Human Proteome Data Analysis Protocol Obtained via the Bacterial Proteome Analysis

Kyung-Hoon Kwon[1], Gun Wook Park[2], Jin Young Kim[2], Jeong Hwa Lee[2], Seung Il Kim[2] and Jong Shin Yoo[1]

[1]*Mass Spectrometer Development Team, Korea Basic Science Institute, Daejeon, Korea*

[2]*Proteomics Team, Korea Basic Science Institute, Daejeon, Korea*

Email : jongshin@kbsi.re.kr

ABSTRACT: In the multidimensional protein identification technology of high-throughput proteomics, we use one-dimensional gel electrophoresis and after the separation by two-dimensional liquid chromatography, the sample is analyzed by tandem mass spectrometry. In this study, we have analyzed the *Pseudomonas Putida* KT2440 proteome. For the protein identification, the protein database was combined with its reversed sequence database. From the peptide selection whose error rate is less than 1%, the SEQUEST database search for the tandem mass spectral data identified 2,045 proteins. For each protein, we compared the molecular weight calibrated from 1D-gel band position with the theoretical molecular weight computed from the amino acid sequence, by defining a variable MW_{corr}. Since the bacterial proteome is simpler than human proteome considering the complexity and modifications, the proteome analysis result for the *Pseudomonas Putida* KT2440 could suggest a guideline to build the protocol to analyze human proteome data.

1 INTRODUCTION

Proteome analysis using tandem mass spectrometry is the technique to produce high-throughput protein data. Before measuring the molecular weight with tandem mass spectrometer, the sample is digested and fractionated to separate it into the fractions each of which contains only several peptides. Usually, tandem mass spectrum keeps the information on the fragment ions of one peptide. A tandem mass spectrum offers the most possible peptide sequence through the protein database search, whose theoretical mass spectrum peaks match well with the experimental peaks to win the highest score. For lower quality mass spectra, the peak match is poor and the match scores are low, while high quality spectra get the high scores. For the marginal match scores, we cannot estimate how well the identified peptide sequence is confident and it is difficult to distinguish a marginal match score from true or false peptide assignments. The database search softwares such as SEQUEST [1], Mascot [2] use different scoring algorithms and they suggest the threshold match score over which the search result seems to be true. Actually the threshold score depends on a variety of experimental environments and it should be estimated at each experiment, respectively. The decoy approach of Elias *et al.* [3] using reversed sequence database enables us to determine the dynamic threshold score to satisfy our error rate requirement.

When we identify proteins by the database search, the protein molecular weight can be another criterion to make a decision that the identification result is true. When we have 1D-gel or 2D-gel image, the protein marker positions are the reference points to calibrate the experimental molecular weight of proteins. But the poor reproducibility of gel image and many modifications of proteins become the obstacle to making the molecular weight information as one of the criterions to estimate true assignments. Here we have adopted a variable MW_{corr} [4] that represents the correlation between the experimental molecular weight interpolated from the protein marker positions of the gel image and the protein molecular weight computed from the amino acid sequence. For the *Pseudomonas putida* KT2440, the proteins that acquired high match scores were shown to have the MW_{corr} values around 1.

Pseudomonas putida KT2440 is one of the bacteria, that adjust itself very well to diverse environments. It is metabolically versatile. Especially, this bacterium has been attractive because of its biodegradability for the various aromatic compounds. [5] In addition to this biological worth, we have focused on the fact that the proteomes of such bacteria are simpler than human proteome in protein characteristics. By applying the decoy approach at the database search with tandem mass spectral data and computing MW_{corr}'s after the protein identification, we could find the MW_{corr} distribution of confident proteins. This result became a criterion to grouping human proteome according to MW_{corr} values. By comparing the MW_{corr} distributions for different samples, we could characterize their proteomes.

2 MATERIAL AND METHODS

2.1 Sample preparation

Pseudomonas putida KT2440 was purchased from ATCC (www.atcc.org). *P. putida* KT2440 was pre-cultured in 50mM potassium phosphate buffer (pH 6.25) containing 3.4 mM $MgSO_4$, 0.3 mM $FeSO_4$, 0.2 mM $CaCO_3$, 10 mM NH_4Cl and 10 mM sodium succinate (KT2440-S) and then transferred into same fresh media or 5 mM benzoate (KT2440-B) media. The cultured bacteria were harvested as soon as growth reached the late exponential phase and then were stored at -80°C until 2-DE analysis.

