Fast and accurate protein false discovery rates on human proteome study scale with Percolator 3.0

Matthew The,
Science for Life Laboratory,
School of Biotechnology,
Royal Institute of Technology - KTH,
Box 1031, 17121 Solna,
Sweden

Michael J. MacCoss,
Department of Genome Sciences,
School of Medicine,
University of Washington,
Seattle, Washington 98195,
United States of America

William S. Noble,
Department of Genome Sciences,
School of Medicine,
University of Washington,
Seattle, Washington 98195,
United States of America

Lukas Käll
Science for Life Laboratory,
School of Biotechnology,
Royal Institute of Technology - KTH
Box 1031, 17121 Solna,
Sweden

March 16, 2016

Keywords: mass spectrometry - LC-MS/MS, statistical analysis, data processing and analysis, protein inference, simulation

Abstract

Percolator is a popular tool to assign reliable statistics, such as q-values and posterior error probabilities, to peptides and peptide spectrum matches (PSMs) using search results from mass spectrometry-based proteomics experiments. Percolator's processing speed has been sufficient for typical data sets with hundreds of thousands of PSMs. With our new scalable approach, we can now also analyze millions of PSMs in a matter of minutes on a commodity computer. Furthermore, with the increasing awareness for the need for reliable statistics on the protein level, we compared several easy-to-understand protein inference methods and implemented the best method in the Percolator package. We used Percolator 3.0 to analyze the data from a recent study of the draft human proteome containing 20 million spectra (PM:24870542).

Introduction

Percolator [6] has played a prominent part in the analysis pipelines of shotgun proteomics experiments for the last decade, by post-processing the results from database search engines such as SEQUEST [3], MASCOT [1], X!Tandem [2] and MS-GF+ [9]. Not only does Percolator often give a significant boost in the number of significant peptide spectrum matches (PSMs) or peptides, it also provides a consistent statistical framework in which to interpret the search results. As Percolator's running time is usually much lower than that of the search engine, applying it as a post-processing step should be a no-brainer. As part of the continuous development and support of the Percolator package, we present two major additions aimed at supporting analysis of studies on the scale of the human proteome [8, 14].

As advances in technology are causing shotgun proteomic experiments to become progressively easy and affordable to carry out, the amount of data per study will keep rising steadily. While previous versions of Percolator are able to process the data from the vast majority of current studies in a decent time frame, certain limitations have come into sight for laboratories without access to an above average commodity computer. When processing millions of PSMs, the majority of Percolators processing time is spent on training support vector machines (SVMs). One could, however, surmise that the performance of the SVM would plateau pretty fast as a function of the number of input PSMs. Here, we propose to use Percolators semi-supervised learning algorithm to train SVMs on only a random subset of the PSMs and used the resulting score vectors to evaluate the rest of the PSMs.

Second, protein-level accuracy estimates have been on our feature wish list for quite some time now. One of the major obstacles was the question of how to deal with shared peptides and protein grouping. An implementation of Fido [13] has been part of the Percolator package for quite some time now and addressed these two issues, but is too computationally intensive on large-scale datasets with many shared peptides. On the other hand, large-scale studies have a deep coverage of the present peptides and therefore identify many peptides that uniquely identify a protein. This provides the option of ignoring shared peptides altogether and makes the task of protein inference much simpler and intuitive. Here, we compared several easy-to-understand protein inference strategies that only use these peptides that uniquely identify a protein and implemented the best candidate in the Percolator package.

Methods

We downloaded a set of spectra, comprising 2212 runs on 17 adult tissues, 7 fetal tissues, and 6 hematopoietic cell types with a total of 21 million spectra from [8]. The investigated peptides were analyzed on an LTQ Orbitrap Velos and Elite (Thermo Scientific) equipped with an Easy-nLC II nanoflow LC System (Waters). We will refer to this set as the *pandey* set.

