Full Report

ibc Interdisciplinary Bio Central

Open Access, Open Review Journal

Reinterpretation of the protein identification process for proteomics data

Kyung-Hoon Kwon[1], Sang Kwang Lee[1], Kun Cho[1], Gun Wook Park[1], Byeong Soo Kang[2], and Young Mok Park[1,3,*]

[1]Division of Mass Spectrometry Research, Korea Basic Science Institute, Ochang, Chungbuk, Republic of Korea
[2]The I-BIO graduate program and National Core Research Center for Systems Bio-Dynamics, POSTECH, Pohang, Kyungbuk 790-784, Republic of Korea
[3]Graduate School of Analytical Science and Technology, Chungnam National University, Daejeon, Korea

Subject areas: Bioinformatics (Proteomics, protein identification, database search)

Author contribution: Kyung-Hoon Kwon, Data Analysis
Sang Kwang Lee, Cell preparation
Kun Cho, Mass Spectrometry
Gun Wook Park, Byung Soo Kang, Database search
Young Mok Park, Project Investigator

***Correspondence** and requests for material should be addressed to Y.M.P. (ympark@kbsi.re.kr).

Reviewer: Sang Yun Cho, Yonsei University, Republic of Korea. Je-Yoel Cho, Kyungpook University, Republic of Korea.

Editor: Keun Woo Lee, Gyeongsang National University, Republic of Korea

Received July 21, 2009;
Accepted July 28, 2009;
Published July 29, 2009

Citation: Kwon, K., et al. Reinterpretation of the protein identification process for proteomics data. IBC 2009, 1(3):9, 1-6. doi:10.4051/ibc.2009.3.0009

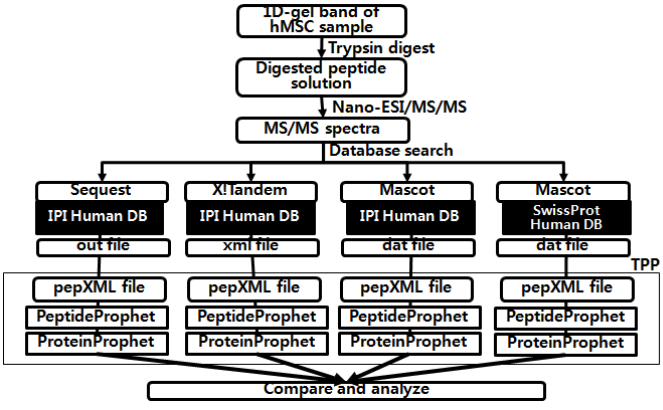
Supporting online materials:

Funding: Korea Science and Engineering Foundation grant (#2006-04605, Y.M.Park, #2006-04104, K.-H. Kwon).

Competing interests: All authors declare no financial or personal conflict that could inappropriately bias their experiments or writing.

SYNOPSIS

***Introduction*:** In the mass spectrometry-based proteomics, biological samples are analyzed to identify proteins by mass spectrometer and database search. Database search is the process to select the best matches to the experimental mass spectra among the amino acid sequence database and we identify the protein as the matched sequence. The match score is defined to find the matches from the database and declare the highest scored hit as the most probable protein. According to the score definition, search result varies. In this study, the difference among search results of different search engines or different databases was investigated, in order to suggest a better way to identify more proteins with higher reliability.
***Materials and Methods*:** The protein extract of human mesenchymal stem cell was separated by several bands by one-dimensional electrophorysis. One-dimensional gel was excised one by one, digested by trypsin and analyzed by a mass spectrometer, FT LTQ. The tandem mass (MS/MS) spectra of peptide ions were applied to the database search of X!Tandem, Mascot and Sequest search engines with IPI human database and SwissProt database. The search result was filtered by several threshold probability values of the Trans-Proteomic Pipeline (TPP) of the Institute for Systems Biology. The analysis of the output which was generated from TPP was performed.
***Results and Discussion*:** For each MS/MS spectrum, the peptide sequences which were identified from different conditions such as search engines, threshold probability, and sequence database were compared. The main difference of peptide identification at high threshold probability was caused by not the difference of sequence database but the difference of the score. As the threshold probability decreases, the missed peptides appeared. Conversely, in the extremely high threshold level, we missed many true assignments.
***Conclusion and Prospects*:** The different identification result of the search engines was mainly caused by the different scoring algorithms. Usually in proteomics high-scored peptides are selected and low-scored peptides are discarded. Many of them are true negatives. By integrating the search results from different parameter and different search engines, the protein identification process can be improved.