Harvested cells were suspended in 20 mM Tris-HCl buffer (pH 8.0) and disrupted by a French pressure cell (SLM AMINCO, Urbana, IL, USA) at 20,000 lb/in^2. The crude extracts were separated by centrifugation at 15,000 x g for 45 min. The supernatant (buffer-soluble fraction) was collected and used for SDS gel electrophoresis.

2.2 Tandem Mass Spectrometer Analysis

All 42 samples were analyzed using multidimensional protein identification technology.[6-9] The digested peptide mixtures from different 1D gel bands were loaded separately onto a micro capillary column packed with C18 and SCX cation exchange materials. All spectra were produced from LTQ/MS/MS experiment using a Thermo-Finnigan (U.S.A.) LTQ ion trap mass spectrometer.

2.3 Database Search

The database search for the tandem mass spectral data was performed by SEQUEST whose version is Turbo-SEQUEST v. 3.1 SR1 (ThermoFinnigan, U.S.A.). Keratin peptide assignments were eliminated in total peptide assignments. For the database search, the *P. putida* protein database (PPDB) was downloaded from the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/). The protein database includes 12,463 protein sequences. To assess the false assignment distribution, the amino acid sequences of PPDB were completely reversed to create a reversed sequence database (RSDB).[10] While PPDB consists of the sequences of the known bacterial ORFs, RSDB is a database of retro-proteins that don't exist in *Pseudomonas Putida* KT2440 and can be hardly synthesized. If a peptide sequence from RSDB is found by the database search, we conclude that this identification result will not be correct.

2.4 Molecular Weight Analysis

The 1DE analysis of *P. Putida* KT2440 was performed with the positions of eight standard protein markers. By referring to the marker positions, we measured the molecular weight range of each band. From a linear regression analysis using the marker protein band positions, we obtained approximately the relationship between the protein molecular weight and band position. In general, the electrophoretic mobility (R_f) of SDS-protein complexes in gels is proportional to the logarithmic value of the polypeptide molecular weight, MW_{exp} in the region except the end of the gel. The molecular weight of a polypeptide can be estimated by comparing its mobility with those of the standards. For the *P. Putida* KT2440 sample, the eight marker positions provided a formula for the electrophoretic mobility.

$$R_f = -0.40 \times \log MW_{exp} + 4.74$$

As a variable to measure differences between the experimental molecular weight MW_{exp} calibrated from the electrophoretic mobility and the calculated protein molecular weight MW_{cal} computed from the protein sequence, we have introduced a variable MW_{corr} of molecular weight correlation. It is defined as

$$MW_{corr} = \frac{\log MW_{exp}}{\log MW_{cal}}$$

2.5 False Positive Analysis

After SEQUEST, we get the .out files where peptide sequences and their match scores are listed. Figure 1 is the plot of the peptide search result that is the graph of X_{corr} values versus MW_{corr} values. X_{corr} is the match score of SEQUEST. At Figure 1(a) that is the score distribution of peptides found from the PPDB, the peptides that won high scores appeared near $MW_{corr} = 1$, differently from the Figure 1(b), the distribution from the RSDB.

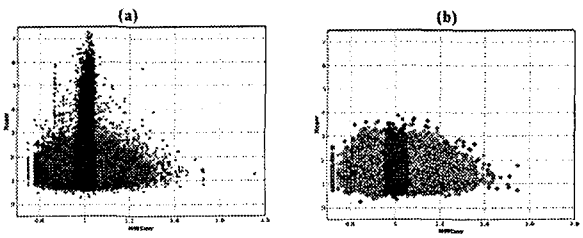

Figure 1 Match score distributions for PPDB and RSDB. (a) *Xcorr* vs. *MWcorr* for peptides found from PPDB (b) *Xcorr* vs. *MWcorr* for peptides found from RSDB. The red points are in the *MWcorr* region where high match scores are located at PPDB search result.

This figure allows us to assure two facts. One is that the lower X_{corr} scores in the peptide identification from PPDB distributes in the same shape as those from RSDB. The other is that the true assignments of Figure 1(a) are concentrated on the region of high X_{corr} scores around $MW_{corr} = 1$ and the distribution of RSDB is the same with the false assignment distribution of PPDB, which is the point mentioned at Elias et al.[3]. Therefore we can regard the peptide distribution of RSDB as the false assignment distribution of PPDB. Then, the true assignment distribution of PPDB is computed by subtracting the false assignments from the whole peptide distribution of PPDB.

