Assessment of an amalgamative approach to protein identification

Iain Rogers

Bioinformatics Solutions Inc.

Introduction

When studying proteins using mass spectrometry, researchers can identify which proteins are in a sample by matching measured masses to the calculated masses of peptides and sequence tags in a protein sequence database. Because of large databases and experimental data sets, this process is necessarily automated using protein identification software. However, because of instrumental and experimental limitations, analysis is made difficult by noise, contamination and inconclusive data.

The problem becomes one of validation. The researcher must accept the software's suggestion and scoring scheme, or spend countless hours manually validating the results. Conclusions based on imperfect data, processed by imperfect software and inferred from non-validated results will always be suspect.

Overview

The following shows how two or more protein identification tools used in chorus, each confirming the results of the others, can improve quality of and confidence in results. A large amount of MS/MS data is run through several protein identification programs. Consensus is tabulated, and the quality of consensus results is quantified. The value of an automated results validation system is demonstrated. Each protein identification program is assessed, both on an individual level, and in terms of contribution to consensus results.

Methods

Five protein identification programs were used, representing a variety of approaches to MS/MS protein identification. OMSSA[4], X!Tandem[2] and SEQUEST compare MS/MS fragment masses directly to masses calculated from a sequence database. SPIDER[5] uses de novo sequences to search a sequence database, while allowing for errors in de novo sequencing. PEAKS[6] uses a combination of sequence searching and fragment mass matching. Each program's default data processing parameters was used, along with standard error tolerance values. PEAKS de novo was employed to generate de novo sequences for both SPIDER and PEAKS protein identification.

Two data-sets made up the material for this analysis:
- Keller et al's[1] benchmark data set consists of 22 separate runs (totaling 37044 spectra) of 18 standard proteins with an "LCQ" ion-trap mass spectrometer. SEQUEST results were available for Keller et Al's benchmark, and these results were used in the analysis.
- 17mix_test2[3] consists of one run (totaling 1389 spectra) of 17 standard proteins through a "QTOF Ultima" quadrupole-time-of-flight mass spectrometer. OMSSA was substituted for SEQUEST when analyzing 17mix_test2.

The resulting peptide matches and scores were tabulated, with one row representing one spectrum and containing proposed peptides from all the programs. A simple Visual Basic script was written to look for consensus and correctness on each row/spectrum. Consensus was defined as agreement between two or more programs on a proposed peptide. Confidence scores as provided by each program were disregarded unless to clear up a conflict in consensus. An exact sequence match, between the proposed peptide and a protein known to be in the sample, determined correctness.

Special Thanks to:
Weiming Zhang, Bin Ma, Virginia Yang, Gilles Lajoie.

Copyright notice: Bioinformatics Solutions Inc. holds the exclusive right to copy or distribute this work. This work, and any images, text, graphs and figures contained herein, may not be reproduced without permission from either the author or Bioinformatics Solutions Inc.

Consensus Results: LCQ

Figure 1 shows the amount and quality of different types of consensus (2-way, 3-way and 4-way consensus) between PEAKS, SEQUEST, SPIDER and X!Tandem on the LCQ data. Figure 3 summarizes the amount and quality of each program's contribution to the consensus results. Notably, 2985 peptides were determined by consensus between programs. Of these 2893 (97%) were correct. Percentage correctness was high and fairly uniform where two of SEQUEST, PEAKS or X!Tandem were involved. 3-way consensus between X!Tandem, SEQUEST and PEAKS made up the bulk of the consensus results. X!Tandem was, marginally, the largest individual contributor to consensus results. SPIDER contributed only a small portion, and this may be because of the lower quality of ion trap data, and consequently less precise de novo sequences. Evaluation of consensus and correctness on all 37044 spectra took a total of ~8 minutes.

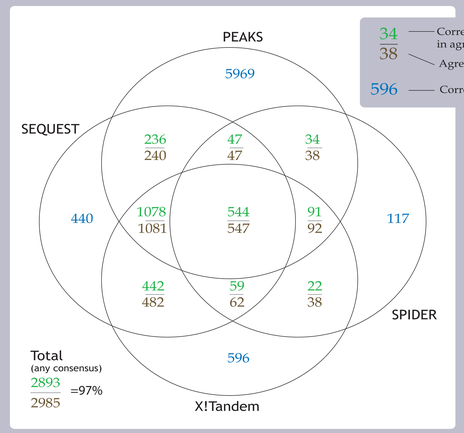

Figure 1: Consensus results on LCQ data

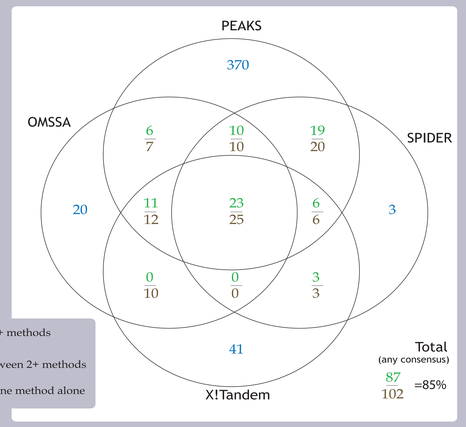

Figure 2: Consensus results on QTOF data

Consensus Results: QTOF

Figure 2 shows the amount and quality of different types of consensus (2-way, 3-way and 4-way consensus) between PEAKS, OMSSA, SPIDER and X!Tandem on the QTOF data. Figure 4 summarizes the amount and quality of each program's contribution to the consensus results. 85% of the 102 peptides reached by consensus among programs were correct. 4-way consensus and 2-way SPIDER-PEAKS consensus made up the bulk of the consensus results. PEAKS was the largest individual contributor to number of consensus results, and X!Tandem the lowest. Peptides reached by consensus where PEAKS was involved were correct 95.6% of the time. Peptides reached by consensus where PEAKS and SPIDER were involved were correct 96% of the time. Evaluation of consensus and correctness on all 1389 spectra took a total of ~20 seconds.

