Bioinformatics and Biosystems 2007, Vol. 2, No. 2, pp. 46-51

Proteomics Data Analysis using Representative Database

Kyung-Hoon Kwon[1], Gun Wook Park[1], Jin Young Kim[2], Young Mok Park[1], Jong Shin Yoo[3*]

[1]Systems Biology Core Facility ,[2]Proteomics Team ,[3]Division of Instrument Development , Korea Basic Science Institute, Eoeun-dong, Yusung-gu, Daejeon 305-333, Republic of Korea

Email: khoon@kbsi.re.kr, jongshin@kbsi.re.kr

ABSTRACT

In the proteomics research using mass spectrometry, the protein database search gives the protein information from the peptide sequences that show the best match with the tandem mass spectra. The protein sequence database has been a powerful knowledgebase for this protein identification. However, as we accumulate the protein sequence information in the database, the database size gets to be huge. Now it becomes hard to consider all the protein sequences in the database search because it consumes much computing time. For the high-throughput analysis of the proteome, usually we have used the non-redundant refined database such as IPI human database of European Bioinformatics Institute. While the non-redundant database can supply the search result in high speed, it misses the variation of the protein sequences. In this study, we have concerned the proteomics data in the point of protein similarities and used the network analysis tool to build a new analysis method. This method will be able to save the computing time for the database search and keep the sequence variation to catch the modified peptides.

1. INTRODUCTION

In the bottom-up proteomics, the sample proteins are digested into peptides and the amino acid sequences of the peptides are obtained through the analysis of tandem mass spectra for their fragment ions. This proteomics technology has contributed largely to the protein analysis. Especially, the development of high-throughput proteomics using multi-dimensional liquid chromatography [1] has enhanced the accuracy and performance in the protein quantitation and characterization as well as protein identification.

In the high-throughput proteomics approach, a set of tandem mass spectra for each fraction are used to search the peptide sequences from the protein sequence database, by the software such as Sequest[2], Mascot[3], Phenyx[4], X!Tandem[5] and Spectrum Mill(Agilent, UK). The database search software generates the *in silico* tandem mass spectra of peptide sequences of protein database. These calculated mass spectra are compared with each experimental real spectrum and the best matched one is selected for the protein identification. Several database search software accompany the protein identification result with the identified peptide sequences. For example, Mascot looks for the matched peptide sequences and the peptide sequence matches give the protein list which is scored based on the probability. The software Sequest does not supply the protein list. It only output the peptide sequence list. In order to get the proteins from Sequest result, we need additional program to integrate the Sequest result. The programs such as DTASelect (Scripps Institute, San Diego, USA) gather the MS/MS ion search result and analyze peptide sequence list to get most probable proteins.

The software DTASelect uses the simple process to get proteins from peptide sequences. It just looks for all the proteins containing the identified peptides. It does not care whether it is significant or not. In many cases, we have the peptide sequences which are found at many different proteins. If we are to filter the false positive proteins, we need to make more analysis on the proteins in the sample.

The database search process uses the established sequence database. The database search software suggests the most probable proteins among the protein sequences in the given database. If there is no match in the database, we can use *de novo* sequencing method[6]. But, it spend much computing time and it requires high quality spectrum to reach the correct sequence. The *de novo* sequencing method is not the good solution for the high-throughput proteomics. Instead, we are

Corresponding author: Jong Shin Yoo (Email:jongshin@ kbsi.re.kr)
This work was supported by a grant to Y.M.Park from the Korea Institute of Science & Technology Evaluation and Planning (M6-0403-00-0154) and a grant to J.S.Yoo from the 21C Frontier Functional Proteomics Project from Korean Ministry of Science & Technology (FPR-05-A2-300).

looking for the optimal way to find out the correct peptide sequences as many as possible. Since the proteins are identified from a set of peptide sequences, it is possible to throw away the incorrect peptide sequences in the protein identification.

In order to improve the confidence of protein identification, there have been many efforts to estimate the identification result[7-11]. Particularly, the decoy approach using reversed sequence database[12] made it possible to evaluate the false positive probability and it has been applied to many research[13,14]. The PeptideProphet / ProteinProphet[11,15], PROT_PROBE[8,9] proposed systematic analysis methods for the protein identification by applying machine learning of neural network technique and statistical analysis. The probabilistic approach by the statistical methods is aimed to assign different weight value to the each protein containing common peptide sequences. We can pick the peptide sequences that have won the highly reliable match score by applying the statistical analysis methods and we can choose the proteins with many peptides matches. A series of these analysis processes enables us to enhance the confidence for the identified protein list. However, still we have some problems to solve for the protein identification in bottom-up proteomics.

