Computational diagnostics based on proteomic data- review on approaches and algorithms

Sreevatsa A.N.[1], Badrunnisa S.*[2], Shaukath Ali M.[3], Vinitha R Pai[4]
[1]BITM, Bellary, Karnataka, India, sreevatsa.agrahara@gmail.com
[2]Department of Biotechnology, BITM, Bellary, Karnataka, India, badrunnisa.s@gmail.com
[3]College of applied medical science, Jazan University, Jizan, shaukath.m@gmail.com
[4]Department of Biochemistry, Yenopoya Medical College Darelakatte Mangalore, Karnataka, India, vinitharpai@gmail.com

Abstract - Protein identification using mass spectrometry is an indispensable tool for proteomics which in recent days has evolved to give better understanding of the biology of cell and its functioning. Proteomics has wide application in diagnosing diseases such as cancer, Alzheimer's disease etc. The data obtained from the diagnostic tools like LC-MS is to be interpreted accurately so as to obtain the correct qualitative and quantitative information about the peptides present in the biological sample. Such interpretation requires and exhaustive knowledge and review about different tools that can be employed and their comparison. This article focuses on comparison of different proteomic tools available for the MS data processing and interpretation. The accuracy demanded during protein identification can be fulfilled by tag based approaches, than PMF or PFF systems. Although, there is a need of standardized matrices for the comparison of the protein identification tools, identifying the single best package for each application from the available literature is at present extremely difficult as each package has its own advantage over other.
The datasets and thresholds used in these kinds of comparisons have a critical importance on the outcome of such experiments, and that the high variability in machine and experimental setups complicates analysis. The state of data standards and lack of benchmarks therefore makes it difficult to make an effective comparison. While the increasing availability of data in public repositories and tightening standards will no doubt ameliorate the problem, until this basic benchmarking problem is overcome, no single package or approach can conclusively be declared to outperform all others, expect, perhaps, in the specific circumstances used in particular studies.
Key words: Review on approaches and algorithms, diagnostic tools, Protein identification tools.

Introduction

Early diagnosis many diseases are currently an essential requirement in the medical sector so as to provide patient an appropriate medication before the disease reaches its chronic stage. Ovarian cancer is one such disease whose early diagnosis is currently unavailable and the disease is diagnosed only when it reaches its chronic stage. Biomarkers are the biological compounds which allow us to diagnose the disease, presence of these biomarkers or presence in higher concentration above normal refers to the presence of the suspected disease in the patient. Discovery and identification of such biomarkers are very much important in diagnostic sector. These biomarkers can be any biological compounds like proteins, hormones, biochemical compounds etc which can be linked with the presence of a particular specific disease. Protein identification is a key and essential step in the field of proteomics. The examination of patterns of protein expression alone can, of course, lead to important discoveries, including, for example, classification of samples on the basis of a particular pattern. However, without identifying the proteins known to be critically involved in the system under investigation, it is not possible to develop into the biological explanation for these patterns to develop hypotheses as to the underlying biology of the system of interest. Thus, while protein identification may often be overlooked or taken for granted, it remains the key initial step in elucidating the biology of an organism by studying its protein expression. Our ability to maximize the benefit of proteomics to life science research is often dependent on our ability to accurately, quickly, and completely identify the full complement of proteins found in our samples of interest. Proteomic tools like LC/GC-MS are extensively used for biomarker discovery or diagnosis. The speed and accuracy of these machines make them amenable to the high-throughput applications required not just in proteomics, but also in many other areas of the life sciences, resulting in rapid developments in hardware, software, and data management in the last decade. When we consider the use of mass spectrometers for protein identification, these rapid developments have lead to a bewildering number of instrument configurations, analysis algorithms, and data formats. This insight into protein identification algorithms is important because often the results may be ambiguous, and the biases chosen to make the problem computationally tractable can radically affect the result. The data obtained by the equipment being very large as the biological samples may contain a very huge number of proteins (of orders of 109) and varies with the source of the sample collected. Despite the improvements in mass spectrometry hardware and the reliability of modern protein identification software, several

studies involving a range of mass spectrometers, datasets, and identification algorithms have shown in each case that fewer than half of the proteins in a complex proteomic sample can be identified. Given the critical role of protein identification in proteomic analysis, this review aims to explore this apparent upper limit on the effectiveness of current protein identification algorithms and to give relevant background information and practical suggestions to computational biologists and life scientists so the best possible protein identifications can be realized.

Review of the tools used

The overall process of computational diagnosis is done by collecting data obtained from the proteomic tools such as LC-MS, MS/MS, SELDI,MALDI etc. and subjecting the data for processing wherein the data is refined to reduce noise levels. Secondly the processed data is subjected to identification of the peptides fragments; the peptide information so obtained gives even quantitative information of the protein present in the sample after its quantification. The consistent and transparent analysis of LC/MS and LC/MS/MS data requires multiple stages. Overview of disease diagnosis steps involved, using proteomic tools indicated in the figure (1). Modules solving each one of these tasks should be integrated into a linear process like the Trans-Proteome Pipeline, which allows smooth processing of the data through the different stages (4). These stages are explained below.

1. DATA PROCESSING
2. PEPTIDE IDENTIFICATION AND VALIDATION
3. PROTEIN IDENTIFICATION AND VALIDATION
4. QUANTIFICATION

In this review we try to find on the efficiency of various scoring systems and approaches used during peptide/protein identification process.