Converting the RAW files to MS1 and MS2 files was done with Proteowizard [7]. Next, we assigned

high-resolution precursor masses and charges using information from the precursor scans with Hardklör [4] followed by Bullseye [5], through the Crux 2.0 package interface [10]. This was followed by a database search against the human Swissprot and Swissprot+Ensembl databases (accessed: 2015 Nov 12) using the Tide search engine, again through the Crux interface. The same search parameters as in [8] were used, with the exception of not using cyclization of N-terminal glutamine as a variable modification, using semi-tryptic searches and using Tide's default fragment tolerance. For the decoy proteins, we reversed the target protein sequences. Separate searches were done on the target and decoy protein database, resulting in a total of 73 million target and decoy PSMs.

The normal Percolator's semi-supervised learning algorithm randomly splits the training set in 3 folds and computes 3 scoring vectors, each trained on 2 of the 3 folds and tested on the remaining fold. The final scores are then calculated using the scoring vector where the PSM was in the test set. To implement subset scoring, we simply applied the normal training algorithm on a random subset of the PSMs, resulting in 3 scoring vectors. However, instead of scoring using only a single scoring vector, we now calculate each PSM's score as the average of the scores from the 3 scoring vectors.

To characterize the behavior of the scoring vectors based on subsets of the PSMs, we evaluated the performance for different sizes of the random subset. Preliminary results showed that including target and decoy PSMs belonging to the same spectrum together during the selection of random subsets gave a more stable performance than sampling without taking this into account. Therefore this strategy was applied in the random sampling process. For each random subset size, we calculated the mean and standard deviation over 10 randomized runs of the number of PSMs and peptides with q value below 0.001.

Before applying the protein inference methods, we used the approach to handling shared peptides from *Nesvizhskii et al.* [11]. Here, proteins are grouped that are indistinguishable based on their theoretical proteolytically digested peptides, rather than their experimentally discovered peptides. We retain the peptides that are unique to such a protein group, rather than to a single protein. Especially for databases containing many proteoforms for a gene, this can decrease the number of shared peptides considerably.

We compared several straightforward protein inference methods: Fishers method for p-value combination, the picked target-decoy strategy [12], and the product of peptide-level posterior error probabilities (PEPs). We assessed the performance of the methods on the draft human proteome data, as well as on a standard protein mix. Using the standard protein mix we could characterize the accuracy and stability of the false discovery rate estimates, and the draft human proteome data could give us an indication of the performance on real data. We calculated false discovery rate (FDR) estimates from p-values for Fishers method, and using target-decoy models for all other methods.

Results

Searching all 73 million target+decoy PSMs resulted in 300 000 unique target peptides. Using subsets of even just 100 000 PSMs (0.1%) for SVM training did not reduce the number of identified peptides and generally even slightly increased this number. The standard deviation of the randomized runs for a fixed subset size did increase when taking increasingly smaller subsets, but this effect was limited. By using a subset of 500 000 PSMs to train the SVM, Percolators runtime was reduced from several hours to under 10 minutes.