Keywords: proteomics, database search, mass spectrometry, probability, trans-proteomic pipeline, protein identification

Introduction

In the beginning of 1990s the data analysis protocols using sequence database started to be published for the proteomics research through tandem mass (MS/MS) spectrometry. (Eng, et al., 1994). The MS/MS spectrum of peptide ion enables us to find amino acid sequence of the peptide. By aligning the experimental MS/MS peaks to the expected MS/MS ion peaks of possible peptides from protein sequences in the database, the best matched peptide is assigned to each MS/MS spectrum. There are some *de novo* sequencing methods (Dancik, et al., 1999) which does not mine peptide sequence from the sequence database but compute peptide sequence from the mass difference of peaks. However, the database search has been more convenient and common solution for the high-throughput proteomics.

Nowadays many analysis tools are available for the protein identification, characterization and quantitation. For the database search, there are several softwares such as Mascot (Perkins, *et al*., 1999), Sequest (Eng, *et al*., 1994), X!Tandem (Craig, et al., 2004), and OMSSA (Geer, et al., 2004). Because they adopted different scoring methods, their search results are not the same with each other. (Kapp, et al., 2005). It means the protein list that was obtained from database search depends on the search algorithm. This fact made biologists confused. Their question to be answered was which proteins were included in their samples. The database search method should have improved to supply more reliable protein list.

The data analysis methods such as the probability-based scoring algorithm (Perkins, et al., 1999; Nesvizhskii, et al., 2003) and the false discovery rate (FDR) estimation by decoy approach (Elias, et al., 2004) could satisfy biologists partially, because they brought the protein list including the value of probability which represented how much their list was reliable. If we take the protein list of very low false positive rate, for instance, 1% FDR, we can confirm that the search result is true within error rate 1% and the protein may be included in the sample by the probability 0.99. However, this validation method is not enough to screen the full list of proteins detected by mass spectrometry. Still the database search engines have affected significantly on the protein list, although we let false positive rate down to very small percentage. Smaller false positive rate could reflect larger true negative rate. The scoring algorithm dependence of protein list might come from the large amount of true negatives. True negative is the protein which got low score and discarded, while it is the true protein which we caught up its signal by mass spectrometer.

Some research group suggested the meta score which combines the search scores of several search engines and developed the data analysis software such as Scaffold (Proteome Software, Portland), ProteinScape (Bruker Daltonics, Germany) where the meta score was defined. When considering meta score, the search result could be refined to get more accurate result. (Alves, et al., 2008).

In the shotgun proteomics, the sample of protein mixture was analyzed by mass spectrometer to get hundred thousands of MS/MS spectra. From the usual database search, about only 20% spectra succeed to identify peptide sequences. When we consider the post-translational modification, additional hits can be found. And it is reported that the use of additional database search engines can expand the hit rate. (http://www.matrixscience.com/pdf/2009WKSHP1.pdf).

In this study, we tried to compare the peptide sequences identified for one MS/MS spectrum by different search engines or by different threshold probability. At first, it was checked whether one

