False positive assessment in database search

TP647

Wilfred H. Tang, Sean L. Seymour, Alpesh A. Patel, Applied Biosystems|MDS Sciex, 850 Lincoln Centre Dr., Foster City, CA 94404

ABSTRACT

- Compare two typical strategies for false positive assessment: (1) annotation-based strategy, and (2) decoy database searching
- Draw the distinction between cumulative false positive rate and instantaneous false positive rate

INTRODUCTION

Database search engines provide an automated, high-throughput way to analyze the large amounts of data generated by many mass spectrometry experiments. Properly interpreting the results, however, requires a good understanding of the prevalence and disposition of false positives in the results. Here, we discuss and compare some methodologies for assessing false positives.

MATERIALS AND METHODS

A simple protein mixture was created by mixing together proteins (nominally 20 in number) purchased from Sigma. The resulting sample was denatured, reduced with dithiothreitol, alkylated with iodoacetamide, and digested with trypsin. Using a Tempo™ nanoLC system (Applied Biosystems/MDS Sciex), the resulting peptides were separated using a 75 µm ID PepMap column (Dionex) and analyzed via electrospray ionization with a QSTAR® Elite Hybrid LC/MS/MS System (Applied Biosystems/MDS Sciex) using Information Dependent Acquisition (IDA). A total of 3 runs were performed, each with a 90 minute LC gradient at 300 nL/min.

The MS/MS data were processed with ProteinPilot™ software (http://download.appliedbiosystems.com/proteinpilot) using two different database search algorithms – Mascot® and Paragon™. All database searches were performed against the April 12, 2005 version of Swiss-Prot (168,405 protein entries). The Paragon algorithm search was performed with "rapid" search effort with no special factors or ID focus. All other parameters for the Paragon algorithm search were set exactly as described in the sample description above. The search parameters for Mascot were chosen to mimic the Paragon algorithm search as much as possible – fixed modifications: carboxamidomethyl on C; variable modifications: deamidation on N, Q and oxidation on M; mass values: monoisotopic; peptide mass tolerance: ± 0.2 Da; fragment mass tolerance: ± 0.2 Da; max missed cleavages: 1; instrument type: ESI-QUAD-TOF. Mascot results were analyzed under two different report types – the default report, and the report obtained by using "require bold red," which is a way to remove duplicate homologous proteins from a report.

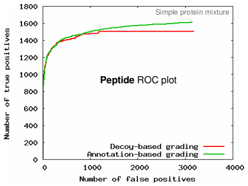
To illustrate the difference between instantaneous and cumulative false positive rates, we use a sample with a large number of proteins in order to get good statistics at the protein level. A brief description of the experiment is provided here; for further details, see reference 1. Lung tumors resulting from metastasis of Lewis lung cancer cells from mice were harvested and subcultured. Four conditions were analyzed by adjusting two variables – control cells vs. cells transfected to overexpress the receptor tyrosine kinase ErbB2 (also known as Her2/Neu); cells cultured in the presence and in the absence of fibronectin. After cell lysis, proteins were isolated (100 µg), digested with trypsin, and labeled with iTRAQ™ reagents according to the Applied Biosystems protocol. The sample was separated into 40 fractions by strong cation exchange, and each fraction was analyzed by LC-MS/MS using a C18 column (75µm x 15cm, LC Packings; 5-30% acetonitrile over 30 min) on a Tempo LC System coupled to a QSTAR Elite system. The MS/MS data were processed with the Paragon algorithm ("rapid" search effort, no special factors or ID focus, all other parameters set as described in the sample description above).

GRADING STRATEGIES – DECOY VS. ANNOTATION

DECOY database searching involves the creation of decoy proteins constructed by reversing or shuffling all the proteins in the search database. False positives are assumed to be equally likely in the decoy and non-decoy (target) database portions. A number of decoy methodologies have been proposed. Here, we search a composite database consisting of two sections – (1) the target section, consisting of all the normal protein sequences (from, for example, Swiss-Prot), and (2) the decoy section, consisting of the reverse of all the sequences of the first section.[2]

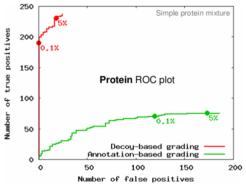
The **ANNOTATION**-based strategy involves the use of a "known" sample together with its associated annotation consisting of a list of all proteins in the sample. Database search results are evaluated by comparison to the annotation. A challenging aspect of this approach is the difficulty in obtaining a reliable annotation. Even "known" samples, such as the protein mixture studied here, generally contain contaminant proteins. Our protein mixture has been exhaustively characterized through repeated acquisition on a variety of instruments. Care has been taken to minimize redundancy in the list of proteins in our annotation in that multiple isoforms of a protein are included only if there is good distinct evidence for each isoform.

