. 999–1005.
[37] Rovshan G Sadygov, Daniel Cociorva & John R Yates III, Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book
, Nature Methods 1, 195 - 202 (Nov 2004)
[38] SEQUEST patent: United States Patent 6,017,693 Yates, III , et al. January 25, 2000
[39] Chamrad DC, Korting G, Stuhler K, Meyer HE, Klose J, Bluggel M. Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. Proteomics, 2004 Mar;4(3):619-28.
[40] Schirmer EC, Florens L, Guan T, Yates JR 3rd, Gerace L., Nuclear membrane proteins with potential disease links found by subtractive proteomics.Science. 2003 Sep 5;301.
[41] Adam BL, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, Semmes OJ, Schellhammer PF, Yasui Y, Feng Z, Wright GL Jr. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. Cancer Res. 2002 Jul 1;62(13):3609-14.
[42] Petricoin, E.F. et al. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 359, 572–577 (2002).
[43] Medical Research Lab, Johann Wolfgang Goethe University, Frankfurt, Germany. Journal of the American Medical Association 28th January 2004 Volume 291 pages 435-441.
[45] Eschenbach A, Director, National Cancer Institute, Clinical Proteomics: Developing Standardized Tools for Cancer Research, March 2005.