The process to identify proteins from peptide sequences includes the ambiguity when the peptide sequence exists commonly at the different proteins. In the high-throughput proteomics, we use the protein mixture samples. Since the protein mixture is digested into peptides, it is hard to distinguish the protein from which a peptide comes. When one protein is identified by a set of peptide sequences and another protein contains the same set of peptide sequences or a part of the peptide set, all these proteins are possible to exist in the sample. Sometimes, tens of proteins are listed for one set of matched peptides.

Nesvizhskii et al.[16] considered and analyzed the shared peptides in the protein quantitative analysis and he suggested a conceptual way to determine the amount of proteins from the quantitative analysis ofshared peptides. However, the accuracy of peptide quantities measured by mass spectrometer is not enough to discriminate the expressions of proteins with shared peptides in the current analysis technology. Identifying proteins from shared peptides is one of the intrinsic problems in bottom-up proteomics. In the journal 'Molecular & Cellular Proteomics', they published the guideline to submit papers about proteomics research. They asked authors to follow the parsimony rule in deciding protein list[17]. They recommend to select the minimal proteins that can explain the set of peptide sequences.

In this study, we start from the assumption that the proteins sharing several peptides will consist of similar sequences. We tried to convert the protein database into the protein group database where similar protein sequences are assigned to one group and each group has a representative protein. This database of representative proteins is called as representative database. Such a representative database of protein sequences is designed for improving the efficiency of BLAST search[18]. Since the BLAST search looks for the sequences similar to the query sequence, the representative database has worked well[19]. However, using representative database at the bottom-up proteomics can make a loss of valuable sequence data. The database search in proteomics relies on the molecular weight value of fragment ions obtained from the peptides. If the peptide sequence of the sample is a little bit different from the peptide sequence of the database, the peptide sequence in the database does not satisfy the tandem mass spectra and it cannot be found as a match. When we generate the representative database, we only select one representative sequence in a group and abandon the other sequences.

Although such a loss could be occurred, we tried to use representative database. In the database search of the proteomics, we look for the confident protein list. If a protein is not false positive, we expect that we can find many peptide sequences in that protein. It is called the protein of multiple peptide hit. If a protein is identified by only one peptide, it is in doubt to be included in the sample. We abandon single peptide hit because it includes many false positives. It is the reason why we could use the representative database in the proteomics data analysis. If the representative protein is similar to the protein of the sample, it will contain at least one of peptide sequences of the sample protein that would be identified from the tandem mass spectra. We made the database search with representative database and took the representative protein list. In this search, we also took the proteins of the single peptide hit. After the database search, we could make a candidate protein list from the identified representative proteins. And we could expect that the real protein will be included in this secondary protein list. By the database search with representative proteins, we classified the proteins related to the representative proteins identified from the tandem mass spectra. This analysis results to the fact that most proteins with shared peptides can be grouped by the sequence similarity.

The second database search can be done for the selected protein groups related to the representative proteins which were identified in the first database search. This search process will be useful to identify the real protein sequences.

Another purpose of implementing the representative data-

base is to suggest more efficient way to get protein list from mass spectral data and to search the peptides modified by post-translational modification. In the database search program, the post-translational modification can be applied as one of the search parameters. The possible post-translation modification makes the database search process much slower and it can bring many false positive matches. Most researches on the post-translational modification are performed with the restricted number of proteins. It is not appropriate for the high-throughput analysis. For more efficient search of the post-translational modification, it will be better to reduce the database size by filtering out the irrelevant proteins.

Usually the protein database contains tens or hundreds of thousand proteins and we take several thousands proteins as meaningful data from the protein database in the proteomics approach. Only a few percents of the proteins in the database are useful. Moreover, they can be grouped by the sequence similarity. Therefore, if we identify protein groups by the representative protein database, we can confine the candidate proteins as the protein groups selected. We construct smaller protein database only including these candidate proteins. Once we manifest the protein groups included in the sample, we can make database search with this new database. The representative database allows us to perform database search with post-translational modification for more probable and smaller protein database.

