Bioinformatics and Biosystems 2006, Vol. 1, No. 2, pp. 89-94

Informatics for protein identification by tandem mass spectrometry; Focused on two most-widely applied algorithms, Mascot and SEQUEST

Chang Ho Sohn[2], Jin Woo Jung[1], Gum Yong Kang[1] , Kwang Pyo Kim[1]

1Department of Molecular Biotechnology, Institute of Biomedical Science and Technology, Bio/Molecular Informatics Center, Konkuk University, Seoul 143-701, Korea
2 Department of Chemistry, Seoul National University, Seoul 151-742, Korea

Abstract

Mass spectrometry (MS) is widely applied for high throughput proteomics analysis. When large-scale proteome analysis experiments are performed, it generates massive amount of data. To search these proteomics data against protein databases, fully automated database search algorithms, such as Mascot and SEQUEST are routinely employed. At present, it is critical to reduce false positives and false negatives during such analysis. In this review we have focused on aspects of automated protein identification using tandem mass spectrometry (MS/MS) spectra and validation of the protein identifications of two most common automated protein identification algorithms Mascot and SEQUEST.

Keywords: database searching, protein identification, proteomics, mass spectrometry

Introduction

Within the past decade, understanding on proteins and their cellular functions has boosted at an exponential rate (Aebersold and Mann, 2003). This is due to revolutionary emergence of proteomics which in general deals with the large scale determination of cellular function of genes directly at the protein level. The field of proteomics is composed of various analytical techniques including cell imaging (Heazlewood et al., 2005; Kleiner et al., 2005; Schubert, 2006), array (Lopez and Pluskal, 2003; Angenendt, 2005 Clarke and Chan, 2005; Hudelist et al., 2005; Speer et al., 2005), yeast two-hybrid assay (Cho et al., 2004; Colland and Daviet, 2004; Ramachandran et al., 2005) and phage display (Hallborn and Carlsson, 2002; Hartley, 2002; Lopez and Pluskal, 2003; Sidhu et al., 2003). Another powerful proteomic approach is based on mass spectrometer which can analyze proteins or protein population isolated from cells or tissue to collect amino acid sequence information of proteins. Remarkable developments in various fields of related technology such as soft ionization methods of protein, high performance hybrid analyzers, ion activation techniques like electron capture dissociation (Zubarev et al., 1998) or recently reported electron transfer dissociation (Syka et al., 2004), powerful analysis softwares and sensitive separations, all these techniques have lead to an appearance of high-throughput methodology. Now, one allows identifying thousand of proteins at very short time compared with traditional biochemical analysis. So it is not surprise that whole proteome analysis is truly possible. Most notably, the soft ionization techniques for proteins were acknowledged by the 2002 Nobel prize in chemistry (Aebersold and Mann, 2003). Such studies, which adopt mass spectrometry to analyze proteomes, typically pose enormous challenges owing to the high degree of complexity and high range of dynamic range of cellular proteomes and low abundance of many proteins which regulate important cellular functions. Mass spectrometry (MS)-based proteomics is a discipline that is largely dependant on the availability of gene and genome sequence databases. Already, MS-based proteomics has established itself as an essential technology to interpret the information encoded in genomes. So far, protein analysis by MS has been most successful when applied to small sets of proteins isolated in specific functional contexts. The systematic analysis of the much larger number

Corresponding Author : Kwang Pyo Kim (E-mail : kpkim @konkuk.ac.kr)
This work was supported by a grant from the National R&D Program for Cancer Control, Ministry of Health & Welfare, Republic of Korea (0520070-1) and Kore Research Foundation (KRF 2004-F00019).
Abbreviations used: MS/MS, tandem mass spectrometry; TOF, time-of-flight instrument; MS, mass spectrometry;

of proteins expressed in a cell is now also rapidly advancing due to the achievements of experimental and computational advances.

Typically, tandem mass spectra derived from proteolytic peptides of complex protein samples are processed by a certain computer software designed to identify various peptides by returning the best match to the experimental data from a given protein database (Enget al., 1994; Perkins et al., 1999). At this stage, it is very critical to distinguish between true positive and false positive identifications (Keller et al., 2002; Peng et al., 2003; Sadygov and Yates, 2003). Although it has been suggested that manual validation of each spectrum could eliminate these spurious hits, it is almost impossible to validate all proteome data set with manual confirmation due to their enormous sizes. It is therefore required for most of the proteomic researchers to design experiments that maximize confident protein identifications using available instrumentation and computation resources. Even though manual validation could not be used to eliminate all of false positives, the use of reverse database searching seems more robust than manual validation and gives more statistically significant results when large databases are considered.