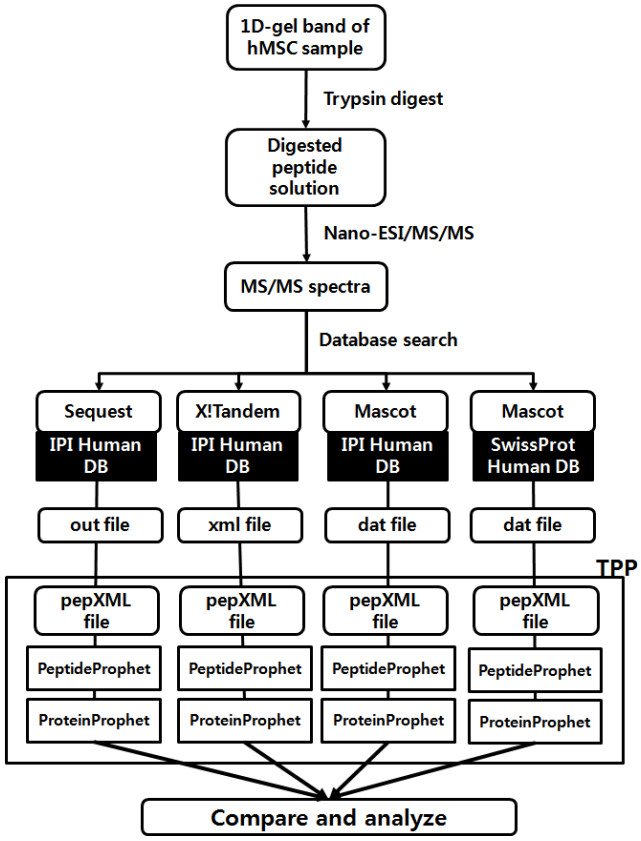


Figure 1. Analysis process of protein identification performance. For the different search engines of Sequest, X!Tandem and Mascot, for the different probability threshold of PeptideProphet of 0.99, 0.95, 0.90, 0.80, 0.50, and 0.20, the peptide / protein identification results were listed and compared.

MS/MS spectrum could be identified by different peptide sequences with low error rate in different search engines. Secondly, when two different search engines identified the same peptide sequence for one MS/MS spectrum, we compared the threshold probability from which the peptide sequence appears at each search engine.

Result and discussion

In this analysis, three major search engines of Mascot, Sequest and X!Tandem were used. As the sequence database, IPI human database v3.49 (EBI, UK) and Swiss-Prot database v51.6 (EBI, UK) were chosen. They are less redundant appropriately for the database search of proteomics experimental data than NCBI nr database. Especially, IPI database (Kersey, et al., 2004) is the standard database which was strongly recommended for the proteomics database search in the international collaboration projects of Human Proteome Organization. (Omenn et al., 2005) Swiss-Prot is a curated protein database keeping a minimal level of redundancy. (O'Donovan, et al., 2002).

We analyzed the database search result of Mascot, Sequest and X!Tandem with IPI database and Mascot search result with Swiss-Prot database. Their output files were converted to XML files and selected as input file of TPP pipeline including PeptideProphet and ProteinProphet. PeptideProphet computes the probability value for each peptide hit by neural network technology. ProteinProphet integrates these PeptideProphet result to assign probabilistic scores

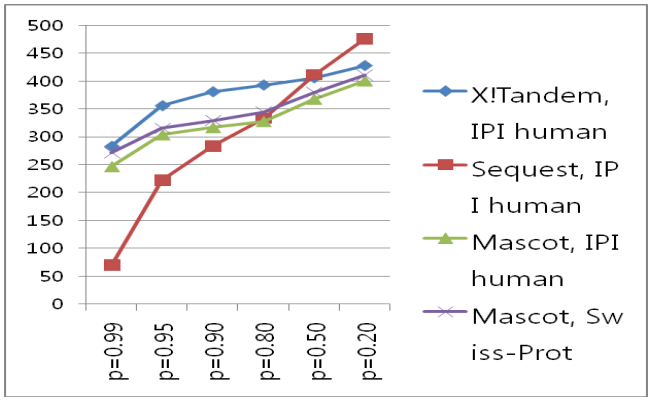


Figure 2. Number of peptides which were identified for the given threshold probability p and search engine, database denoted. X!Tandem identified more peptides than the others at threshold probability p>0.8. Comparing the Swiss-Prot database with IPI human database, Swiss-Prot found a little bit more peptides at the same threshold probability. Concerning with the search engine, Sequest identified the least peptides at higher probability threshold. On the other hand, Sequest identified the most peptides at lower probability threshold. The numbers of peptides were listed at Table 1.

to the identified proteins. The proteins which were obtained from ProteinProphet are grouped. If one protein share some peptides with another protein, then they are classified as a protein group.