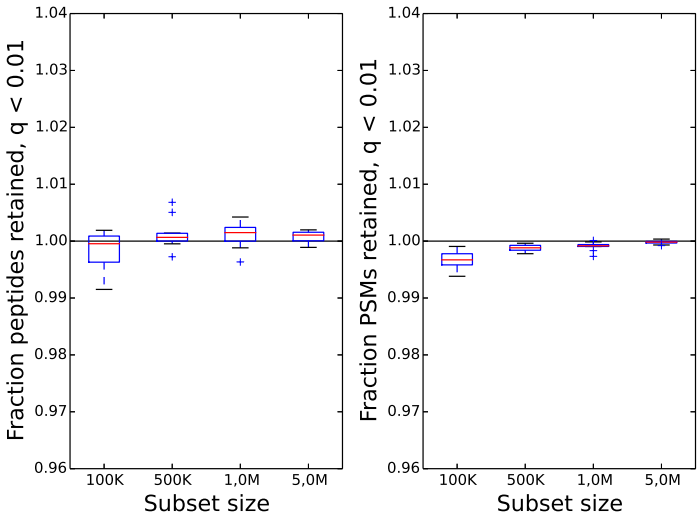

Figure 1: **Using subsets of the 73 million target+decoy PSMs for SVM training retains the number of significant PSMs and peptides as training using the full set.** Training the SVM using all 73 million PSMs resulted in 7 928 551 PSMs and 298 095 unique peptides at a q value threshold of 0.01. We used 10 random subsets each of 100 000, 500 000, 1 000 000 and 5 000 000 PSMs to train the SVMs and scored all 73 million PSMs using the resulting support vectors. The number of significant PSMs and unique peptides does not drop significantly for even subsets of 100 000 PSMs.

We assessed the accuracy and stability of FDR estimates using three standard protein mixes. While Fishers method gave the most robust FDR estimates, it identified $5-10\%$ fewer proteins than the picked target-decoy strategy and product of PEPs. These two strategies did give reasonably accurate FDR estimates but were more sensitive to errors in the peptide identification.

For the draft human proteome set, at 1% FDR, the picked target-decoy strategy identified 12 300 proteins, multiplication of PEPs 11 600 and Fishers method only 8 800. From these results we concluded that the picked target-decoy strategy was the superior alternative and implemented this in the newest Percolator package.

Discussion

Acknowledgements

References

[1] John S Cottrell and U London. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.

[2] Robertson Craig and Ronald C Beavis. Tandem: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, 2004.

[3] Jimmy K Eng, Ashley L McCormack, and John R Yates. 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.

[4] Michael R Hoopmann, Gregory L Finney, and Michael J MacCoss. High-speed data reduction, feature detection, and ms/ms spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. *Anal. Chem.*, 79(15):5620–5632, 2007.

[5] Edward J Hsieh, Michael R Hoopmann, Brendan MacLean, and Michael J MacCoss. Comparison of database search strategies for high precursor mass accuracy ms/ms data. *J. Proteome Res.*, 9(2):1138–1143, 2009.

[6] Lukas Käll, Jesse D Canterbury, Jason Weston, William Stafford Noble, and Michael J MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods*, 4(11):923–925, 2007.

[7] Darren Kessner, Matt Chambers, Robert Burke, David Agus, and Parag Mallick. Proteowizard: open source software for rapid proteomics tools development. *Bioinformatics*, 24(21):2534–2536, 2008.

[8] Min-Sik Kim, Sneha M Pinto, Derese Getnet, Raja Sekhar Nirujogi, Srikanth S Manda, Raghothama Chaerkady, Anil K Madugundu, Dhanashree S Kelkar, Ruth Isserlin, Shobhit Jain, et al. A draft map of the human proteome. *Nature*, 509(7502):575–581, 2014.

[9] Sangtae Kim, Nitin Gupta, and Pavel A Pevzner. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.*, 7(8):3354–3363, 2008.

[10] Sean McIlwain, Kaipo Tamura, Attila Kertesz-Farkas, Charles E Grant, Benjamin Diament, Barbara Frewen, J Jeffry Howbert, Michael R Hoopmann, Lukas Käll, Jimmy K Eng, Michael J MacCoss,

and William S Noble. Crux: rapid open source protein tandem mass spectrometry analysis. *J. Proteome Res.*, 13(10):4488–4491, 2014.

[11] Alexey I Nesvizhskii, Andrew Keller, Eugene Kolker, and Ruedi Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Analytical chemistry*, 75(17):4646–4658, 2003.

[12] Mikhail M Savitski, Mathias Wilhelm, Hannes Hahne, Bernhard Kuster, and Marcus Bantscheff. A scalable approach for protein false discovery rate estimation in large proteomic data sets. *Molecular & Cellular Proteomics*, pages mcp–M114, 2015.

[13] Oliver Serang, Michael J MacCoss, and William Stafford Noble. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *Journal of proteome research*, 9(10):5346–5357, 2010.

[14] Mathias Wilhelm, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M Savitski, Emanuel Ziegler, Lars Butzmann, Siegfried Gessulat, Harald Marx, et al. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–587, 2014.