An Improved Target-decoy Strategy for Evaluation of Database Search Engines and Quality Control Methods in Shotgun Proteomics

Xiao-dong FENG[1,2], Jie MA[1], Cheng CHANG[1], Kun-xian SHU[2*] and Yun-ping ZHU[1*]

[1]Institute of Radiation Medicine, State Key laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences, Beijing 102206,China.

[2]Chongqing University of Posts and Telecommunications, Chongqing 400065, China

*To whom correspondence should be addressed:

Keywords: Proteomics, Tandem mass spectrometry, Entrapment sequences method, Target-decoy, Quality control.

Abstract. With the advance of mass spectrometry and experimental techniques, proteome research has broken through the bottleneck of data generation, and a huge amount of mass spectrometry (MS) data has been accumulated rapidly in the past few years. Meanwhile, the lack of efficient data analysis and quality control methods has greatly hindered proteome development. Target-decoy searching strategy has become one of the most popular strategies to control the false identification in MS/MS data analysis. While this strategy can estimate the false discovery rate (FDR) within a dataset, it cannot directly evaluate the false positive matches in target identifications. In this study, we developed an improved target-decoy strategy: the entrapment sequences method, to set up an objective standard to evaluate the performance of quality control methods and database search tools. We started with a preliminary study of the size of entrapment sequences and found ten times the sample sequences could be a reasonable size of entrapment sequences. Then,we went on to give thedefinition and equation of Estimated FDR and Actual FDR. We found the entrapment sequences method can be a good supplement to target-decoy strategy.

Introduction

In the era of postgenomics, proteomics research has become increasingly important to the research of life sciences. Mass spectrometry based proteomics research can provide a large amount of information on protein identification and quantification because MS/MS can analyze protein mixtures in a high throughout manner and provide sequence information for peptides and proteins.[1]Target-decoy searching strategy has become one of the most popular strategies to control false identification in MS/MS data analysis.[2]In target-decoy strategy, sample sequences are usually used as target sequences before being reversed or randomized as decoy sequences. One key assumption is that the number of PSMs from the decoy database equals the number of false identifications from the target database, which permits the FDR estimation. However, the target-decoy strategy can estimate the false discovery rate (FDR) within a dataset, but it cannot directly evaluate the false positive matches in target identifications. In this study, we introduced the entrapment sequences method to formulate an objective standard to evaluate the performance of quality control methods and database search engines. As shown in Fig. 1, in the entrapment sequences method, the target sequences are composed of sample sequences (A) and entrapment sequences (B), which are of low homology with the sample sequences,

then the combined target sequences (A+B) are reversed to construct decoy sequences (A'+B'). By using different labels, we can separate the PSMs into different kinds and calculate the actual FDRs for the PSMs.

Figure 1. Database used in entrapment sequences method.

Target sequences are comprised of sample sequences (A) and entrapment sequences (B).Then we reverse both of them (A+B) as decoy sequences (A'+B').

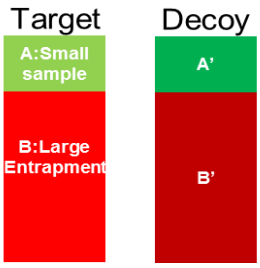


Figure 1.

In our previous work, we used entrapment sequences method to evaluate the performance of four quality control methods (BNP model, PeptideProphet, cutoff-based method and nonparametric method)[3].Granholm*et al.* used a similar method to evaluate several commonly used score functions.[4] In Marc Vaudel*et al*'s work for quality control of the FDR estimation, they defined the FDR calculated on the target-decoy strategy as the "Estimated" FDR and the FDR calculated using only the hits from the target database as the "Reference" FDR[5].But we would argue that it is more appropriate to call the "Reference"FDR "Actual" FDR. In the year of 2012, Marc Vaudel*et al.*also used a similar method to prove that *pyrococcusfuriosus*sample can be used as a reliable reference sample of known content,[6] which is better than the widely used standard samples of known content.[7]

All the above work used much larger entrapment sequences than sample sequences, only to ignore the random hits in sample sequences. In Granholm*et al*'s work[4],he said "An entrapment database infinitely larger than the sample partition is likely to capture all top-scoring PSMs, making it equivalent to a normal decoy database.The ideal proportion of sample and entrapment sequences,for the purpose of creating the optimal null model,thus remains to be elucidated." So in this work, we started with a preliminary study on the size of entrapment sequences, hopingto solve the following two problems: (1) At what proportion of entrapment/sample sequences can the random hits in sample sequences be ignored? (2) If the random hits in sample sequences cannot be ignored, how can we estimate the number of false matches in sample sequences? Then, we proceeded to the definition and equation of Estimated FDR and Actual FDR. We demonstrated that the entrapment sequences method could be an excellent strategy to assess each step of the mass spectrometry data analysis process.