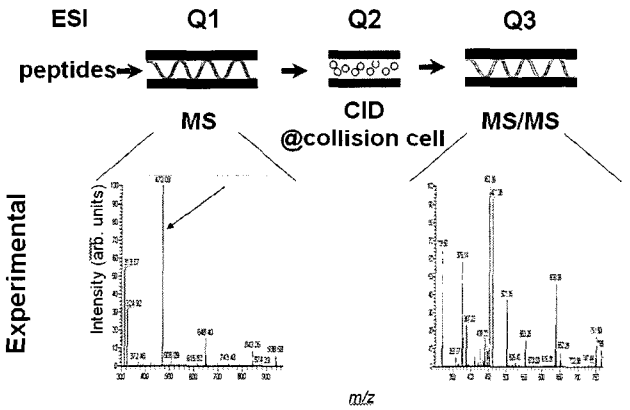
In this review, we present and compare two widely used MS/MS spectra interpretation algorithms, Mascot (Perkins et al., 1999)and SEQUEST (Eng et al., 1994). Traditionally, MS/MS spectra acquired with ion-trap mass spectrometer are interpreted with SEQUEST, whereas Mascot is used to sequence spectra obtained with time-of-flight (TOF) type mass spectra. Based on recent studies (Keller et al., 2002; Elias et al., 2005), it would appear that the two algorithms yielded similar results at least for ion-trap acquired spectra.

These two algorithms apply basically similar approaches to assign experimentally obtained MS/MS spectra to theoretically predicted peptides in a sequence database by comparing the mass to charge ratio of fragment ions generated by experimental and theoretical procedures (Sadygov et al., 2004). However, Mascot and SEQUEST use fundamentally different principles in their mathematical scoring operations. Generally, Mascot integrates all of the following information such as molecular weight of peptides, MS/MS spectra and combination of mass data acquired from a protein database. Its scoring algorithm is probability based metric to assess the likelihood of a fragmented peptide to an acquired spectrum.

MS/MS DATABASE MATCHING

Peptide fragmentation

Peptide fragmentation mass spectra, which are generated through sequential MS/MS, contain rich information on the sequence of the peptide. The information on the sequence of each peptide enables the identification of a protein form a single peptide. The use of MS/MS spectra is also the method of choice for identifying complex protein mixtures, unlike peptide mass fingerprinting methods which match the experimentally observed masses of peptides to those calculated from the sequence databases. MS/MS spectra are generated by automatic selection of the MS ion signals and fragmentation of each selected precursor ion, so-called "data-dependent" analysis through MS/MS (Blyn et al., 1996). The general procedure (Figure 1) is illustrated. The peptides are separated by HPLC and eluted into an electrospray ion source. After ionization charged peptides enter the mass spectrometer and a mass spectrum of the peptides ionized at that moment is obtained. The selected peptides by a "data-dependent" analysis are further fragmented at collision cell to record MS/MS spectra. The MS/MS spectra obtained are stored, reduced and extracted from the binary files to perform database search. The extracted MS/MS spectrum firstly is filtered through where candidate peptides in the database are identified if the peptide molecular weight is within preset mass tolerance. Given such a list of candidates, the experimental MS/MS spectrum is compared to a virtual MS/MS spectrum generated for each candidate peptide by calculation. Various scoring methods are then used to validate each match. Each match is scored that reflect how well each virtual MS/MS spectrum matched with the experimental spectrum.



MASCOT

Mascot is one of the most prevail searching programs among MS-based protein identification. It is based on the MOWSE algorithm, which uses average properties of the proteins in the database, but in addition it uses probability-based scoring algorithm for fragment information. 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

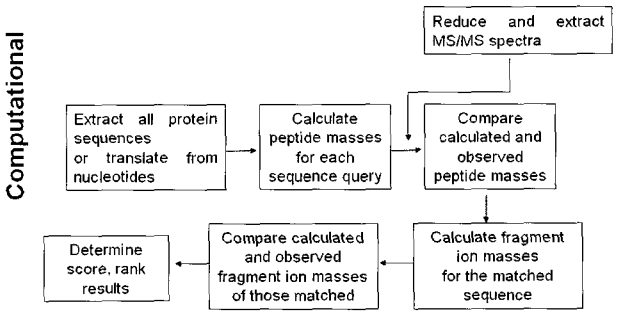


Figure 1. Schematic of the approach used by the computer algorithm to match MS/MS spectra of peptides to sequences in the proteins database.

