Reinterpretation of the database search result using Trans-Proteomics Pipeline in mass spectrometry-based proteomics

Kyung-Hoon Kwon
khoon@kbsi.re.kr

Sang Kwang Lee
lskwang@kbsi.re.kr

Kun Cho
chokun@kbsi.re.kr

Gun Wook Park
cancun@kbsi.re.kr

Byeong Soo Kang
bskang@postech.ac.kr

Young Mok Park
ympark@kbsi.re.kr

1 Mass Spectrometry Research Division, Korea Basic Science Institute, 804-1, Yangcheong-ri, Ochang-eup, Chungwon-kun, Chungbuk 363-883, Republic of Korea
2 Graduate School of Analytical Science and Technology, 79, Daehak-ro, Yusung-gu, Daejeon 305-764, Republic of Korea
3 The I-BIO graduate program and National Core Research Center for Systems Bio-Dynamics, POSTECH, Pohang, Kyungbuk 790-784, Republic of Korea

Keywords: proteomics, mass spectrometry, protein identification, TPP pipeline, probability score

1 Introduction

The proteomics using mass spectrometry has been a powerful tool to identify, characterize and quantify the proteins included in a sample. The mass spectral data which was produced by the high-throughput and high-resolution mass spectrometer is used for the database search from which we can find the best match of protein sequence with mass spectrum peaks. The database search algorithm defines their own match score and display the scores with matched sequences. Since many search algorithms have been released and their results are somewhat different from the other results. For the different result, we need to confirm the identified proteins. In order to get coincident result, here we tried to compare the search result from different search algorithms. We need additional tools to compare the results from different search algorithms. The Trans-Proteomics Pipeline (TPP) of Institute for Systems Biology supports the probabilistic scores for search software of Mascot, Sequest and X!Tandem.[1] The probabilistic score can be useful to measure the false positive rate of search result.

In this study, we got the proteomics data from human mesenchymal stem cell and made database search under the different softwares such as Mascot, Sequest and X!Tandem. By changing the threshold score, we compared the results from these three softwares.[2]

2 Method and Results

The protein extract of human mesenchymal stem cell was used for the proteomics analysis. By the separation of one-dimensional electrophorysis, the proteins were separated by gel-bands which were excised. The proteins of each band were digested by trypsin nd analyzed by LC/MS/MS with FT LTQ (Finnigan). Finally we got tandem mass spectra of peptide ions. The database search of Mascot, Sequest, and X!Tandem was used for the protein identification and Trans-Proteomics Pipeline(TPP) compared these results. (Figure 1)

3. Results

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 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. TPP pipeline defines the probability score to compare the results from different software. When we compare the peptide identification result for the same threshold probability score from different search software, the difference in search result appears. Some proteins which were identified at one software were missing at the other software. However, when we shift threshold probability scores, many of missed proteins came up.

Figure 1: Workflow for the experiment and data analysis

3 Discussions

In order to get more proteins while keeping the high confidence level, it is necessary to make database search with different softwares and integrate the result by threshold probability value. However, from our result, we found that the threshold probability should be corrected in some softwares. For example, the probability score of Sequest seems to be underestimated. We need the recalibration of probability score for several database search software.

References