Material and methods

The*Pfu* dataset wasproduced by analyzing*Pyrococcusfuriosus* sample on LTQ OrbitrapVelos (Thermo Scientific) [6], and used as anstandard dataset here.All data searching was performed using combined target-decoy strategy. The target database comprised sample sequences and entrapment sequences.The decoy database was created by reversing all the target sequences. Three kinds of databases were

downloaed fromUniProt[8] database on 5th Jan. 2016. 1) *Pyrococcus furiosus*protein sequences of Swiss-Prot(abbreviationPfu, containing 2,045 sequences).2) *Homo sapiens*protein sequencesof Swiss-Prot(abbreviation HomoSp, containing 20,187 sequences).3) Homo sapiens of TrEMBL(abbreviation HomoTr, contraining49,889sequences).We expanded human protein sequences by randomizing HomoSp and HomoTr using Matrix Science Company's perl module, which could be downloaded free at http://www.matrixscience.com/help/decoy_help.html.To evaluate the best proportion of entrapment/sample sequences, we constructed 13 database of increasing sizes by combining*Pfu* protein sequences with different numbers of human protein sequences. The datails were shown in Table 1.All mzML files were converted from raw files using the msconvertmodule[9] in the Trans-Proteomic Pipeline (TPP v4.7.0)[10].The monoisotopic mass was used for both peptide and fragment ions with fixed modification (Carbamidomethyl, +57Da) on Cys and variable modification (oxidation, +16Da) on Met. Tryptic cleavage at only Lys or Arg was selected. Only b and y fragment ions were taken into account, so were fragment ions with +2, +3 or +4 charge.

The Appropriate Proportion of Entrapment and Sample Sequences

A commonly-used method to demonstrate that a quality control method or a database search engine is reliable is to test it on proteomic standards of knowncontent. These analyses allow the discrimination of true andfalse positives based on the known sample composition andthus allow the estimation of reference metrics like a referencefalse discovery rate (FDR). The PSMs pointing to the proteinsactually in the sample are considered true positive matches,whereas all other hits are suspected to be errors.Marc Vaudelet al[6].introduced two reference metrics of controlled samples: (1) no true positive PSMs shall hit the entrapment database, and (2) no random match shall hit the sample sequences. In this study, we extended these two reference metrics for evaluation. Meeting metric (1) can be achieved using sample sequences and entrapment sequences with low homology. Using blast local server(v2.4.0+),[11] we compared the homology of 2,045 *Pfu* sequences and 70,076 human sequences, and found a very low homology. Meeting metric (2) can be achieved by using a much larger entrapment database than the sample database, for the possibility of a random match hitting the sample database is negligible. But how large should the entrapment database be? For the entrapment database which is not so large as the sample database, can we use the hits in this entrapment database to estimate the false matches in the sample database? So in this section, we did preliminary research on the appropriate proportion of entrapment and sample sequences. In the next section,we gave the definition and equation of Estimated FDR and Actual FDR.

Pfu dataset was searched by MS-GF+ using combined target-decoy strategy. We used 13 kinds of target database illustrated in Table 1.

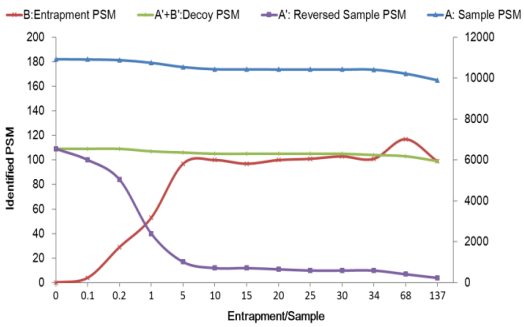
Table 1.Target database used in this work

Database	Pfu (Sample)	Homo (Entrapment)	Proportion(entrapment/sample)
(1)	2045	0	0
(2)	2045	205	0.1
(3)	2045	409	0.2
(4)	2045	2045	1
(5)	2045	10225	5
(6)	2045	20187	10
(7)	2045	30675	15
(8)	2045	40900	20
(9)	2045	51125	25
(10)	2045	61350	30
(11)	2045	70076	34
(12)	2045	140152	68
(13)	2045	280304	137

We used 2045Pfu sequences as sample sequences and different kinds of Homo sequences as entrapment sequences. So the proportion of entrapment sequences/sample sequences various from 0 to 137.

Then we reversed the target database as decoy database. So the proportion of entrapment sequences/sample sequences ranged from 0 to 137. The results were shown in Figure 2:When the proportion of entrapment/sample sequences was small, there were few entrapment PSMs identified. As the proportion grew, more entrapment PSMs were identified and graduallyon the verge of identified decoy PSMs. But the number of sample PSMs was somewhatreduced. We found the appropriate entrapment/sample sequences proportion was10-30. For the purpose of getting more sample results, we recommended the entrapment/sample sequences proportion to be 10 in our study.

Figure 2.The appropriate proportion of entrapment and sample sequences.



When the proportion of entrapment/sample sequences is small, very few entrapment PSMs are identified. With the proportion grows, more entrapment PSMs are identified and gradually on the verge of identified decoy PSMs. But the number of sample identifications keep stable.