METHODS

As a reference experimental data of the proteome, we have used the proteomics data of human brain tissue and we have made SEQUEST search with the whole IPI human database. The experimental details were explained at Y.M. Park et al.[21] Among the human brain proteome data, we have chosen the dataset of better quality. The PE fraction of the sample was analyzed. The PE fraction is one obtained from cell fractionation process and it is known that the membrane and structural proteins are included[22]. The sample is separated into 50 gel bands by the one-dimensional electrophoresis. We excised the bands and digested each band with the trypsin. After the tryptic digestion, the fractions are separated by liquid chromatography of SCX column and reverse phase column. From the succession of separation process of electrophoresis and liquid chromatography, we could get the better quality of tandem mass spectra. The tandem mass analysis was performed by LTQ-MS (Thermo Scientific, Inc.) with a nano-ESI source. LTQ-MS is the high resolution mass spectrometer whose instrumental resolution is reported to be several ppm. From the database search of tandem mass spectral data, we could have 10,486 peptide sequences identified from which we obtained 5,584 proteins[21]. In the database search of SEQUEST, we have used precursor ion mass tolerance as 1.5 Da, maximum missed cleavage is 1. The modifications of carboamidomethyl and methionine oxidation were allowed. The database search was restricted to the tryptic peptide sequences which are cleaved only at the C-terminal of Arginine or Lysine.

For the network analysis of proteomics data, we have used the database clustering software named CD-HIT[20]. CD-HIT is the software to cluster the similar sequences and make representative sequence. When it clusters the similar sequences as a group, it selects the longest sequence as its representative sequence.

In order to investigate the sequence similarity characteristics for the identified proteins, we constructed the representative database from IPI human database v.3.15. By CD-HIT, the clustering software, we made protein groups whose elements are collected by the condition that their similarity was higher than 60%. This representative database contained 24,120 proteins among 48,193 proteins.

RESULT AND DISCUSSION

In Figure 1 and Figure 2, we displayed the first database search result with the representative protein database and the second database search with related grouped protein database, respectively. For the visualization, we have used network analysis tool named Pajek [23].

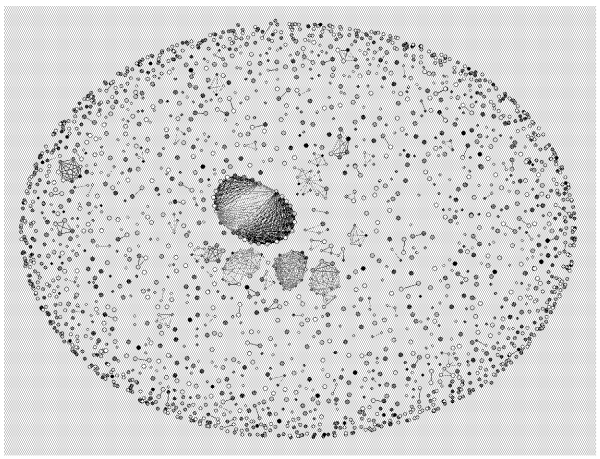

Figure 1. Protein similarities in IPI representative database

Each node of the network stands for the identified protein and its color denotes the protein group. For the representative database, all the nodes colored differently. We have identified 2,136 nonredundant representative proteins. For the grouped protein database, the nodes of the same color show the protein elements of the same group. In Figure 2, we can see many nodes coloured by the same colour. In the second search, we have identified 2,934 nonredundant proteins. In the network diagram, we connected two nodes by a black line if one protein has the same set of matched peptides with the other protein. If one protein shows the partial matches with peptides of the other protein, they were connected by a gray line.

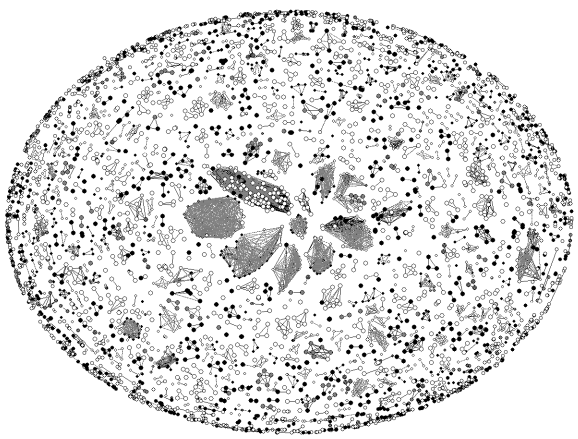

Figure 2. Protein similarities in IPI human database related to the selected representative proteins of Figure 1.

In Figure 1, although we have selected only one protein from the group whose sequence similarity is higher than 60%, there are many nodes connected. This means that some peptides are shared by many different proteins. The proteins containing common peptide sequences can belong to the same category of proteins or the different category. From Figure 2, for each representative protein, its group elements were found as the proteins sharing the peptide sequences with the representative protein. Figure 2 shows the similar proteins of the same group were connected by black or gray lines. The nodes with the same colors are the proteins whose sequence similarity is higher than 60%.