Figure 1 shows our analysis procedure. For each database search, the probability values of 0.99, 0.95, 0.90, 0.80, 0.50 and 0.20 were assigned as the minimum values of PeptideProphet probability. The minimum probability 0.99 collects very reliable peptide sequences, while the probability 0.2 contains very many incorrect assignments. For each minimum probability, the identified peptides and proteins were listed in the supplementary materials. (Supplementary Table 1 and 2).

A dataset of 4487 MS/MS spectra from a one-dimensional gel band of human mesenchymal stem cell was performed the database search. Among these spectra, the number of identified

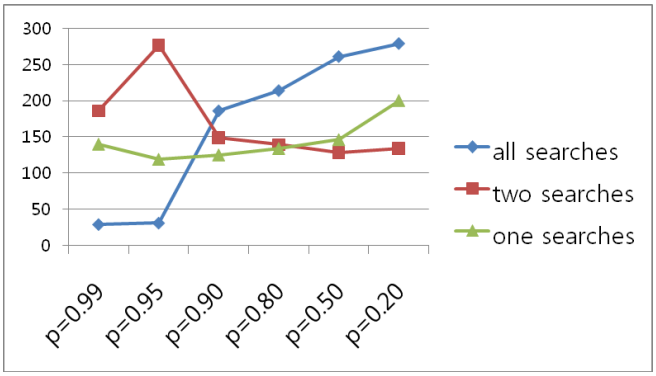


Figure 4. The number of peptides which were identified at only one search, two searches and all three searches, respectively. The number of peptides which were overlapped by three searches increased suddenly at p=0.90.

peptides of each database search are listed at Table 1 and drawn at Figure 2 and Figure 3. As the probability threshold decreased, new peptide sequences appeared by lower score. Some peptide sequences were scored low at one search engine, although they were scored high at another search engine. Some of these sequences might be true negatives at the former search. Figure 4 explains such tendency. It shows the three curves for the number of peptides which appeared at only one search, two searches and all three searches among X!Tandem, Sequest and Mascot, respectively. When the threshold probability decreases, the peptides of three search matches increased instead of the decrease of two search matches. When the probability became lower, the peptides which were obtained from two searches were identified at the other search engine and became the peptides which matched at all three searches.

Considering the difference among search engines, X!Tandem identified more peptides than the others, while Sequest identified much less at the higher threshold probability. And there were 34 peptide sequences which were found only at Sequest for p>0.9. Among these peptides unique at Sequest result, 15 peptides were identified with high confidence for p>0.99. Concerning with the low-