Peptide identification and validation

Scoring system

Scoring system is very important in protein identification. Mass spectrometric data of an unknown protein is compared with theoretical data of known protein, and a score is assigned on how well the two data compare. If the score is above an arbiter threshold then it is called “hit”, if it is below the threshold value then the protein remains unidentified. Scoring system has been developed by adapting preexisting statistical models such as Bayesian probability [1, 2], expectation maximization [3], and machine learning [4], to name just a few. Progressively more sophisticated scoring systems have since been built by improving and combining standard scoring systems and by introducing novel statistical and search methods [5, 6].

The limiting factor on all protein identification tools is the tradeoff between false positives and false negatives. It is absolutely essential to keep false positives to a minimum during protein identification because identifying the wrong protein can lead to a costly waste of time and resources. At the same time, it is clearly desirable to identify as many proteins as possible to draw maximum benefit from the experimental data. The ability of the system to identify a protein is called sensitivity and ability to differentiate between correct true positive and false positive is called specificity. A researcher should bear in mind that the threshold value given determines a balance between specificity and sensitivity of the system.

As may be expected, there is a tradeoff between the two, embodied in a numerical threshold often called the confidence level, above which proteins are classed as identified. This is important for the researcher to bear in mind, since the balance between sensitivity and specificity will have a bearing on the threshold above which they are prepared to accept a protein as ‘‘identified.’’ For example, Chen et al. [7] report results for the popular peptide fragment fingerprinting (PFF) package called Mascot in a large cross-species study identifying human proteins in Escherichia coli databases using data collected on a high performance LC-MS/MS LCQ ion trap mass spectrometer. They find correct proteins to have scores between 20 and 117, and incorrect proteins to have scores of up to 60. This demonstrates a fundamental property of protein identification software. As shown in Figure 1, the separation of true from false protein identifications based on a score is never perfect, and the general effectiveness of all protein identification algorithms should be viewed with this in mind.

Mass based approaches

In this approach every protein in the database is theoretically subjected to similar environment as the sample protein undergoes. The sample protein undergoes enzymatic digestion and secondary fragmentation this allows to obtain mass spectrum of all the protein fragments in the database. These theoretical mass spectra are compared with the experimental spectrum. In theory, any method of comparison between two spectra can be a candidate for a scoring system, and in practice a variety of methods are used. Most basic method for comparison, one can consider the shared peak count. The shared peak count, as the name implies, counts the number of peaks in the same position (shared) in both the experimental and theoretical spectra. The theoretical spectrum with the highest shared peak count is then said to be the closest match. Peptide mass fingerprinting this method uses theoretical spectra each comprising the list of

masses expected by an enzymatic digestion of each protein sequence in the reference database. The experimental spectrum consists of the masses of the digested protein fragments detected by the mass spectrometer. PMF is popular and works well because it is relatively fast to search using PMF data against a database. With high quality sampling, PMF can in many cases produce protein identifications with high confidence, especially in organisms with smaller genomes. . Unfortunately, sometimes a sample spectrum does not resemble any theoretical spectra in the protein database closely enough to make a confident identification. This can happen for many reasons, such as unexpected post-translational or chemical modifications, splice variants, individual sequence variants (single nucleotide polymorphisms [SNPs], etc), or omissions and errors in the database. As more sophisticated methods for scoring PMF have been developed, more proteins can now be identified with confidence. This corresponds to a better separation of true from false positives using the scoring system. The next step is to specify a threshold above which the protein is termed to be identified. Determination of a threshold value should be carefully done because if the threshold is set low then the false positives may be getting identified and the result will be corrupted, whereas a high threshold masks the correct protein from getting identified. Statistical methods have been developed so as to find a correct threshold value which can identify correct proteins [8, 9]. Limitation of PMF is that the database size affects the sensitivity of the method. Most of the software use PMF as a first screening method wherein if the method succeeds in identifying protein then the result is accepted and if not other method such as PFF is employed. Most popular packages are definitely easier to understand and use due to user friendly graphical interface used. Few of such popular packages are introduced below. Aldente [10] is hosted on the ExPASy Proteomics Server as one of a suite of bioinformatics tools. Released in 2004, Aldente uses a robust Hough transform to speed searches and find straight lines hidden in the data, making this tool more robust to noise than other PMF packages. A number of additional constraints can be input by the user, such as isoelectric point and molecular weight to restrict the effective database size. Unlike most other PMF packages, the user is able to select the parameters contributing to the final score and their proportions in order to ‘‘fine-tune’’ the search engine to a particular experiment. The details of this tunable scoring scheme are available on the ExPASy Web site [11] along with supporting documentation. A threshold for identification is set after processing random sequences in the same parent mass range. The random sequence with the highest score becomes the threshold above which a protein is said to be identified. Mascot [2] uses a proprietary scoring algorithm but is known to be based on the MOWSE algorithm [12], first described in 1993. By calculating the distribution of tryptic peptide lengths across the entire search database, a probability can be calculated for each observed peak for this match being purely random. Perkins et al. [2] describe in general terms the basis of this probabilistic scoring system, giving the user of this package an insight into how to interpret data generated via Mascot: ‘‘The fundamental approach is to calculate the probability that the observed match between the experimental data set and each sequence database entry is a chance event. The match with the lowest probability is reported as the best match. Whether this match is also a significant match depends on the size of the database. To take a simple example, the calculated probability of matching six out of ten peptide masses to a particular sequence might be 10_5. This may sound like a promising result but, if the real database contains 106 sequences, several scores of this magnitude may be expected by chance. A widely used significance threshold..... is p, 0.05.For a database of 106 entries, this would mean that those with significant matches were those with probabilities of less than 5 3 10_8. .... we have adopted a convention often used in sequence similarity searches, and report a score which is 10Log10 (P), where P is the probability. A significant match is typically a score of the order of 70.’’ This means searches in smaller protein databases, such as bacterial databases, will generally have lower threshold scores for confidence than those conducted in larger databases for higher organisms. We can also infer that for noisy experimental spectra, for example those with contamination, these extra peaks contribute to the possibility of a random match, and thus raise the confidence score threshold for a given probability. Mascot automatically returns a score threshold with its results calculated to represent a confidence level of p, 0.05. MASCOT search for smaller databases such as bacterial databases will generally have lower threshold scores for confidence than those conducted in larger databases for higher organisms. Example of input data format is available at matrix science website [13]. MS-Fit [14] is also a probabilistic algorithm, again based on MOWSE, but runs over FASTA format [15] databases. MS-Fit first bins proteins according to the parent mass weight. Within each of these bins, a series of bins are created according to the tryptic peptide masses. This is done so that when calculating the probability of a random tryptic peptide match, it is calculated specifically for the distribution of these peptide masses for a given parent mass,