Table 1 is the list of the number of identified proteins for each database search. For the representative database search, the ratio between the redundant proteins and nonredundant proteins is much smaller than for the grouped protein database search. 109% of nonredundant proteins are identified as redundant proteins. It is because the representative database consists of less similar protein sequences. Concerning with the proteins identified by the whole IPI database, we got 2,944 non-redundant proteins. Following our representative database approach, we could get 2,934 proteins. We have lost only 0.3% proteins, compared to the whole database search.

Table 1. Summary of the human brain proteome analysis result with IPI human database v3.15.

	IPI human database	IPI human representative	IPI human selected groups	NCBI nr selected groups
Unfiltered proteins	48,193	24,120	6,860	32,916
Redundant proteins	5,584	2,336	5,288	22,895
Non-redundant proteins	2,944	2,136	2,934	4,090
Redundant peptides	10,486	5,585	11,066	17,500
Non-redundant peptides	6,124	5,177	6,569	6,580

The last column of the Table 1 is the database search result using NCBI nr database. NCBI nr database is the protein database served by the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). It is the integrated database of the several other database and contains a variety of protein sequences. Usually, we can hardly use the whole NCBI nr database for the proteomics data analysis, because it is the huge database.

For the human proteins, NCBI nr database contains about 300,000 proteins. However, the representative database approach enable us to make a database search of NCBI nr database. We made protein groups from NCBI nr database. Based on the IPI representative database, we connected NCBI nr protein sequences to their most similar IPI representative proteins. This process was done by CD-HIT, too. From our representative database approach, we could reduce the possible protein candidates among 283,548 proteins of NCBI nr human database down to 32,916 and we made the database search to get 4,090 non-redundant proteins from NCBI nr.

As an example of the additional information obtained from this approach of NCBI nr database, we could find a cleaved protein. The IPI human database is filtering the redundant protein sequences and keeps only the larger sequences. If we have two protein sequences of whole protein and cleaved protein, IPI database abandon cleaved protein. However, we can find that cleaved protein in the NCBI nr database. The protein, 'Protein disulfide-isomerase A6 precursor' was included in the IPI human database as the following sequence :

>IPI00644989.1 Protein disulfide-isomerase A6 precursor
MALLVLGLVSCTFFLAVNGLYSSSDDVIELTPSNFNREVIQSD
SLWLVEFYAPWCGHCQRLTPE···

And NCBI nr database had its cleaved sequence where N-terminal peptide was cleaved.

>gi|1710248 Protein disulfide isomerase-related protein 5
LYSSSDDVIELTPSNFNREVIQSDSLWLVEFYAPWCGHCQRLTPE
.....

From the tandem mass spectra, we could identify the peptide sequence of 'LYSSSDDVIELTPSNFNR' in the NCBI nr database by the high match score, while it could not be found in IPI human database. In the IPI human database, this peptide could not assigned as a tryptic peptide because of the N-terminal peptide attached.

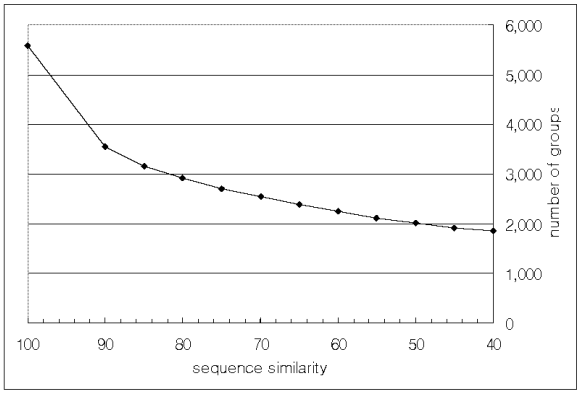

Figure 3. The change of the number of clusters for human brain proteome for the different similarity thresholds.

For the supplementary analysis, we have considered the various sequence similarity of identified proteins. In order to see the sequence similarity distribution of proteins identified by the mass spectrometry, we used CD-HIT with the identified protein list for the similarities 90%. By this high similarity, we got 4,025 representative sequences. Next, by the clustering of this representative proteins for the similarity 85%, we got 3,161 representative proteins.

By decreasing the similarity thresholds from 90% down to 40%, we could see how the number of clusters decreases. Figure 3 shows the number of clusters for the different similarity value. The slope of the curve becomes smooth around the point of similarity 70%. When the similarity threshold decreases from 100% to 70% we could bind many proteins into protein groups. Decreasing the similarity from 70% does not affect the number of protein groups.