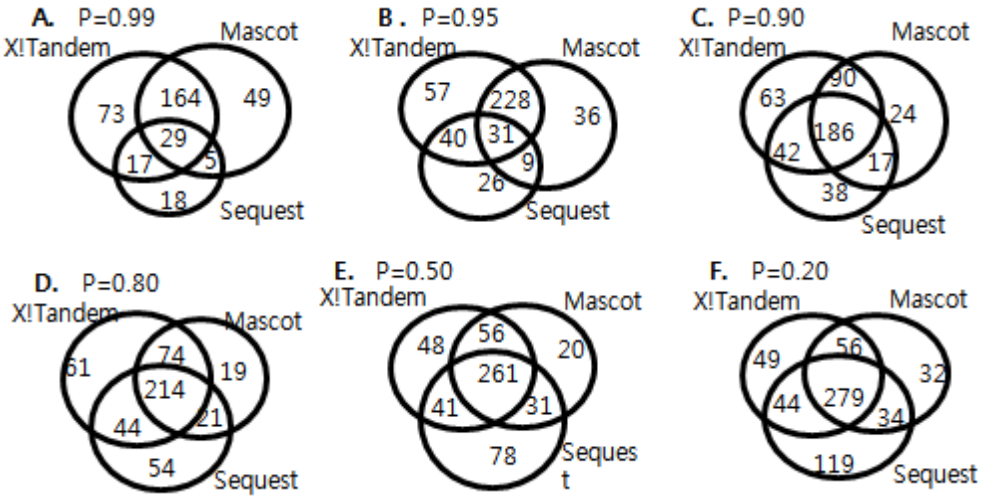


Figure 3. Venn diagram of the peptide sequence distribution for each threshold probability values. At p=0.99, only 29 peptides (8.2%) among 355 peptides were overlapped from three different database search of X!Tandem, Mascot and Sequest. At p=0.95, 31 (7.3 %) among 427 peptides were overlapped. At p=0.90, the number of overlapped peptides increased suddenly upto 186 peptides (40.4%) among 460 peptides, while only 33 more peptides were identified than at p=0.95. When the threshold probability value increased from 0.99 to 0.90, the peptide sequences which had been identified by only one search engine at higher probability appeared at other search engines as the threshold probability was lowered.

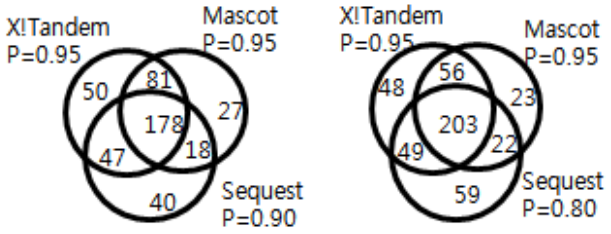
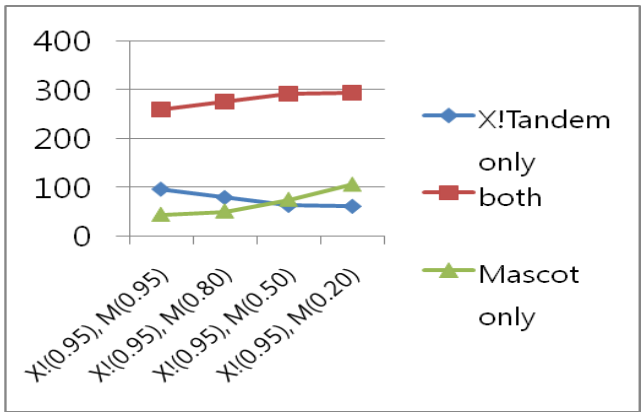


Figure 5. From the assumption that the PeptideProphet probability of Sequest is underestimated and compared the Sequest result of p=0.90, p=0.80 with those of the other search engines of p=0.95, throughput of Sequest at high threshold probability, probably PeptideProphet underestimated Sequest score and assigned the PeptideProphet probability lower. At Figure 5, we tried to compare Sequest p=0.90 result with X!Tandem and Mascot p=0.95 result. By slightly lowering Sequest threshold probability, we have got much more peptides which were also found at X!Tandem and Mascot by increasing only small amount of Sequest-unique peptides.

The difference of search result between IPI human and Swiss-Prot database was not serious. When we compare IPI human database search by Mascot with Swiss-Prot human database search by Mascot, only 24 peptides among 412 peptide identifications were uniquely identified at Swiss-Prot. Moreover, 14 peptides of them appeared only at low probability p=0.2. On the other hand, 14 peptides were uniquely identified at IPI human database. Among these peptides, 6 peptides were identified at p>0.90. It is noticeable that 24 peptides were found not at Mascot result of IPI human database but at Mascot result of Swiss-Prot database and among them 11 peptides were identified at X!Tandem and/or Sequest search of IPI human database with high score.