effectively reducing the size of the search databases MS-Fit also allows for the input of a number of possible contaminant masses. This allows the user to pre-filter any likely contaminants from the spectrum, thus increasing the quality of the spectrum against which a search is to be performed.

Profound [10] uses a Bayesian probability scoring system to score hits, using additional information outside of the set normally used by PMF algorithms, such as enzyme cleavage chemistry information, provisions for the knowledge that particular amino acids are present (or absent) in the sample protein, and previous experiments on the sample protein. Each piece of information functions as an additional constraint upon the search space of database proteins, therefore reducing the effective size of the database against which the search is conducted. Profound uses Gaussian distributed measurement errors in the probability calculations to more closely model real error, as opposed to the simple bounded ‘‘tolerance’’ error measurements used in other PMF algorithms. A recent study by Chamrad et al. [10] using matrix-assisted laser desorption–ionization time-of-flight (MALDI-TOF) mass spectrometry data from a project mapping genes onto mouse chromosomes used expert interpretation of the spectra to identify 70% of the proteins, thus forming a reference set for PMF algorithm comparison. This study found the performance of Mascot and Profound to be similar, correctly identifying around 53% of proteins from the reference set at a 5% significance level, with MS-Fit identifying only 32% using the same input parameters. This study also looked at the effects of various parameters for Mascot and Profound queries. Profound performs better over the entire range of parameter settings including taxonomy restriction, mass accuracy variation, variable modifications, and missed cleavages. Mascot showed a slightly better performance only in the case where mass accuracy was better than 25 ppm. Overall, Profound identified slightly more proteins, showed a better separation between true and spurious identifications, and generated not a single random match above the 5% significance level throughout the experiment. A study in yeast using 266 spectra gathered on MALDI-TOF instruments from three different manufacturers [2] found Mascot to outperform Profound, with Mascot identifying 45% of proteins while Profound identified 33% of the proteins. However, Mascot did give a single false positive identification, while Profound did not. This suggests that lifting the threshold for a Mascot identification to avoid all false positives, thus making the results comparable with the Profound result, would reduce the percentage of proteins identified using Mascot. The claims for the highest rate of protein identifications belong to groups using consensus methods. These methods submit the query data independently to multiple search engines and combine the results. The rationale for this process is that marginal identifications may be corroborated or rejected by complementary packages. Experimentalists have been doing this independently for some time [16], usually in an ad hoc manner. However, recently, more rigorous statistical methods have been applied to the integration of the scores returned by each engine. One well-known example is ProteinScape [17]. This software is designed to accommodate a number of different proteomics workflows, including 2-D LC-MS/MS, LC–eletrospray ionization, and LCMALDI. ProteinScape’s consensus method claims an increase of identified proteins of up to 10% by taking a meta-score of Profound, Mascot, MS-Fit, and/or other algorithms. Details of the algorithm are not in the public domain, and the vendor provides only a short description of the meta-score as an ‘‘intelligent combination of scoring schemes.’’ Such consensus methods are now being adopted by large-scale projects [18], but are still not popular in smaller labs because these consensus programs are not free, and there are additional complexities in terms of running multiple PMF search engines. Growing advancements in mass spectrometry hardware, database size and computational advances necessitates the need of higher accuracy and sensitivity in protein identification. These needs resulted in emergence of the new high throughput technique called peptide fragment fingerprinting. Fig (2) shows an Example of a PFF spectrum from the HUPO Brain Proteome Project. The bold numbers associated with each peak give the m/z value, while the italic numbers associated with the peak show the intensity value. This ‘‘stick’’ spectrum has been processed from the raw output of the mass spectrometer. Available at http://www.ebi.ac.uk/pride/viewSpectrum.do?mzDataAccession¼1717&spectrumReference¼32494.