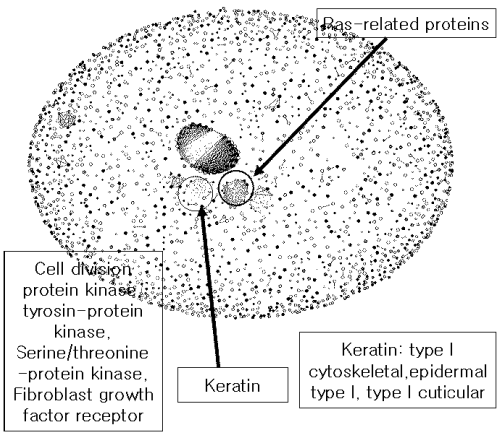

Figure 4. The proteins including common peptide sequences in the representative IPI human database for the human brain proteome.

In the protein groups clustered by the sequence similarity 60%, we analyzed their functions. Figure 4 shows the protein names of some larger clusters in the protein list obtained from the representative IPI database. The largest cluster was the group of the cell-division protein kinase, tyrosine-protein kinase, Serine/Threonine protein kinase, Fibroblast growth factor receptor. These proteins share one peptide '[LI]ADFGLAR'.

CONCLUSION

In this study, we have suggested the new method of database search in the proteomics data analysis. We have prepared the representative database of IPI human database and analyzed the human brain proteome data. The representative database was generated by the representative sequences of the protein groups where the group elements are similar to each other by the sequence similarity higher than 60%. We have achieved the two level database search. The first search was done by the representative database to get the categories of proteins that were included in the sample. With the first search result, we built the secondary database of the proteins belonging to the groups whose representative proteins were selected in the first search. By the next search of secondary database, we could find the various peptide matches to the tandem mass spectra. This approach was extended to the NCBI nr database. We grouped NCBI nr database by the IPI representative proteins. It was useful to overcome the limit of non-redundant database, without sacrificing the computing time.

REFERENCES

[1] A.J. Link, Trends Biotechnol. 20:S8-S12, 2002.
[2] J.K. Eng, A.L. McCormack, J.R. Yates III, J. Am. Soc. Mass Spect. 5: 976-989, 1994.
[3] D.N. Perkins, D.J. Pappin, D.M.Creasy, J.S. Cottrell, Electrophoresis, 20:351-3567, 1999.
[4] Colinge, J., et al., Proteomics, 4: 1977-1984, 2004.
[5] R. Craig, R.C.Beavis, Bioinformatics, 20: 1466-1467, 2004.
[6] Mo, L., Dutta, D., Wan, Y., Chen, T., Anal. Chem., 6, 2007, in press.
[7] D.C. Anderson, W. Li, D.G. Payan, J. Proteome Res. 2: 137-146, 2003.
[8] R.G. Sadygov, J.R. Yates III, Anal. Chem. 75:3792-8.
[9] R.G. Sadygov, H. Liu, J.R. Yates III, Anal. Chem 76: 1664-1671, 2004.
[10] F. Li, W. Sun, Y. Gao, J. Wang, Rapid Commun. Mass Spectrom., 18: 1655-9, 2004.
[11] A. Keller, A.I. Nesvizhskii, E. Kolker, R. Aebersold, Anal. Chem., 74: 5383-5392, 2002.
[12] Elias, J.E., Gibbornsm F.D., King, O.D., Roth, F.P., Gygi, S.P., Nat. Biotechnol. 22: 214-219, 2004.
[13] Park, G.W., et al., Proteomics 6: 1121-1132, 2006.
[14] Stephan, C., et al., Proteomics 6: 5015-29, 2006.
[15] A. Nesvizhskii, A. Keller, E. Kolker, R. Aebersold, Anal. Chem, 75: 4646-4658, 2003.
[16] A. Nesvizhskii, R. Aebersold, Mol. Cell. Proteomics, 4(10):1419-40, 2005.
[17] Bradshaw, R.A., (Editorial), Mol. & Cell. Proteomics 1223-1225, 2005.
[18] Holm, L. Sander, C., Bioinformatics, 14, 423-429, 1998.
[19] Park, J., et al., Bioinformatics, 16(5), 458-464, 2000.
[20] W. Li, L. Jaroszewski, A. Godzik, Bioinformatics, 17:282-3, 2001.
[21] Y.M. Park, et al., Proteomics 6: 4978-4986, 2006.
[22] Klose, J., et al., Nat. Genet. 30, 385-393, 2002.
[23] V. Batagelj, A. Mryar, Connections 21, 2: 47-57, 1998