After extracting the true and false distributions, we could compute the sensitivity $S(F)$ and the reliability $R(F)$ for the search result, which are defined as

$$S(F) = \int_F^{\infty} dF' P_{true}(F') / \int_{-\infty}^{\infty} dF' P_{true}(F')$$

$$R(F) = 1 - [\int_F^{\infty} dF' P_{false}(F') / \int_F^{\infty} dF' P_{true}(F')]$$

Here, $P_{true}(F)$ is the peptide distribution function of true assignment for the match score F, and $P_{false}(F)$ is the function of the false assignment.

By selecting the threshold score as the value whose reliability is 99%, we could get a set of confident peptides whose reliability is higher than 99%.

2.6 Proteome characterization

The highly confident peptide list filtered by the reversed sequence database was used to identify proteins by the program DTASelect. When we plot the MW_{corr} distribution, we can see how many proteins are found out of the region where the experimental molecular weight is similar to the theoretical molecular weight. Since bacterial proteome is much simpler than human proteome, we expected that most

of the proteins would belong to MW_{corr} ~1 region and the range where the most proteins are located can be interpreted as the MW_{corr} error range, that is, the experimental molecular weight error range. Because we have calibrated the molecular weight by using the marker protein positions in 1D-gel bands, there should be the difference from the exact molecular weight. The MW_{corr} distribution of this bacteria would suggest the error range for MW_{exp} . With such a proteome analysis method, we could find the MW_{corr} where rather simple proteins belong. By applying this information to the human proteome containing many variations in protein status, we have characterized human proteome.

3 RESULTS

From the X_{corr} distribution, we got the true assignment distribution for PPDB. Figure 2(a) shows the peptide number distribution identified for each database. Figure 2(b) is the true and false assignment distribution of PPDB that is obtained indirectly from the peptide distribution of RSDB.

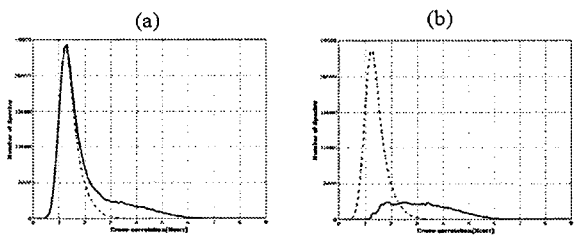

Figure 2 Peptide distribution for the cross-correlation value X_{corr}. (a) The solid line is the distribution for the peptides identified from PPDB. The dashed line is the peptide distribution from RSDB. (b) The dashed line is the peptide distribution from RSDB and reinterpreted as the false assignment distribution of PPDB. The solid line is the true assignment distribution of PPDB that was obtained by subtracting the distribution of RSDB from the distribution of PPDB.

This reversed sequence based analysis can be applied to any match scores for the database search using tandem mass spectral data. For a better discrimination of true and false assignement, we have tried another score F that was introduced at Keller et al. [11].

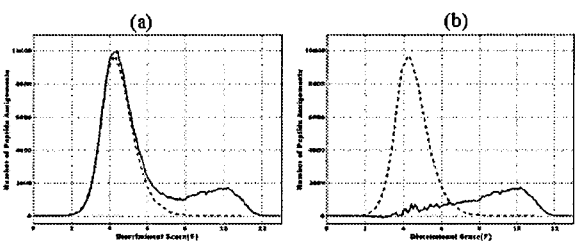

Figure 3 Peptide distribution for the discriminant score F. (a) The solid line is the distribution for the peptides identified from PPDB. The dashed line is the peptide distribution from RSDB. (b) The solid line and dashed line are the true assignment and the false assignment distribution of PPDB, respectively.

Figure 3 shows the distribution for the discriminant score F. Comparing it with Figure 2, F looks to separate true peptides from false ones better than X_{corr}. In order to get more confident peptide identification, the true distribution should be separated from the false distribution as well as possible. Therefore from now on we proceed the further analysis with the score F.

In Figure 4, the protein distributions for PPDB are shown for several reliability values, with varying MW_{corr} values. By increasing the reliability, the tails of the curve far from MW_{corr} = 1 disappear, while the curve around MW_{corr} = 1 changes little. It is because higher score peptides are rarely found far from MW_{corr} = 1.

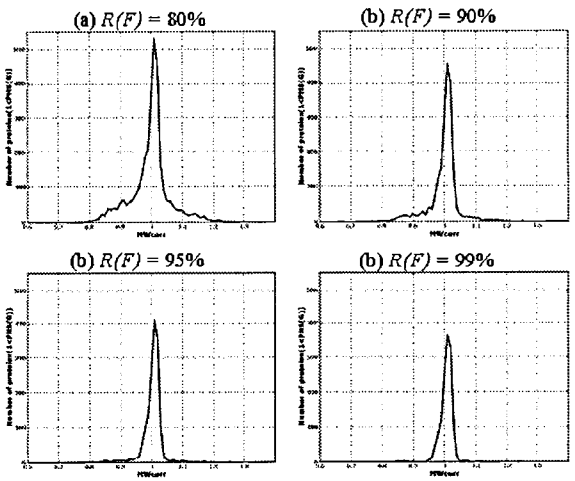

Figure 4 MW_{corr} distribution for threshold scores according to the different reliability values.