Figure 6 shows the protein group distribution for different threshold probability values. The identified proteins at TPP pipeline were grouped by shared peptides. Usually the proteins which belonged to one protein group were isoforms of similar sequences. When the threshold probability changed from 0.99 to 0.95, much more protein groups were identified. However, until p=0.80, there

A



B

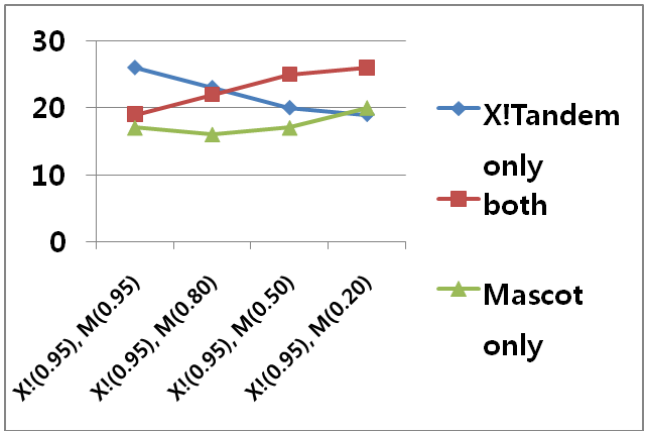


Figure 7. Comparison of protein identification result for X!Tandem result of fixed probability, p=0.95 and Mascot result of several probabilities such as p=0.95, p=0.80, p=0.50 and p=0.20. For lower probability of Mascot, the number of proteins unique at Mascot increases and the number of proteins overlapping at X!Tandem and Mascot also increases. Many proteins which were identified only at X!Tandem at p=0.95 were identified at Mascot at lower probability value. (**6A**) Distribution of the number of peptides (**6B**) Distribution of the number of protein groups.

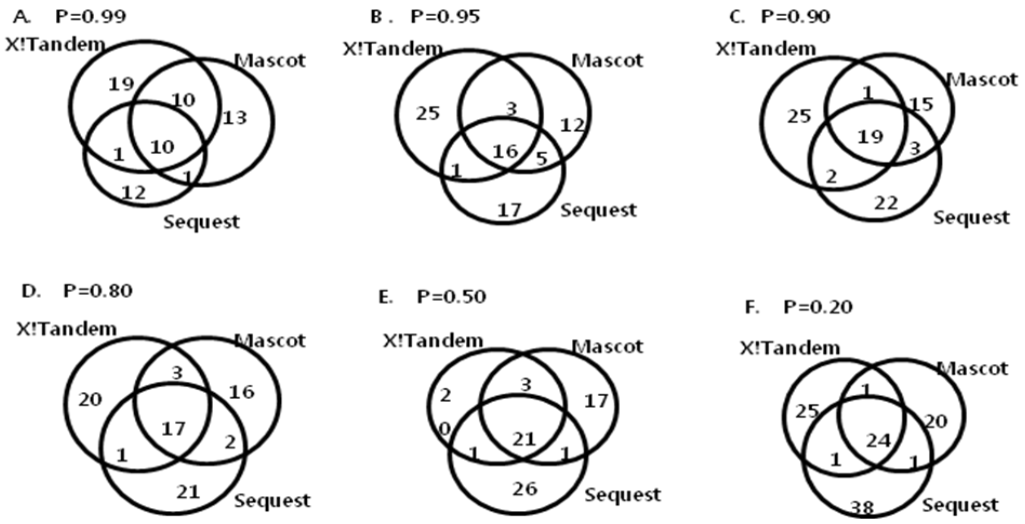


Figure 6. Protein group distribution for each threshold probability. Differently from the peptide distribution of Figure 3, the number of protein groups was not increased rapidly. From p=0.95 to p=0.80, the identified protein groups does not change much.

Table 1. The number of identified peptides for each probability threshold and each search engine.

Search engine	database	P=0.99	0.95	0.9	0.8	0.5	0.2
X!tandem	IPI human	283	356	381	393	406	428
Sequest	IPI human	69	222	283	333	411	476
Mascot	IPI human	247	304	317	328	368	401
Mascot	Swiss-Prot human	271	315	329	344	380	411

occurred no remarkable increase of protein groups. Such behavior is different from the case of identified peptide distribution of Figure 3. At Figure 3, p=0.90 was a point where distinct change occurred in the number of overlapped peptides. Considering the identified proteins not peptides, such abrupt change disappeared. It is expected that the overlapped peptides of p=0.90 at Figure 3 contributed to improve the identified peptides of proteins, not to identify new proteins, from Figure 3B, 3C, 6B and 6C.