Figure 1. No significant difference in decoy vs. annotation approaches for assessing false positive peptides (Mascot search results – default report)



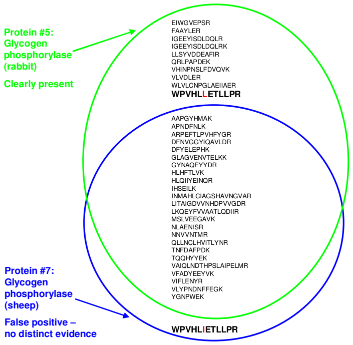
Peptide counting ROC plot of the Mascot search results (default report).

Figure 2. Large discrepancy in decoy vs. annotation approaches for assessing false positive proteins (Mascot search results – default report)



Protein counting ROC plot of the Mascot search results (default report). While the ROC plot for decoy-based grading (red curve) looks much better, it is actually **wrong**, as will be explained in figures 3 and 4. As points of reference, the 0.1% and 5% nominal error rates declared by Mascot are marked on the graph.

Figure 3. Protein #7 – true positive or false positive?



Protein #5: Glycogen phosphorylase (rabbit)

Clearly present

EIWGVEPSR
FAAYLER
IGEEYISDLDQLR
IGEEYISDLDQLRK
LLSYVDDEAFIR
QRLPAPDEK
VHINPNSLFDVQVK
VLVDLER
WLVLCNPGLAEIIAER
WPVHLLETLLPR

AAPGYHMAK
APNDFNLK
ARPEFTLPVHFYGR
DFNVGGYIQAVLDR
DFYELEPHK
GLAGVENVTELKK
GYNAQEYYDR
HLHFTLVK
HLQIIYEINQR
IHSEILK
INMAHLCIAGSHAVNGVAR
LITAIGDVVNHDPVVGDR
LKQEYFVVAATLQDIIR
MSLVEEGAVK
NLAENISR
NNVVNTMR
QLLNCLHVITLYNR
TNFDAFPDK
TQQHYYEK
VAIQLNDTHPSLAIPELMR
VFADYEEYVK
VIFLENYR
VLYPNDNFFEGK
YGNPWEK

Protein #7: Glycogen phosphorylase (sheep)

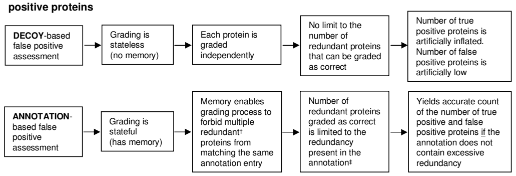
False positive – no distinct evidence

WPVHLIETLLPR

Protein #7 (the 7th highest-scoring protein in the Mascot search results list) is the highest-scoring protein where the two grading strategies differ in their assessment. **This protein is deemed a true positive in the decoy approach and is deemed a false positive in the annotation approach.**

The Venn diagram shows that the only peptide of protein #7 that is unaccounted for by protein #5 is WPVHLIETLLPR with the only difference being an L->I substitution, which is generally indistinguishable by mass spectrometry. Hence, there is no distinct evidence for protein #7. In addition, glycogen phosphorylase (rabbit) is known to be in the sample by construction, but there is no reason to suspect that glycogen phosphorylase (sheep) is in the sample. All in all, there is absolutely no reason to believe that protein #7, glycogen phosphorylase (sheep), has actually been detected in our experiment, and, therefore, **protein #7 is a false positive**.

Figure 4. Explanation for the discrepancy in annotation vs. decoy approaches for assessing false positive proteins



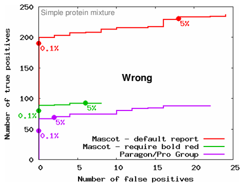
DECOY-based false positive assessment → Grading is stateless (no memory) → Each protein is graded independently → No limit to the number of redundant proteins that can be graded as correct → Number of true positive proteins is artificially inflated. Number of false positive proteins is artificially low

ANNOTATION-based false positive assessment → Grading is stateful (has memory) → Memory enables grading process to forbid multiple redundant† proteins from matching the same annotation entry → Number of redundant proteins graded as correct is limited to the redundancy present in the annotation‡ → Yields accurate count of the number of true positive and false positive proteins if the annotation does not contain excessive redundancy

† Here, we define "redundant" proteins as proteins that have significant amounts of (sub)sequence identity due to their evolutionary relationship. For example, in the previous figure, glycogen phosphorylase (rabbit) and glycogen phosphorylase (sheep) are considered redundant. Strictly speaking, "near-redundant" or "near-redundant due to homology" may be more accurate terms.