Peptide fragment fingerprinting. Approaches using PFF data are the current mainstream of high-throughput protein identification. Proteins are first digested with an enzyme, and then individual peptides are selected to undergo further fragmentation to yield PFF spectra such as the one shown in Figure 2. The set of these spectra, along with information such as the parent mass of these fragmented peptides, are then used in the database search. There are many dozens of scoring systems described in the literature, but in most cases these consist of two steps: (1) attributing a score for each protein in the database and (2) calculating a measure of confidence that the top-ranking identified protein is not a false positive—such as in the case where the protein being investigated does not exist in

the database. PFF is the method of choice for high-throughput applications due to the additional information gained from secondary fragmentation. This information makes the protein identification process less sensitive to effects such as protein modifications and can generate higher statistical confidence in the correct identification than traditional PMF. Some of the more popular PFF packages are listed in Table 1. Sequest and Mascot are arguably the two most popular packages for protein identification using PFF. Mascot is probabilistically based while also using some heuristics to improve scoring, but, like Spectrum Mill, Protein Prospector, and the most recent commercial PFF package, Phenyx, the details of the scoring process have not been published. Mascot, however, is known to be based on the probabilistic MOWSE algorithm [12], which uses the parent mass and the relative abundance of peptide masses for that parent mass as constraints on the search space. More than a decade has passed since the MOWSE algorithm was published, and Mascot now includes parameters not related to the features described in the original paper, such as selecting the type of mass spectrometer the input data comes from. It is therefore impossible to describe in any detail the process by which Mascot scores are generated, and a comparison with other engines can only be made empirically by analysis of the benchmarking papers discussed in this review. Sequest uses a patented scoring algorithm utilizing a cross correlation approach. Figure (2) shows a simplified flowchart of the Sequest peptide identification process as described in U.S. patent 6,017,693. and fig (3) shows Simplified Flowchart for the Sequest Algorithm Showing the Process by which Sequest Provides Scores Used to Identify Peptides Flowchart shows process as described in United States patent 6,017,693 [19]. Note that information from the mass spectrum is used three times: (1) as a filter to select only peptides from the database sharing a similar parent ion mass with the unknown peptide; (2) during a preliminary Sp ‘‘closeness-of-fit’’ filter to select the top 500 peptide candidates; and (3) through a correlation function to produce the final scores.

Tag based approaches

Tag-based approaches begin with an attempt to extract peptide sequence information directly from the peptide fragmentation spectra. These methods are based on casting the problem into one of finding a maximum path length through a graph, a problem already known to have efficient solutions, and are based on a seminal paper by Dancik et al. [20]. The process of inferring protein sequence from MS/MS data is known as de novo sequencing. Due to the high complexity of most MS/MS spectra, de novo sequencing tools often return short, ambiguous sequences known as ‘‘tags.’’ These tags are then searched against a database. Although many of these tags may randomly align with sections of protein sequence right across the genome, the correct protein identification is expected to have multiple alignments with sequence tags derived from the unknown protein. Tag based approaches have been successfully used to identify proteins from larger EST databases that are more inclusive than curated databases. They have also been used for finding homologous proteins in other species [21], an area where mass-based approaches, and particularly PMF, have been shown to have limited applicability [22]. Not surprisingly, tag-based approaches appeared as the first de novo sequencing methods were becoming available [23]. A number of popular packages available for de novo sequence interpretation and subsequent tag-based searching are listed in Table 2.

GutenTag [24] is a popular tag-based package released in 2003 by the same group responsible for the popular PFF package Sequest, and is available free for nonprofit organizations. Lutefisk is available as source code in C, allowing the experimenter to tailor aspects of the scoring function or any other aspect of reporting and calculation. It works by first identifying ‘‘significant’’ ions, followed by the collection of evidence for N- and C-terminal ions from the spectra. A list of candidate sequences is generated for passing onto a tag-based program for alignment of these candidate sequences with proteins in a database. InsPecT [25] is a recently introduced tag-based package based on a probability model for assessing the accuracy of candidate sequence tags. PEAKS are a proprietary package, and as such have not published details of its implementation. MSBLAST and FASTA do not infer de novo sequence but are popular alignment programs evolved from DNA alignment roots. De novo sequencing: PEAKS is currently a standard tool for any de novo sequencing task before its submission to protein sequencing. It is found to be more accurate than any software packages and can outperform Lutefisk. For quadrupole TOF (QTOF) data across a range of spectrum qualities, the authors claim 41% perfectly correct sequences and 94% of sequences to have six consecutive correctly sequenced amino acids. De novo sequencing quality is highly dependent on the precision of the mass spectrometer and the quality of the spectra. Advances in hardware accuracy and precision have a great effect on the ability of de novo algorithms to correctly and accurately infer longer stretches of protein sequence. Quality spectra as well as high precision greatly constrain the possible sequences capable of generating the observed spectrum. Thus, the short list of possible