then ranked with decreasing probability of being a random match to the experimental data. The scoring algorithm is probability based, which gives several advantages over other programs. Most of all, the guidance to determine the possibility of wrong identification is quite straightforward in comparison with SEQUEST. In addition, Mascot supports three different types of search: peptide mass fingerprint, sequence query and MS/MS ions search. However, its reliability of results relatively relies on the size of the database unlike SEQUEST. For successful searching through Mascot, some parameters should be treated very carefully. For example, a mass error window, either peptide mass fingerprinting or MS/MS ions search, is a particularly important parameter. This has directly influence on the final score and statistically significance. At the same time, the specified database set gives the better score than the whole database set searching. Any combination with sequence query search, Mascot can improve the result by using amino acid composition and its partial or incomplete sequence information acquired from either biochemical analysis or de novosequencing of MS/MS spectrum. A more extensive support beyond other searching program about post-translation modification (PTM) of peptides is an obviously big advantage to proteomic researchers. Yet the exact identification of such a labile PTM like phosphorylation is still on its way to improvement due to its complexity in MS/MS spectrum contaminated by neutral loss of phosphoric acid (-98Da) or phosphate (-80Da) and computational load to consider modifications during MS/MS ions search, especially in the low energy collision-induced dissociation (CID) MS/MS.

SEQUEST:

SEQUEST uses solely un-interpreted peptide tandem mass spectra to identify a certain protein. The most distinctive feature of SEQUEST compared with Mascot is cross-correlation scoring algorithm. Its statistical meaning could be somewhat unfamiliar to ordinary proteomic researchers, however, its performance is worth compensating its users for the difficulty of understanding. While Mascot scores each MS/MS spectrum by random matching algorithm, SEQUEST evaluates its score by the discrete cross-correlation analysis through fast Fourier transformation. The correlation function measures the coherence of two signals which are the experimental MS/MS spectrum and the theoretically predicted MS/MS spectrum. That is to say, the cross-correlation scores how close the experimental spectrum fits to the ideal spectrum. By changing the displacement value through the spectrum, the cross-correlation sums overlap of the two signals at each different displacement value. If the two different signals have no offset, the cross-correlation function should maximize at the displacement value = 0. The final score assigned to each predicted peptide sequence is the value of the function when the displacement value = 0 minus the mean of the cross-correlation function over the range of the displacement value from -75 to 75. The scores are normalized to 1.0 and termed Cn or Xcorr. Unlike the score from random event probability of matching, normalized cross-correlation scores (Xcorr) are comparative with different database set. Namely, it is possible to conduct a direct comparison among several searches based on each different database set. With the absolute value of Xcorr, normalized difference (ΔCn) between that of first-ranked and second-ranked peptide is a very useful criterion to distinguish false positive identification among specious hits. Several published criteria for both Xcorr and ΔCn are well summarized at else where (Elias et al., 2004). At recent, multi-enzymatic digest is preceded before interpretation of obtained MS/MS spectra using SEQUEST, which improves the coverage of protein identification from a complex mixture.

Quality of search results

The software tools for protein identification using mass spectrometric information will give a top-ranking candidate even if all the matching peptides are random matched. It is important to determine what the probability is that the identified protein is a false positive. Using a composite reversed database search strategy (Moore et al., 2002; Peng et al., 2003; Elias et al., 2004), it is possible to estimate the false positive rates of applied score filter criteria, allowing comparisons of multiple data sets with similar false positive rates. The basis of false positives is not due to any single database search algorithm or instrument type. This is a problem related to the fundamental aspects of algorithms that handle enormous amount of data set and instruments that analyze proteome samples composed of high dynamic range of component proteins. This method has been expanded to use alternative information, such as isoelectric point and molecular weight, to remove more false positives while providing greater sensitivity during filtering (Cargile et al., 2004a; Cargile et al., 2004b).

Most probability-based calculations do not take into account enough complex biological factors, such as divergent evolution and organism-specific codon bias, to provide a real measure of the validity of data. Also there are several numerical problems with these algorithms, which include the fact that several combinations of amino acids are indistinguishable (i.e., glycine-glycine and aspartate at nominal mass of 114, valine-serine and tryptophan at nominal mass of 186, and isoleucine and leucine at nominal mass of 113) by mass only. These factors are presumably the reasons of the fact that such algorithms return a number of virtual proteins from the reversed database at a high confidence search criteria.

Researchers often restrict their confident protein identifications to those identified by two or more peptides (Cargile and Stephenson, 2004), as proteins identified by single peptides exhibit higher false positive rates. However, when include removal of this peptide class as another filtration step, it decreased one third of the number of identified proteins, 85-95% of which were estimated to be correct (Elias et al., 2005).

