Drastically increased coverage by using four search engines for Protein Identification

Iain Rogers[1], Will Haskins[2]
[1]Bioinformatics Solutions Inc., Waterloo ON., [2]Genentech Inc., S. San Francisco, CA.

Bioinformatics Solutions Inc.

Introduction

Several algorithmic approaches have been devised to aid scientists in the identification of proteins and peptides from tandem mass spectrometric data. Most involve comparing the masses of peptide fragments to theoretical masses, calculated from protein sequence databases. Because of poor quality data, unforeseen post-translational modifications, and sequence variations, these database search engines are unable to confidently all spectra. A search engine may be able to identify only 5% of spectra in a sample. Because of the possibility of false positive assignments, even these peptide assignments must be verified.

This research proposes that two or more search engines, when used together, can not only provide suitable automatic validation for peptide assignments, but can double the number of confident peptide assignments.

Overview

The following shows how four protein identification tools used in chorus, each confirming the results of the others, can substantially improve the number of spectra to which a peptide sequence can confidently assigned. 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. Each protein identification program is assessed individually and in terms of its contribution to consensus results.

Methods

Four protein identification programs were used, representing a variety of approaches to MS/MS protein identification. MASCOT, X!Tandem[2] and SEQUEST compare MS/MS fragment masses directly to masses calculated from a sequence database. PEAKS[6] uses a combination of sequence tag 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 the sequence tag search portion of PEAKS protein identification.

- Keller et al's[1] benchmark data set was used, consisting 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.

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.

Consensus Results

Figure 1 summarizes each programs ability to assign peptide sequences to spectra by database search, as compared to the consensus results. By using a consensus approach, rather than an individual search engine, we can increase the number of spectra that are confidently explained by at least 50%. This increase is solely gained by using agreement between two search engines as the only measure of confidence. In this way low scoring (but nevertheless correct) matches returned by two search engines are given high confidence. These low scoring matches would otherwise have been rejected by any of the single search engines. Furthermore, each search engine returns a small number of unique, high-scoring and correct matches, that we can add to the consensus results to further improve coverage. Concurrence between X!Tandem and Mascot had a high rate of false positives, as such, consensus results where only Mascot and X!Tandem agreed were rejected from analysis.

Figure 2 summarizes the amount and quality of each program's contribution to the consensus results. Notably, 3079 peptides were determined by consensus between programs. Of these 2981 (97%) were correct. Percentage correctness was high and fairly uniform where two of SEQUEST, PEAKS or X!Tandem were involved. 4-way consensus made up the bulk of the consensus results. X!Tandem was, marginally, the largest individual contributor to consensus results, and consensus results involving PEAKS had the lowest incidence of incorrectness. Evaluation of consensus and correctness on all 37044 spectra took a total of ~8 minutes.

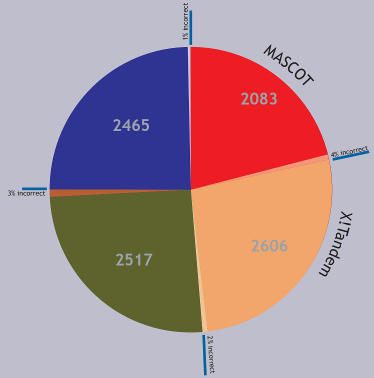

Figure 2: the number of consensus results each program contributed to (and of those, how many were incorrect).

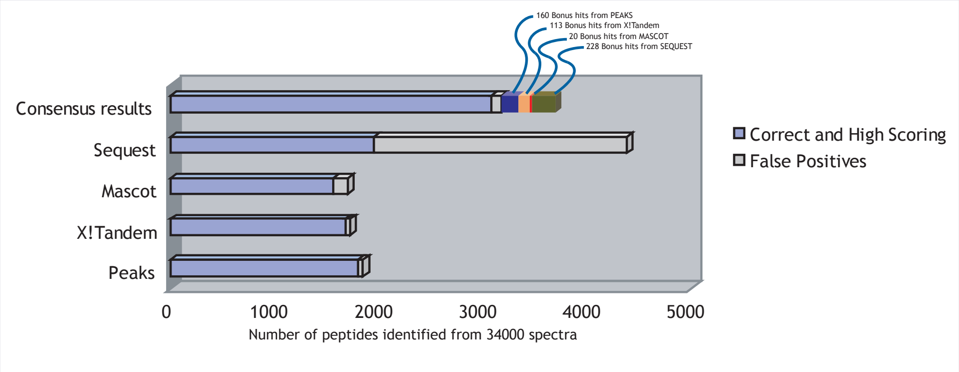

Figure 1: Comparative performance of PEAKS, MASCOT, X!Tandem and SEQUEST and an amalgamative approach involving all four search engines

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 search tools together. 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 benefit of this approach is improved sensitivity in identification of peptides from MS/MS data, without sacrificing accuracy.

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).
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).
6. PEAKS software demo available by request at www.bioinformaticssolutions.com