Estimated FDR and Actual FDR

To calculate Estimated FDR, we used the results from both target database and decoy database without distinguishing sample sequences from entrapment sequences in the target database. We can calculate the false discovery rate (FDR) using equation (1).Where N_{fp} stands for the number of false positive PSMs in target database and N_{target} stands for all the number of PSMs in target database.

$$FDR = \frac{N_{fp}}{N_{target}} \tag{1}$$

As we cannot count the number of false positive PSMs in target database directly, we can use the number of decoy PSMs (N_{decoy}) to estimate the number of false positive

$$FDR_{est} = \frac{N_{decoy}}{N_{target}} \tag{2}$$

PSMs in the target database （N_{fp}=N_{decoy}）. So the Estimated FDR can be calculated by equation (2).

To calculateActual FDR, weused the results from the target database only and distinguished sample sequences from entrapment sequences in the target database.As illustrated by Marc Vaudel*et al*[6]., the number of random matches insample sequences (A) can be estimated by reversed sample PSM (A'). And in an ideal situation, the number of random matches in (A) should be 0. Our work showed that as the proportion grew, the number of identified PSMs declined significantly.Even at the highest proportion of 137 (2045 sample sequences+280304 entrapment sequences), there still existed PSMs in (A'), suggesting that in sample sequences there still exist randem hits that can be ignored when the entrapment/sample sequences proportion is large.We found that the number of random hits in sample sequences (N_{fps}) could be estimated by the number of PSMs in entrapment sequences (N_{trap}), which is illustrated by equation (3). Where S_{sample} stands for the size of sample sequences, S_{trap} stands for the size of entrapment sequences.

$$N_{fps} = \frac{S_{sample}}{S_{trap}} N_{trap} \tag{3}$$

So all the false postive PSMs in the target database (N_{fp}) can be estimated by equation (4).

Compared with equation (1), we can calculate Actual FDR (FDR_{act}) using the target

$$N_{fp} = N_{trap} + N_{fps} \tag{4}$$

database alone, which is illustrated in equation (5).

$$FDR_{act} = \frac{N_{trap} + N_{fps}}{N_{target}} = \frac{N_{trap}}{N_{target}}(1 + \frac{S_{sample}}{S_{trap}}) \tag{5}$$

$$FDR_{act} = \frac{N_{trap}}{N_{target}} \tag{6}$$

For a target database which has a large entrapment/sample proportion, the value of (S_{sample}/S_{trap}) can be ignored. In this case, the Actual FDR (FDR_{act}) can be calculated by equation (6).

Discussion

In this study, we developed an entrapment sequences method to assess each step of the mass spectrometry data analysis process.By using different labels, we can separate the PSMs into different kinds. So that we can set up an objective andexecutable metric for the evaluation of database search engines, quality control methods and protein assembling tools. As shown in our previous study, we found the appropriate proportion of entrapment/sample sequences is 10-30. And even at the proportion of 137, the number of sample identificationsreduce a little. Entrapment sequences method can perform evaluations easily and cause less reduction of sample identifications,this is why it’s recommended for laboratory use.

Acknowledgements

This work was supported by grants from Projects of International Cooperation and Exchanges(2014DFB30010), National High Technology Research and Development Program of China (2015AA020108, 2013CB910800), National Natural Science Foundation of China (21275160, 21475150),andChongQing postgraduate scientific research and innovation projects(CYS14154).Xiao-dong FENG and Jie MA contributed equally to this work.

References

[1] Hernandez P, Muller M, Appel R D. Automated protein identification by tandem mass spectrometry: issues and strategies. Mass spectrometry reviews, 2006, 25(2): 235-254.

[2] Elias J E, Gygi S P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nature methods, 2007, 4(3): 207-214.

[3] Zhang J, Ma J, Dou L, et al. Bayesian nonparametric model for the validation of peptide identification in shotgun proteomics. Molecular & cellular proteomics : MCP, 2009, 8(3): 547-557.

[4] Granholm V, Noble W S, Kall L. On using samples of known protein content to assess the statistical calibration of scores assigned to peptide-spectrum matches in shotgun proteomics. Journal of proteome research, 2011, 10(5): 2671-2678.

[5] Vaudel M, Burkhart J M, Sickmann A, et al. Peptide identification quality control. Proteomics, 2011, 11(10): 2105-2114.

[6] Vaudel M, Burkhart J M, Breiter D, et al. A complex standard for protein identification, designed by evolution. Journal of proteome research, 2012, 11(10): 5065-5071.

[7] Klimek J, Eddes J S, Hohmann L, et al. The standard protein mix database: a diverse data set to assist in the production of improved Peptide and protein identification software tools. Journal of proteome research, 2008, 7(1): 96-103.

[8] Apweiler R, Bairoch A, Wu C H, et al. UniProt: the Universal Protein knowledgebase. Nucleic acids research, 2004, 32(Database issue): D115-119.

[9] Kessner D, Chambers M, Burke R, et al. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics, 2008, 24(21): 2534-2536.

[10] Deutsch E W, Mendoza L, Shteynberg D, et al. A guided tour of the Trans-Proteomic Pipeline. Proteomics, 2010, 10(6): 1150-1159.

[11] Altschul S F, Gish W, Miller W, et al. Basic local alignment search tool. Journal of molecular biology, 1990, 215(3): 403-410.