At Table 2, the number of protein groups collected from TPP pipeline was written for each database search result. The single hit number denotes the number of proteins which were imported by a single peptide. Single hit is evaluated as less confident compared to the multiple hit. As previously mentioned, Table 1 showed that X!Tandem identified more peptides. However, Table 2 represents that X!Tandem has rather higher single hit ratio compared to the Mascot search result at high threshold probability. It means more proteins of X!Tandem are less confident.

Conclusion and Prospects

In this study, we compared the peptides and proteins which were identified from different search engines and filtered by different threshold probabilities. At first, we aimed to check whether two search engines identify different sequences for one MS/MS spectrum. In our case, several cases were discovered where different sequences were identified for one MS/MS spectrum. But in all these cases, one of two sequences was scored by very low value. Finally, we confirmed that one MS/MS spectrum was assigned to one peptide sequence independently of the search engines.

Secondly, we were interested in the threshold probability where the same sequence is identified at another search engine. From the protein and peptide distribution of several different probability levels, we found that many true assignments got low scores and treated as negative. It was observed that many low-scored hits of one search engine were the high-scored hits of another search engine. These hits can be estimated as true negative at the previous search engine. At Figure 7, the low-scored hits of Mascot search were compared with high-scored hits of X!Tandem in the peptide and protein group distribution. Until when the threshold probability of Mascot goes down to p=0.5, the overlapped hits of X!Tandem and Mascot increases. As hybridizing search result of X!Tandem of p=0.95 and Mascot of p=0.5, more proteins were attained and we could distinguish which hits are less confident. At p=0.8, the overlapped hits stopped the increase but the number of Mascot-only hits increased. These hits would be insignificant. At this analysis, p=0.5 seems to be the optimal low probability to compare.

This work was done only for the spectra data acquired from FT LTQ/MS/MS which is one of high-resolution mass spectrometers. Therefore, some of this analysis may be specific to this experiment. In spite of the specificity of the sample, we had analyzed several hundreds of peptide hits and detected a consistent tendency in peptide identification. We expect that this behavior would be common for the data of the shotgun proteomics using high-resolution mass spectrometer.

Materials and Methods

Sample Preparation

The mesenchymal stem cells were isolated from human bone marrow aspirate and cultured in DMEM containing 10% fetal bovine serum, 100 U penicillin, 100 mg/ml streptomycin (Invitrogen, Carlsbad, CA). After the sequential processes of centrifugation, sonification, and incubation, stablized membrane proteins were collected. The regular one-dimensional 12% SDS-PAGE electrophoresis was applied to separate proteins by molecular weight. The gel was stained with Coomassie Brilliant Blue R-250 and excised into 20 bands. We selected one of dark bands to

Table 2. The number of identified protein groups which was computed by ProteinProphet after filtering by PeptideProphet.

search engine		p=0.99	p=0.95	p=0.90	p=0.80	p=0.50	p=0.20
X!Tandem, IPI human	proteins	44	48	50	53	58	68
	single hits	16	17	18	20	25	35
	single hit ratio	0.36	0.35	0.36	0.38	0.43	0.51
Sequest, IPI human	proteins	27	42	49	58	73	88
	single hits	12	15	19	27	40	55
	single hit ratio	0.44	0.36	0.39	0.47	0.55	0.63
Mascot, IPI human	proteins	39	44	45	47	56	76
	single hits	10	13	13	16	25	44
	single hit ratio	0.26	0.3	0.29	0.34	0.45	0.58
Mascot, Swiss-Prot	proteins	42	45	47	49	59	78
	single hits	12	15	16	18	27	46
	single hit ratio	0.29	0.33	0.34	0.37	0.46	0.59

analyze the protein identification performance.