peptides, to be later submitted to a tag-based search, may contain longer and therefore more specific sequences, resulting in more confident identifications. Preparatory methods to improve the quality of spectra intended for de novo sequencing is an active area of research [26-27].
Tag-based search algorithms: Most current tag-based methods use a basic adaptation of the BLAST [28] or FASTA [29] algorithms. These are already in common use in the life sciences for gene and protein sequence alignments. For use in tag-based searching, the algorithms are modified for the much shorter peptide sequences usually generated by MS/MS, typically in the order of eight to 15 amino acids, and to handle the errors and ambiguities resulting from the alternate possible sequence interpretations when de novo sequencing [30]. Tag-based approaches are much faster than PFF searches; Tanner et al. [25] report a two order of magnitude speed-up over the commonly used Sequest through using their tag based method InsPecT, as well as demonstrating a much better scalability when scaled to include added modifications or protein mixtures. The speed-up is due to a more efficient and sensitive use of tags to exclude the vast bulk of potential protein matches considered in a first pass, although the authors note that performance for single protein identifications is not better than the PFF package X! Tandem in terms of speed and sensitivity, as this package has already incorporated a similar filtering system. Tag-based methods have been designed to function in environments where exact matches are not expected—for example, searching against databases of other species—and as such have different methods for determining the statistical significance of a result under these conditions. [21].
Other available packages: There is a great number and variety of protein identification packages other than those listed in this review. Many of these packages have been tailored to provide identifications for particular classes of proteins, or even glycans [31], or use certain techniques and report superior performance to established general protein identification engines listed in this review for their specific application. A quick survey of other available tools and packages in many cases can turn up software ideal for a particular application. Brief descriptions and intended uses of all the packages listed in Tables 1–3 and others can be found in a review by Shadforth et al. [32]. An exhaustive list of protein identification tools can be found at http://www.molecularstation.com/bioinformatics/link/Proteomics/Protein_Identification_Tools
http://www.proteomesoftware.com/Proteome_software_link_software.html

Quantificatoin

Quantification is a further critical step in biomarker studies because the primary focus is on peptides (proteins) that show differences in expression between two sets of samples; peptides that are invariant present much less interest. Systematic quantification of all peptides across multiple data sets is actually a very demanding task that has not yet been fully resolved. Strategies emphasizing the quantitative aspect tend to decouple identification and quantification and perform two independent experiments. Basically two main approaches have been applied. The first is based on stable isotope labeling and requires derivatization of the peptides from the various samples sets with different reagents that have different isotopic composition. The second approach, which is more relevant to larger biomarker studies (i.e. analysis of a larger sets of samples from normal (control) and disease (or treated) patients), analyzes each sample individually and then compares the multiple LC/MS runs subsequently. By performing all these steps we can identify cancer cells. Biomarker discovery projects (as well as many other proteomics studies) are often large experiments generating large data sets, and results might be obtained from concerted efforts of several laboratories. It is essential that data exchange and sharing becomes a transparent process. Standardization of data through wide use of common formats and use of transparent tools for data processing and analysis with well defined parameters is essential [33,34].

Results

Protein identification tools

Many groups have devised metrics in order to gain better comparison within various protein identification tools such as (1) calculating expectation values for the number of hits expected for a given score [11-36]; (2) the hit-ratio (i.e., the ratio of the peaks submitted in the experimental spectrum matched in the theoretical spectrum); and (3) sequence coverage (i.e., the proportion of the protein sequence covered by the peptides matched between experimental and theoretical spectra) [37]. However such metrics are not encouraged in literatures, even though there is no correct methodology on how to compare such protein identification tools. Finding this difficulty in comparison several journals have started to use standards [38-39] and other are tending to follow such standards. The metrics required to be presented along with protein identifications for these publishers include but are not limited to: (1) supporting information detailing the use of all processing steps, experimental design, scoring methods used, software and database versions, and all parameters used in the search; (2) sequence coverage and/or hit rate; (3) measures of certainty such as p-values;

(4) justifying evidence for identifications made on single peptides, for a particular protein within a protein family, or proteins identified in another species; and (5) multiple replicates for complex analyses. The problem of standardizing mass spectrometry–related data formats and vocabulary is being addressed by the HUPO Proteomics Standards Initiative [18]. This group has released a standard format for encapsulation of peak list data (mzData) and has an alpha version of the successor to this format under development. Known as mzML, this format will merge the competing mzXML and mzData formats. Details of the new format can be found on the mzML development page [40]. The same group is also developing a format known as AnalysisXML for the encapsulation of parameters and results from protein identifications. These formats are enjoying increasing support from instrument manufactures and software vendors, and are rapidly being adopted up by the proteomics community.

Discussion

A consistent message found in this review of protein identification algorithms is that the best results for protein identification are extracted through the use of consensus programs used to collect the results from various packages and distill their results. This is particularly the case in the mass-based approaches of PMF and PFF. Through such methods, the strengths of some packages can be exploited, while weaknesses in others are mitigated. The only difficulty of this method is the expense and difficulty in handling multiple search algorithms and scoring systems. However, the advantage of getting accurately identified protein encourages the use of consensus based programs. All of the methods require the use, at some point, of a reference sequence database for identifying the proteins expressed in the sample. This presents extra challenges for researchers working with less well-characterized species. In this scenario, tag-based methods are preferred because of the reduced computational complexity of searching for diverged proteins. Mass-based methods require matching of peptide or peptide fragment masses to their theoretical equivalents derived from a sequence database. A single amino acid change, with the exception of a change between leucine and isoleucine, will change the mass of the peptide or peptide fragment with a resultant effect on the ability of the algorithm to correctly identify the protein. In contrast, with tag-based methods, particularly if the tag-matching process is tolerant of sequence variation, sequence changes have less of an impact on the ability to correctly match database entries. Thus, cross-species databases can be more effectively used to aid in protein identification. Identifying the single best package for each application from the available literature is at present extremely difficult due to a number of factors. Each package claims advantages over a number of others. These claims are often backed up with compelling results. While some of the comparative studies cited above have produced work of excellent scope and quality, many of these results show marginal differences between packages or show contradictory results to other studies. Furthermore, across all the studies, only a small fraction of the available packages have been considered and evaluated. Similarly contradictory results are reported not only in comparisons between various packages, but also between approaches, such as between mass-based and tag based approaches. This indicates the datasets and thresholds used in such comparisons have a critical importance on the outcome of such experiments, and that the high variability in machine and experimental setups complicates analysis. The state of data standards and lack of benchmarks therefore makes it difficult to make an effective comparison, implying the need for sustained directed research on the creation of suitable benchmarks. While the increasing availability of data in public repositories and tightening standards will no doubt ameliorate the problem, until this basic benchmarking problem is overcome, no single package or approach can conclusively be declared to outperform all others, expect, perhaps, in the specific circumstances used in particular studies. For such benchmarking work to be successful, it is important that it be broad, replicable, and routine, because each software package is constantly evolving, so a benchmark can at best produce a comparison likely to quickly become redundant as newer versions of packages are released. This, in turn, points to the need for an ongoing process of benchmark-based testing, in which new algorithms and techniques, or developments in existing packages, are regularly re-evaluated to measure performance and provide guidance to life-science researchers seeking to extract the most from their proteomic experiments.