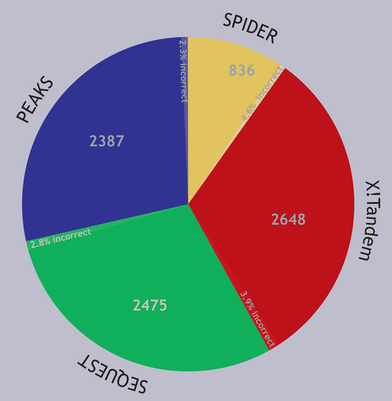

Figure 3: On LCQ data, the number of consensus results each program contributed to (and of those, how many were incorrect).

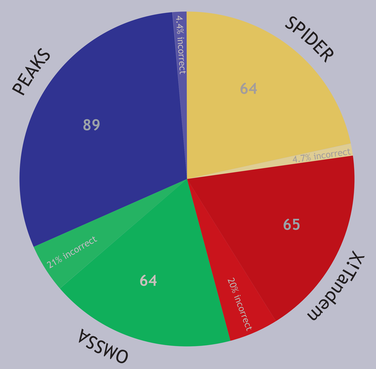

Figure 4: On QTOF data, the number of consensus results each program contributed to (and of those, how many were incorrect).

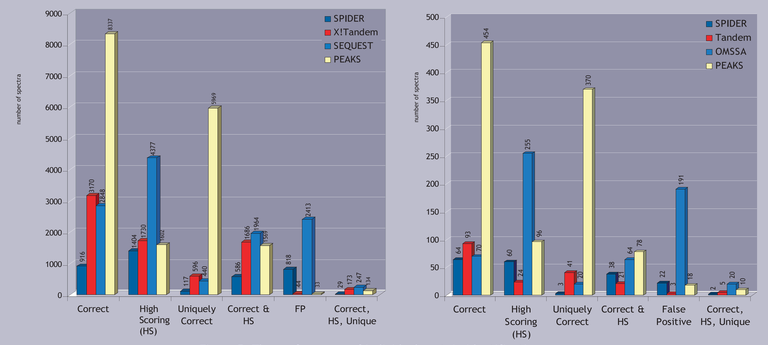

Figure 5: The performance of individual protein identification programs on:
a) The Keller et al LCQ data-set and b) the 17mix_text2 QTOF data set.

Individual program performance

PEAKS demonstrated a remarkable ability to match peptides to the correct proteins where no other program could. Further investigation reveals that these are mostly low scoring matches. Interestingly, SEQUEST and OMSSA showed similar performance, returning fewer uniquely correct results, and an unacceptable level of false positives.

X!Tandem performed well, returning very few false positive results. Comparing the proportions of "correct, high-scoring unique results" to "correct, high-scoring resutls" X!Tandem performed just as well as PEAKS. This demonstrates both programs' ability to find high quality, useful results, where no other software could.

Conclusions

The high percentage of correctness among results obtained by consensus between two or more protein identification programs speaks clearly for the advantage of using many methods in chorus. Automated comparison, even using a script as inefficient as the one used for this analysis, is far quicker than painstaking manual cross-referencing.

Gains in coverage of a protein, by matching more peptides, is another benefit to using more than one protein identification program. Coverage can be gained by considering peptides on which two separate programs agreed, but assigned very low scores. Coverage can also be gained by considering peptides that only one program could identify.

Confidence scores provided by individual programs, while useful for result evaluation in a some cases, can be extremely misleading in others. Agreement between two protein identification programs may provide more definitive answers. A complex scoring algorithm to evaluate the strength of a consensus is not required.

The choice of tools to use in chorus depends on the data to be analyzed. Since SEQUEST was built and trained on the sort of LCQ data used by this study, it is not surprising that it contributed well to the consensus results. On QTOF data, the high percentages of consensus correctness where SPIDER or PEAKS is used demonstrates an advantage of using de novo sequences to aid in protein identification even when the protein is known.

References

1. Keller, A., Purvine S., Nesvizhskii, A.I., Stolyar, S., Goodlett, D.R., and Kolker, E., Experimental Protein Mixture for Validating Tandem Mass Spectra Analysis, (OMICS 6(2), 207-212, 2002).
2. Craig, R., Beavis, R. C., TANDEM: matching proteins with mass spectra, (Bioinformatics, 20, 1466-7, 2004).
3. Data repository publicly available at sashimi.sourceforge.net/repository.html
4. Geer, L.Y., Markey, S.P., Kowalak, J.A., Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W., Bryant SH., Open mass spectrometry search algorithm, (J Proteome Res. Sep-Oct;3(5):958-64, 2004).
5. Han, Y., Ma, B., Zhang, K., Software Protein Identifier, (http://bif.csd.uwo.ca/spider/)
6. PEAKS software demo available by request at www.bioinformaticssolutions.com