Mass Spectrometry

The gel band was digested into peptides by trypsin and analyzed by tandem mass (MS/MS) spectrometry. All MS/MS experiments for peptide identification were performed a Nano-LC/MS system consisting of a Surveyor HPLC system and a 7-tesla LTQ-FT mass spectrometer (Finnigan, San Jose) equipped with a nano-ESI source. Ten microliter of each sample with digested peptides was separated on a homemade microcapillary column of length 100mm packed with C_{18} in 75 μm silica tubing. The mass spectrometer was operated in the data-dependent mode to automatically switch between MS and MS/MS acquisition. Target ions selected for MS/MS were dynamically excluded for 60 seconds.

Data Analysis

For the database search, the IPI human database (IPI.HUMAN.v3.49, EBI, UK) and SwissProt database (SwissProt v.51.6, EBI, UK) were used. Three database search engine of X!Tandem TORNADO (GPMDB, Canada), Mascot v. 2.2 (MatrixScience, UK) and Sequest v.28 (Finnigan, San Jose) were used. The missed cleavage was allowed at most once. The variable modification of methionine oxidation and the fixed modification of carbamidomethyl cystein were assigned as search parameters. The peptide tolerance of 50 ppm and MS/MS tolerance of 1 Da were used. MS/MS search results were analyzed by Trans-Proteomic Pipeline (TPP) (Keller, et al., 2005) of Institute for Systems Biology. Within the TPP user interface, the identified peptides were filtered by PeptideProphet (Keller, et al., 2002) by the probability values, p=0.99, 0.95, 0.90, 0.80, 0.50, and 0.20. And then, the PeptideProphet result for each probability threshold value and each search engine was transferred to ProteinProphet (Nesvizhskii, et al., 2003) to identify proteins by integrating peptide sequences. The proteins were combined to make groups according to the peptide sequences shared with several proteins.

Acknowledgement

We thank Prof. Daehee Hwang at POSTECH for the fruitful discussions about the idea of this work. This work was supported by the KOSEF grant funded by the Korean government (MOST) (#2006-04104, Kyung-Hoon Kwon). This research was supported to YMP by the Ministry of Education, Science and Technology (2009-008146).

References

Alves, G., Wu, W.W., Wang, G., Shen, R.F. and Yu, Y.K. (2008) Enhancing peptide identification confidence by combining search methods. J. Proteome Res. 7(8), 3102-13.

Craig, R., Beavis, R.C. (2004) TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20, 1466-1467.

Dancik, V., Addona, T.A., Clauser, K.R., Vath, J.E. and Pevzner, P.A. (1999) De Novo Peptide Sequencing via Tandem Mass Spectrometry. J. Comp. Biol. 6, 327-342.

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. Biotech. 22, 214-219.

Eng, J.K., McCormack, A.L., Yates, JR III (1994) An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. J. Am. Soc. Mass Spectrom 5, 976-989.

Eng, J.K., Fischer, B., Grossmann, J. and MacCoss, M.J. (2008) A Fast SEQUEST Cross Correlation Algorithm. J. Proteome Res. 7, 4598-4602.

Geer, L.Y., Markey, S.P., Kowalak, J.A., Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W. and Bryant, S.H. (2004) Open mass spectrometry search algorithm, J. Proteome Res. 3(5), 958-64.

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.

Keller ,A., Eng, J., Zhang, N., Li, X. and Aebersold, R. (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol. Sys. Biol. 2, 1-8.

Nesvizhskii, A.I., Keller, A., Kolker, E. and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal. Chem. 75, 4646-4658.

Kapp, E.A., Schutz, F., Connolly, L.M., Chakel, J.A., Meza, J.E., Miller, C.A., Fenyo, D., Eng, J.K., Adkins, J.N., and Omenn, G.S. (2005) An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. Proteomics 5, 3475-90.

Kersey, P.J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E. and Apweiler, R. (2004) The International Protein Index : an integrated database for proteomics experiments, Proteomics, 4(7), 1985-8.

O'Donovan, C., Martin, M.J., Gattiker, A., Gastelger, E., Bairoch, A. and Apweiler, R. (2002) High-quality protein knowledge resource: SWISS-PROT and TrEMBL. Brief Bioinform. 3(3), 275-84.

Omenn, G.S., States, D.J., et al. (2005) Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database, Proteomics 5(13), 3226-45.

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