At present, the greatest focus in improving protein identification software is on the following: (1) developing better scoring metrics or including additional information [41–42]; (2) improving fragmentation models. The inclusion of new metrics [43] and use of new techniques [44] applied to fragmentation modeling allows for better prediction of theoretical spectra. This, in turn, leads to more discriminating scoring systems; (3) data representations for clustering or filtering to improve speed and efficiency [45, 46]. These methods can massively speed searches by reducing the size of the database being searched through the use of statistical methods to cheaply reject the majority of nonmatching database entries, or by improving the speed at which comparisons can be made.

Acknowledgement: Authors would like to acknowledge the Management of BITM Bellary for the facility and support in this study.

Abbreviations:
LC-MS: liquid chromatography- mass spectrometry
PMF : probability mass function
SELDI: Surface-enhanced laser desorption/ionization
MALDI: Matrix-assisted laser desorption/ionization
LCQ : Last Chance Qualifier
MOWSE: Molecular weight search

References

[1] Zhang W., Chait B.T. (2000) *Anal Chem* 72, 2482–2489.
[2] Perkins D.N., Pappin D.J.C., Creasy D.M., Cottrell J.S. (1999) *Electrophoresis* 20, 3551–3567.
[3] Nesvizhskii AI., Keller A., Kolker E., Aebersold R. (2003) *Anal Chem* 75, 4646–4658.
[4] Gay S., Binz P.A., Hochstrasser D.F., Appel R.D. (2002) *Proteomics* 2, 1374–1391.
[5] Bafna V., Edwards N., (2001) *Bioinformatics* 17, S13– S21.
[6] Zhang Z., Sun S., Zhu X., Chang S., Liu X., et al. (2006) *BMC Bioinformatics* 7, 222.
[7] Chen Y., Kwon SW., Kim SC., Zhao Y. (2004) *J Prot Res* 4, 998-1005.
[8] Magnin J., Masselot A., Menzel C., Colinge J. (2004) *J Prot Res* 3, 55–60.
[9] Ganapathy A., Wan XF., Wan J., Thelen J .,Emerich DW. (2004) *Conf Proc IEEE Eng Med Biol Soc* 2, 3051–3054.
[10] Tuloup M., Hernandez C., Coro I., Hoogland C., Binz P-A., (2003) In: Proceedings of the Swiss Proteomics Society 2003 Congress: pp. 174–176.
[11] Rauch., Bellewm., Fitzgibbon., Holzman T., Hussey P., Igra M., Maclean B., Lin C. W., Detter A., Fang R.., Faca V., Gafkenm P., Zhang H., Whitaker J., States D., Hanash D., Paulovich A., Martin W., and McInstoshm M. W. (2006) J. *Proteome. Res*. 5, 112–121
[12] Pappin DJ., Hojrup P., Bleasby A.J . (1993). *Curr Biol* 3, 327–32.
[13] Matrix Science (1999) Mascot. Peptide mass fingerprinting [computer program]. Available: http://www.matrixscience.com/help/pmf_help.html. Accessed 15 December 2007.
[14] University of California San Francisco (1996) UCSF Protein Prospector version 4.27.1 [computer program]. Available: http://prospector.ucsf.edu. Accessed 15 December 2007
[15] National Center for Biotechnology Information. Fasta format description. Available: http://www.ncbi.nlm.nih.gov/blast/fasta.shtml. Accessed 15 December 2007.
[16] Zhang W., Chait B.T. (2000) *Anal Chem* 72, 2482–2489.
[17] Chamrad DC., Korting G., Stuhler K., Meyer HE., Klose J. (2004) *Proteomics* 4, 619–628.
[18] Bruker Daltonics ProteinScape [computer program]. Available:
a. http://www. proteinscape. com Accessed 15 December 2007.
[19] HUPO Brain Proteome Project. Available http://www.hbpp.org. Accessed 15 December 2007
[20] Yates JR III., Eng JK., inventors; University of Washington, assignee (2000 Jan 25) Identification of nucleotides, amino acids, or carbohydrates by mass spectrometry. United states patent 6,017,693. Available: http://www. patentstorm.us/patents/6017693.html. Accessed 10 December 2007
[21] Dancik V., Addona T.A., Clauser K.R., Vath J.E., Pevzner P.A . (1999) *Proceedings of the Annual International Conference Computational Molecular Biology: RECOMB 1999*. pp. 135–144.
[22] Liska A.J., Sunyaev S., Shilov I.N., Schaeffer D.A., Shevchenko A. (2005) *Anal Chem,* 5, 4118–4122.
[23] Wilkins M.R., Williams K.L. (1997) *J Theor Biol,* 186, 7–15.
[24] Mann M ., Wilm M. (1994) *Anal Chem* 66, 4390–4399.
[25] Frank A ., Tanner S ., Bafna V., Pevzner P. (2005) *J Prot Res,* 4, 1287–1295.
[26] Noga MJ ., Lewandowski J.J ., Suder P ., Silberring J. (2005) *Proteomics,* 5, 4367–4375.
[27] Ullmer R ., Plematl A ., Rizzi A. (2006) *Rap Comm Mass Spectrom,* 20, 1469–1479.
[28] Altschul S.F ., Madden T.L ., Schaffer A.A ., Zhang J ., Zhang Z. (1997) *Nucleic Acids Res,* 25, 3389–3402.
[29] Pearson W.R. (1990) *Methods Enzymol*,183, 63–98.
[30] Han Y ., Ma B ., Zhang K. (2005) *J Bioinform Comput Biol* 3, 697–716.
[31] Joshi H.J ., Harrison M.J ., Schulz B.L ., Cooper C.A ., Packer N.H. (2004). *Proteomics,* 4, 1650–1664.
[32] Shadforth I ., Crowther D ., Bessant C. (2005) *Proteomics,* 5, 4082–4095.
[33] Lie X.-J ., Kemp C. J ., Zhang H ., and Aebersold R. (2005) Mol. Cell. *Proteomics,* 4, 1328–1340
[34] Rauch., Bellewm J ., Fitzgibbon ., Holzman T ., Hussey P., Igra M ., Maclean B ., Lin C.W ., Detter A ., Fang