For further analysis of the protein molecular weight, we selected peptides with a reliability $R(F)$ > 99% and collected distinct peptide sequences by discarding the redundant peptide sequences. As a result, we obtained 60,862 peptides with high confidence for protein identification. Using the protein identification software DTASelect, these peptides finally identified 2,045 proteins.

When we identify proteins from the peptide list, some proteins contain many peptides and other proteins are identified by only a single peptide. Moreover, one peptide sequence can sometimes be found in several proteins. In that case, such identifications are much less confident than those obtained for multiple peptides [31]. By defining the weight score $W_P(s)$ of a peptide sequence s as a fractional number of

$$W_P(s) = 1/n(s)$$
$$PHS(G) = \sum_{s \in G} W_P(s)$$

we can extend the number of peptide hits of a protein to a new hit score $PHS(G)$. Here $n(s)$ is the number of proteins that contain the tryptic peptide sequence s, the sum of the fractional numbers of peptides that are used for the identification of protein G. As like the usual filtering methods discarding single-peptide protein identification, we abandoned proteins whose $PHS(G)$ was not greater than 1. For the *P. putida* KT2440 sample, these high-confidence proteins were plotted at Figure 5. Figure 5 is the graph of the number of proteins vs. MW_{corr}, that shows a very sharp peak near MW_{corr} = 1. The center of the peak is not located

exactly at the point where $MW_{corr} = 1$. This is caused by the error that occurred when we adopted the linear regression method for the calibration of MW_{exp} from the eight marker proteins. If we use the higher order polynomial function at MW_{corr} definition, this peak center would be shifted to $MW_{corr} = 1$.

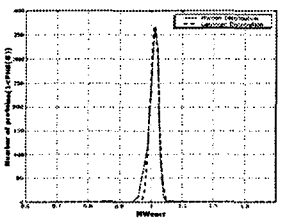

Figure 5 MW_{corr} distribution for proteins with *PHS(G)* larger than 1. The dashed line is the curve fitted to the Gaussian function.

When we interpolate the MW_{corr} distribution curve of Figure 5 into a Gaussian function, we find that the region $0.97 \leq MW_{corr} \leq 1.05$ includes 99.7% of the proteins. Thus, we can define groups G1, G2, and G3 according to their MW_{corr} values as

G1: $1.05 \leq MW_{corr}$
G2: $0.97 \leq MW_{corr} \leq 1.05$
G3: $MW_{corr} \leq 0.97$

Group G2 has a more narrow range than that assigned by Kim *et al.* [4], who selected a 5% error range for MW_{corr} approximately, based on the previous experimental reports. Group G1 is the group in which the theoretical molecular weight of the predicted protein is smaller than the experimental molecular weight. In group G3, the theoretical molecular weight of the protein is larger than the experimental value. Groups G1 contain highly modified proteins, proteins consisting of several chains, protein complexes. G3 is the group of segmented proteins.

Applying the former proteome analysis method to the human sample, we can divide the MW_{corr} range into three groups. The distributions are somewhat different from those of the bacterial proteome. Figure 6 shows the MW_{corr} distribution for the human plasma proteome, filtered by the criterion of *P. Putida* KT2440 proteome with 99% reliability.

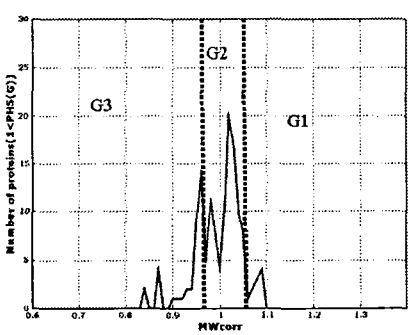

Figure 6 the MW_{corr} distribution of the human plasma proteome

The distribution of the human plasma proteome is quite different from that of *P. putida* KT2440. The human proteins show the distribution more diverse over the MW_{corr} range because of their complex states, which include protein modifications and cleavages. The proteins in group G1 have molecular weights that are larger than those computed from the amino acid sequences in the database. Glycoproteins, proteins containing disulfide bonds and protein complexes are classified into group G1. Group G3 contains the complement proteins and precursor proteins. When the signal peptide is cleaved from the precursor protein, its molecular weight is much lower than the computed molecular weight. The diverse distribution of proteins for MW_{corr} denotes a variety of protein statuses and reflects the fact that many proteins are modified to give large molecular weight differences.