The other important issue that should be considered in any protein identification algorithms is the rate of false negatives that refer the number of correct peptides missed identification by the search threshold. The scores of the missed peptides fall below the limit of the threshold to remove an acceptable degree of false positives from a given results. It is very critical to balance the search threshold between the reduction the amount of false positives and the multiplication of the number of false negatives.

Both SEQUEST and Mascot do not use fragment ion intensity-based scoring algorithm, which only use the information for pre peak selection. Even though the intensity of fragment ion is quite informative due to its distinctive features in most of the MS/MS spectra, it has not been used explicitly way. Recent report by Elias and coworkers (Elias et al., 2004) devised a new algorithm which evaluates a score of MS/MS spectrum based on the intensities of fragment ions using the learning algorithm of MS/MS spectra. This newly invented algorithm shows outranked performance beyond that of either SEQUEST or Mascot with well-known and most recently reported criteria especially in case of combination of the intensity-based algorithm and traditional identification scoring system. To use this scheme with confidence, further search and validation of this method should be done in the near future.

There are also very interesting issues concerned with comparative research between different types of scoring algorithm. Many of biological research have been used Mascot or SEQUEST routinely without much consideration of their unique scoring. Like many users'experiences, coverage of peptides or proteins identification is not quite comprehensive between the two mostly used programs. First wide and extensive comparative evaluation between different algorithms was done by Resing and coworkers (Resing et al., 2004) and showed at least Mascot and SEQUEST are mutually comparable for ion trap-acquired MS/MS spectra. Elias and coworkers (Elias et al., 2005) recently reported comparative evaluation of instrument platforms, scoring algorithm and multiple stage analysis of samples. The comments from their paper is quite similar with previous study, however, more detailed and contingent evaluations are included. They draw three primary conclusions: (i) Mascot and SEQUEST results fairly overlapped (>85%) for LTQ-acquired MS/MS spectra with filtering criteria, but Mascot gave a birth better match with QSTAR MS/MS spectra which is comparable with traditional usage of Mascot program; (ii) multiple times can be made more confident results better than that of single time.; (iii) peptide or protein identification is complimentary between the two different instruments.

Researchers might consider using complementary analytical platform to increase peptide and protein identification. This, however, may be insufficient if platform reproducibility is low. According to recent publications (Durr et al., 2004; Elias et al., 2005), when samples are analyzed multiple times and are measured by complementary instruments, both protein and proteome coverage were increased dramatically.

Conclusion

In the past decade, applications of mass spectrometry to analysis of proteins have brought revolutionary advancement from analysis of simple purified proteins to the whole mixture of proteins and in understanding cellular roles of proteins. Furthermore, it opens totally new biological field, systems biology. It mainly depends on the improvement of mass spectrometry instruments, separation techniques and algorithms to interpret enormous MS/MS spectra. There are a number of algorithms available for performing database searches using MS/MS spectra to identify peptide sequences and proteins.

While database searches are straightforwardly executed and they successfully capture as much identification from input data set as possible, there are currently no golden rules with respect to the validation of protein identifications. Therefore, we may look forward to even more powerful algorithms to automatically validate search results and yield more rich information. Future work has to focus on improving scoring schemes even more to reduce the number of false positive and false negative identification simultaneously and unifying MS/MS and database format to advance the development of global identification system.

References

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

[2] Angenendt, P. (2005). Progress in protein and antibody microarray technology. Drug Discov Today 10, 503-511. Blyn, L.B., Swiderek, K.M., Richards, O., Stahl, D.C., Semler, B.L., and Ehrenfeld, E. (1996). Poly(rC) binding protein 2 binds to stem-loop IV of the poliovirus RNA 5' noncoding region: identification by automated liquid chromatography-tandem mass spectrometry. Proc Natl Acad Sci U S A 93, 11115-11120.

[3] Cargile, B.J., Bundy, J.L., Freeman, T.W., and Stephenson, J.L., Jr. (2004a). Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification. J Proteome Res 3, 112-119.

[4] Cargile, B.J., and Stephenson, J.L., Jr. (2004). An alternative to tandem mass spectrometry: isoelectric point and accurate mass for the identification of peptides. Anal Chem 76, 267-275.

[5] Cargile, B.J., Talley, D.L., and Stephenson, J.L., Jr. (2004b). Immobilized pH gradients as a first dimension in shotgun proteomics and analysis of the accuracy of pI predictability of peptides. Electrophoresis 25, 936-945.

[6] Cho, S., Park, S.G., Lee do, H., and Park, B.C. (2004). Protein-protein interaction networks: from interactions to networks. J Biochem Mol Biol 37, 45-52.