R ., Faca V ., Gafkenm P ., Zhang H ., Whitaker J ., States D ., Hanash D ., Paulovich A ., Martin W ., and McInstoshm M. W. (2006). J. *Proteome Res*, 5, 112–121

[35] Fenyo D., Beavis R.C. (2003) *Anal Chem,* 75, 768–774.

[36] Eriksson J., Chait B.T., Fenyo D. (2000). *Anal Chem*, 72, 999–1005.

[37] Stead D.A ., Preece .A ., Brown J.P .(2006) *Mol Cell Prot*, 5, 1205–1211.

[38] Carr S ., Aebersold R ., Baldwin M ., Burlingame A ., Clauser K. (2004) *Mol Cel Prot,* 3, 531-533.

[39] HUPO Proteomics Standards Initiative. mzMZL development. Available: http://www.psidev.info/index.php?q¼node/257. Accessed 21 December 2007.

[40] Weatherly DB, Atwood JA, Minning TA, Covala C, Tarleton R. (2005) *Proteomics,* 4, 762–772.

[41] Hogan J.M., Higdon R., Kolker N., Kolker E . (2005) *OMICS* 9: 233–250.

[42] Zhang Z ., (2004) *Anal Chem,* 76, 3908–3922.

[43] Arnold R.J ., Jayasankar N ., Aggarwal D.A ., Tang H ., Radivojac P. (2006) *Pac Symp Biocomp,* 11, 219–230.

[44] Wong J.W.H ., Sullivan M.J ., Cartwright H.M ., Cagney G (2007) *B.M.C Bioinformatics,* 8, 51.

[45] Robertson C ., Cortens J.P ., Beavis R.C. (2005) *Rapid Commun Mass Spectrom,* 19, 1844–1850.

[46] Beer I.Barnea E ., Ziv T ., Admon A. (2004) *Proteomics*, 4, 950–960.