4 DISCUSSION AND CONCLUSION

The protein list for each group defined as G1, G2 and G3 should provide clues as to the dynamics of the proteins. Among the G3 proteins, the precursor proteins could be transferred to group G2 if we knew the signal peptide cleavage site. The SignalP software can predict with high accuracy the cleavage sites for precursor proteins. The expected cleavage sites can be used to generate a new protein sequence database, in which each precursor protein has two different protein IDs, one of which represents the protein without the signal peptide and the other represents the whole protein. After updating the protein database with the proteins from which signal peptides are excised, we can revise the protein molecular weights, and the precursor proteins in group G3 can be moved to group G2. In this scenario, group G3 would then contain only the complement component proteins. Regarding the proteins in group G1, proteins with disulfide bonds, such as immunoglobulins, can have different molecular weights depending on their connection states. By inserting into the protein database the possible connections that are related to the disulfide bonds, the correct theoretical molecular weights could be calculated, and some of these proteins could be moved to group G2. These examples show that updating the protein database with the protein molecular weights facilitates the evaluation of proteins identified by proteomics.

Nowadays the protein databases are exploding with the discovery of new proteins. In order to make database search more efficient, they build trimmed database including one representative sequence for one protein. However such sequence based database is not appropriate for the mass spectrum analysis. In the peptide identification, it loses the homologue sequence peptides. Concerning with the intact protein mass, the trimmed database cannot support the correct information on the protein molecular weight. For the proteomics research using mass spectrometry and protein database, we need to design a new structured architecture of protein database.

5 ACKNOWLEDGEMENT

This project was supported by a grant from MOHW through the Biomedical Proteome Research Center (03-PJ10-PG6-GP01-0002).

REFERENCES

[1] J. K. Eng, A. L. McCormack, J. R. Yates III, An Approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, J. Am. Soc. Mass Spectrom., 5(11): 976—989, 1994.

[2] D. J. C. Pappin, P. Hojrup, A. J. Bleasby, Rapid identification of proteins by peptide-mass fingerprinting, Curr. Biol, 3(6): 327—32, 1993.

[3] J. E. Elias, F. D. Gibbons, O. D. King, F. P. Roth, S. P. Gygi, Intensity-based protein identification by machine learning from a library of tandem mass spectra, Nature Biotechnology, 22(2): 214--9, 2004.

[4] J. Y. Kim, J. H. Lee, G. W. Park, K. Cho, K.-H. Kwon, J. S. Yoo, Utility of electrophoretically derived protein mass estimates as additional constraints in proteome analysis of human serum based on MS/MS analysis, Proteomics, 5, 2005, in press.

[5] K. N. Timmis, Pseudomonas putida: a cospopolitan opportunist par excellence, Environ. Microbiol. 4(12):779-781, 2002.

[6] A. J. Link, Multidimensional peptide separations in proteomics, Trends Biotechnol. 20(12 suppl): S8-13, 2002

[7] R. Pieper, Q. Su, C. L. Gatlin, S. T. Huang, N. L. Anderson, S. Steiner, Multi-component immunoaffinity subtraction chromatography: an innovative step towards a comprehensive survey of the human plasma proteome, Proteomics 3(4): 422-432, 2003.

[8] J. Chen, C. S. Lee, Y. Shen, R. D. Smith, E. H. Baehrecke, Integration of capillary isoelectric focusing with capillary reversed-phase liquid chromatography for two-dimensional proteomics separation, Electrophoresis. 23(18): 3143-3148, 2002.

[9] K. Gevaert, J. Damme, M. Goethals, G. R. Thomas, B. Hoorelbeke, H. Demol, L. Martens, M. Puype, A. Staes, J. Vandekerckhove, Chromatographic isolation of methionine-containing peptides for gel-free proteome analysis: identification of more than 800 Escherichia coli proteins, Mol. Cell. Proteomics 1(11):896-903, 2002.

[10] R. E. Moore, M. K. Young, T. D. Lee, Qscore: an algorithm for evaluating SEQUEST database search results, J. Am. Soc. Mass Spectrom., 13(4): 378-386, 2002.

[11] A. Keller, A. I. Nesvizhskii, E. Kolker, R. Aebersold, Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Databas Search, Anal. Chem. 74(20):5383-5392, 2002.