[7] Clarke, W., and Chan, D.W. (2005). ProteinChips: the essential tools for proteomic biomarker discovery and future clinical diagnostics. Clin Chem Lab Med 43, 1279-1280.

[8] Colland, F., and Daviet, L. (2004). Integrating a functional proteomic approach into the target discovery process. Biochimie 86, 625-632.

[9] Durr, E., Yu, J., Krasinska, K.M., Carver, L.A., Yates, J.R., Testa, J.E., Oh, P., and Schnitzer, J.E. (2004). Direct proteomic mapping of the lung microvascular endothelial cell surface in vivo and in cell culture. Nat Biotechnol 22, 985-992.

[10] Elias, J.E., Gibbons, F.D., King, O.D., Roth, F.P., and Gygi, S.P. (2004). Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nat Biotechnol 22, 214-219.

[11] Elias, J.E., Haas, W., Faherty, B.K., and Gygi, S.P. (2005). Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. Nat Methods 2, 667-675.

[12] Eng, J.K., McCormack, A.L., and Yates, J.R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry 5, 976-989.

[13] Hallborn, J., and Carlsson, R. (2002). Automated screening procedure for high-throughput generation of antibody fragments. Biotechniques Suppl, 30-37.

[14] Hartley, O. (2002). The use of phage display in the study of receptors and their ligands. J Recept Signal Transduct Res 22, 373-392.

[15] Heazlewood, J.L., Tonti-Filippini, J., Verboom, R.E., and Millar, A.H. (2005). Combining experimental and predicted datasets for determination of the subcellular location of proteins in Arabidopsis. Plant Physiol 139, 598-609.

[16] Hudelist, G., Singer, C.F., Kubista, E., and Czerwenka, K. (2005). Use of high-throughput arrays for profiling differentially expressed proteins in normal and malignant tissues. Anticancer Drugs 16, 683-689.

[17] Keller, A., Nesvizhskii, A.I., Kolker, E., and Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and

database search. Anal Chem 74, 5383-5392.

[18] Kleiner, O., Price, D.A., Ossetrova, N., Osetrov, S., Volkovitsky, P., Drukier, A.K., and Godovac-Zimmermann, J. (2005). Ultra-high sensitivity multi-photon detection imaging in proteomics analyses. Proteomics 5, 2322-2330.

[19] Lopez, M.F., and Pluskal, M.G. (2003). Protein micro- and macroarrays: digitizing the proteome. J Chromatogr B Analyt Technol Biomed Life Sci 787, 19-27.

[20] Moore, R.E., Young, M.K., and Lee, T.D. (2002). Qscore: an algorithm for evaluating SEQUEST database search results. J Am Soc Mass Spectrom 13, 378-386.

[21] Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J., and Gygi, S.P. (2003). Evaluation of multi-dimensionalchromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J Proteome Res 2, 43-50.

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

[23] Ramachandran, N., Larson, D.N., Stark, P.R., Hainsworth, E., and LaBaer, J. (2005). Emerging tools for real-time label-free detection of interactions on functional protein microarrays. Febs J 272, 5412-5425.

[24] Resing, K.A., Meyer-Arendt, K., Mendoza, A.M., Aveline-Wolf, L.D., Jonscher, K.R., Pierce, K.G., Old, W.M., Cheung, H.T., Russell, S., Wattawa, J.L., Goehle, G.R., Knight, R.D., and Ahn, N.G. (2004). Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. Anal Chem 76, 3556-3568.

[25] Sadygov, R.G., Cociorva, D., and Yates, J.R., 3rd. (2004). Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. Nat Methods 1, 195-202.

[26] Sadygov, R.G., and Yates, J.R., 3rd. (2003). A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. Anal Chem 75, 3792-3798.

[27] Schubert, W. (2006). Exploring molecular networks directly in the cell. Cytometry A 69, 109-112.

[28] Sidhu, S.S., Fairbrother, W.J., and Deshayes, K. (2003). Exploring protein-protein interactions with phage display. Chembiochem 4, 14-25.

[29] Speer, R., Wulfkuhle, J.D., Liotta, L.A., and Petricoin, E.F., 3rd. (2005). Reverse-phase protein microarrays for tissue-based analysis. Curr Opin Mol Ther 7, 240-245.

[30] Syka, J.E., Coon, J.J., Schroeder, M.J., Shabanowitz, J., and Hunt, D.F. (2004). Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc Natl Acad Sci U S A 101, 9528-9533.

[31] Zubarev, R. A., Kelleher, N. L., and McLafferty, F.W. (1998). Electron capture dissociation of multiply charged protein cations. A nonergodic process. J Am Chem Soc 120, 3265-3266.