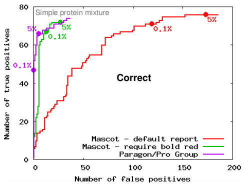
‡ The degree of redundancy in the annotation should be reflective of the sample. Multiple protein isoforms can often be detected in a sample, and that should translate to multiple annotation entries. In creating an annotation, multiple isoforms should be declared only if there is good distinct evidence for every isoform.

Figure 5. WRONG: Proteins lacking good distinct evidence are mistakenly graded as true positives in the decoy-based approach



The decoy-based approach for false positive assessment rewards software that does a poor job of recognizing and eliminating redundant proteins. In the decoy-based approach, the more redundant proteins in the results list, the higher the apparent number of true positives.

Figure 6. CORRECT: Proteins lacking good distinct evidence are properly graded as false positives in the annotation-based approach



The annotation-based approach provides a much more accurate assessment of true positives and false positives when faced with redundant proteins in the results list. Using the correct methodology reveals that **better protein grouping actually yields fewer false positive proteins** (in direct contradiction to the erroneous plots of Figure 5).

The default report generated by Mascot contains a lot of false positive proteins due to the presence of too many homologous proteins in the report. Using the "require bold red" option in the Mascot report greatly reduces the number of false positive proteins reported. Paragon uses the Pro Group™ algorithm, which requires good distinct evidence for every isoform reported. For the Paragon algorithm, the first false positive protein in the results list is protein #53. By contrast, for Mascot, false positive proteins are found even amongst the highest-ranking proteins, although the false positives are much fewer in number overall when the "require bold red" option is used.

INSTANTANEOUS VS. CUMULATIVE FALSE POSITIVE RATE

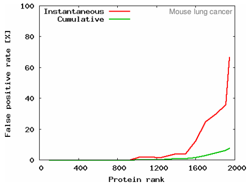


Figure 7. Cumulative false positive rate thresholds usually allow too many false positive proteins

For any protein, one can report the false positive rate at the protein's rank in the results list (instantaneous false positive rate), or one can report the false positive rate at or above the protein's rank in the results list (cumulative false positive rate). This figure shows that there can be a dramatic difference between the two numbers. For example, a cumulative false positive rate of 5% allows a protein rank cutoff of ~1800, but at rank 1800 the instantaneous false positive rate is about 30% – that is, about one in every three proteins is actually wrong!

In evaluating search results, people often calculate the cumulative false positive rate. By contrast, the confidences/probabilities reported by the Paragon and Mascot algorithms (and probably other search engines as well) are designed to be estimates of the instantaneous false positive rate. The instantaneous false positive rate is independent of the number of higher ranking proteins and is generally a more relevant and more robust metric.

CONCLUSIONS

- The annotation-based approach and the decoy-based approach are effective and comparable for assessing false positive **peptides**.
- The decoy database searching approach is a stateless grading system and will invariably be fooled by protein homology. This approach **cannot** be used to count the number of true positive and false positive **proteins** unless care is taken to account for protein homology. Otherwise, the numbers obtained are **completely meaningless**. (Note that if protein homology is properly accounted for, the decoy database searching approach works OK.)
- The annotation-based approach is a stateful grading system and is thus capable of giving accurate counts of the numbers of true positive and false positive **proteins**, although obtaining a reliable annotation is hard work. In creating an annotation, care must be taken to avoid excessive protein redundancy – to declare multiple isoforms, there must be good distinct evidence for every isoform.
- The difference between cumulative and instantaneous false positive rates can be large. The Paragon and Mascot algorithms estimate the instantaneous false positive rate.

Trial version of ProteinPilot software available at http://download.appliedbiosystems.com/proteinpilot

REFERENCES

1. Hunter, C. L.; Chisholm, K.; Xu, Y.; Alaoui-Jamali, M.; Pinto, D. *54th ASMS Conference on Mass Spectrometry*, Seattle, WA (May 28 - June 1, 2006), Poster ThP33.
2. Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. *J. Proteome Res.* **2003**, *2*, 43-50.

ACKNOWLEDGEMENTS

We thank Christie Hunter for providing the MS/MS data discussed here.

TRADEMARKS/LICENSING

For research use only. Not for use in diagnostic procedures.