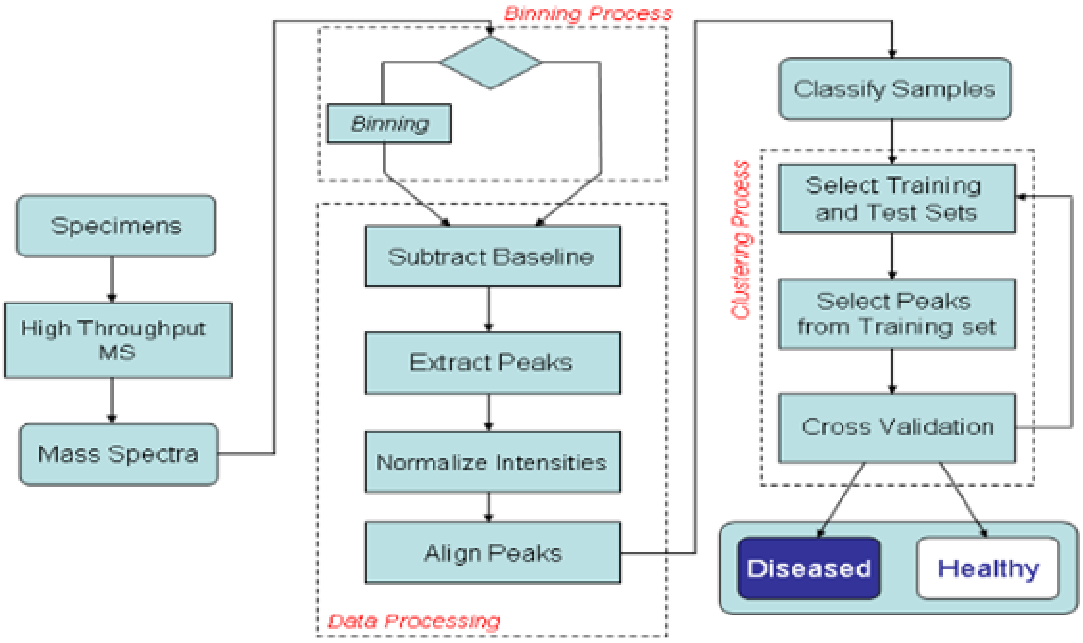

Fig. 1-Overview of disease diagnosis steps involved, using proteomic tools

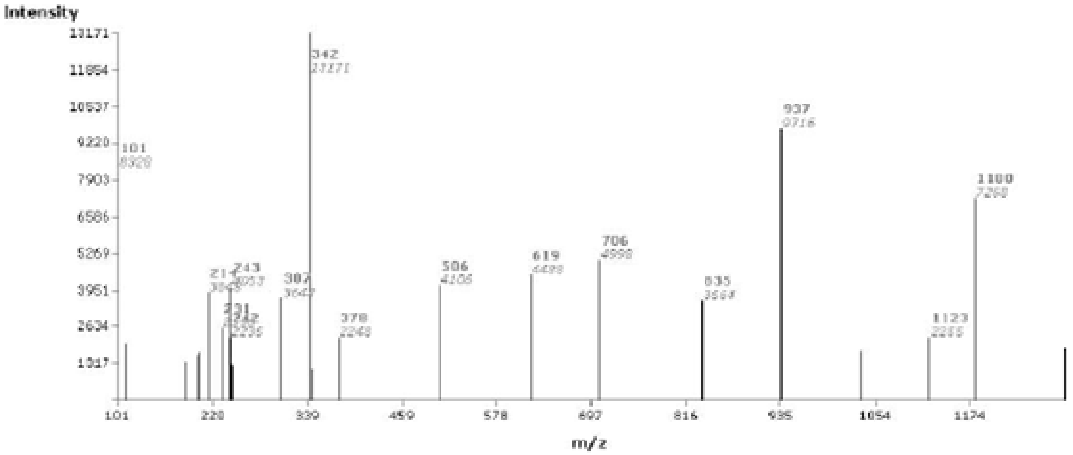

Fig. 2-Example of a PFF spectrum from the HUPO Brain Proteome Project

Protein sample | Mass spectrometer | Experimental spectrum

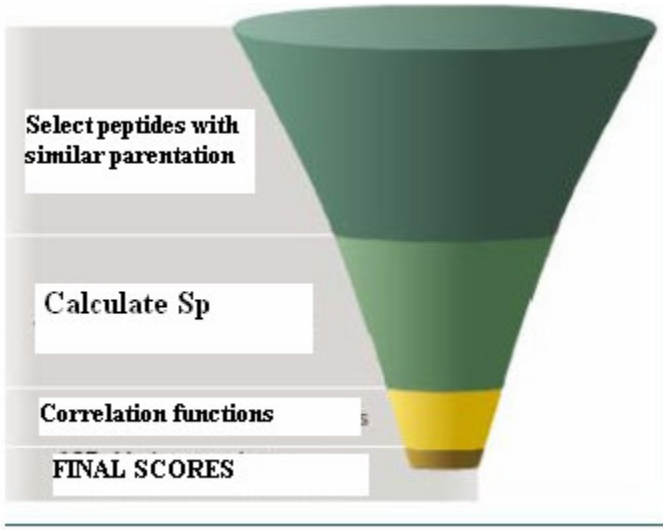

Fig. 3- A Simplified Flowchart for the Sequest Algorithm Showing the Process by which Sequest Provides Scores Used to Identify Peptides Flowchart shows process as described in United States patent 6,017,693 [13]. Note that information from the mass spectrum is used three times: (1) as a filter to select only peptides from the database sharing a similar parent ion mass with the unknown peptide; (2) during a preliminary Sp ‘‘closeness-of-fit’’ filter to select the top 500 peptide candidates; and (3) through a correlation function to produce the final scores.

Table 1- List of PFF packages available in internet

PFF Package	**URL**
Sequest	http://fields.scripps.edu/sequest/index.html
Popitam	http://expasy.org/tools/popitam
Mascot	http://www.matrixscience.com/search_form_select.html
Sonar	http://bioinformatics.genomicsolutions.com/ProteinId.html
Protein Prospector	http://prospector.ucsf.edu
TANDEM	http://prowl.rockefeller.edu/tandem/thegpm_tandemhtml
Phenyx	http://www.phenyx.ms.com
Spectrum Mill	http://www.chem.agilent.com/scripts/pds.asp?lpage=7771

Table 2 -Popular packages for De novo sequencing of MS data using Tag based approach

Tag-Based Package	**URL**
Gnten Tag	http://fields.scripps.edu/Guten Tag/index.html
Inspect	http://peptide.ucsd.edu/inspect.html
Lutefisk	http://www.hairyfatguy.com/lutefisk
PEAKS	http://bioinformaticssolutions.com:8080/peaksonline
MS BLAST	http://dove.embl-heidelberg.de/Blast2/msblast.html
FASTA	http://www.ebi.ac.uk/fasta33