Bioinformatics Solutions Inc.

Cross-Species Search for Accurate and Sensitive Peptide Identification

University of Waterloo

Wei Wu Chen[1], Denis Yuen[1], Dan Maloney[1], Bin Ma [2]
Bioinformatics Solutions Inc, Waterloo, ON [1]
University of Waterloo, Waterloo, ON [2]

Introduction

Database search has proven to be useful for identifying proteins from well characterised organisms [2]; however, it does not deliver the best results when studying proteomes from species that are not fully documented. The combination of de novo sequencing and homology searching has proven to be very useful for the identification of proteins from unsequenced organisms [1,3]; in which case a database from a homologous species must be used as a reference. Unfortunately, conventional homology search tools, such as BLAST, does not model the possible de novo sequencing errors correctly, thus specialized software such as SPIDER is needed to ensure the search sensitivity [4]. Furthermore, no result validation method for cross-species search exists. We propose a method for accurate and sensitive cross-species peptide identification that allows result validation based on false discovery rate (FDR).

Methods

The existing cross-species identification methods first perform de novo sequencing on the MS/MS data, and then utilize a special homology search tool, such as MS-BLAST and SPIDER, to find sequence similarities from the homologous protein database. Our method differs from existing methods in four aspects:

- PEAKS DB is performed after de novo sequencing to identify peptides that can be found without the use of homology search.
- The homology search is only performed on a short list of proteins that are found by sequence tag matching. This increases the search speed.
- The homology search result is re-evaluated by comparing with the original MS/MS spectrum. This allows a more accurate scoring function and better separates the true and false identifications.
- A decoy-fusion method is used for FDR estimation.

Results

The method described above is implemented in PEAKS 6. To illustrate its ability to identify peptides from a homologous proteome instead of the target proteome a human data set was used to search a mouse proteome. A publicly available tryptic digest of HepG2 Homo sapiens hepatocellular carcinoma cell line LC-MS/MS dataset containing 104491 MS/MS spectra analysed with an LTQ-Orbitrap with CID fragmentation was downloaded. This data was analysed using PEAKS 6 automatic de novo sequencing, PEAKS DB, and SPIDER homology search. Two databases were searched: (a) The H. sapiens database contains 20250 proteins from the UniProtKB/Swiss-Prot Release 2012_4; and (b) the Mus musculus database contains 16528 proteins from UniProtKB/Swiss-Prot Release 2012_4. For both protein identification and homology search the decoy fusion method [3] was used to estimate the false discovery rate.

MASCOT, was able to identify 53417 PSMs from the H. sapiens database and 29468 PSMs from the M. musculus database.

Using PEAKS DB and the H. sapiens database, 85993 peptides spectrum matches (PSMs) were identified. 81162 of those were found to have a 1% or lower false discovery rate (FDR). Using the M. musculus database, 59767 PSMs were identified. 46658 of those were found to have a 1% or lower FDR. This shows that peptide identifications are lost when proteins from the species of interest are not included in the database.

Using SPIDER and the H. sapiens database, 3703 extra PSMs were matched. 3639 of those matches were found to be above the 1% FDR cut off. Using SPIDER and the M. musculus database, 12433 extra PSMs were identified that were not identified by PEAKS DB. 11394 of those were found to have a 1% or lower FDR.

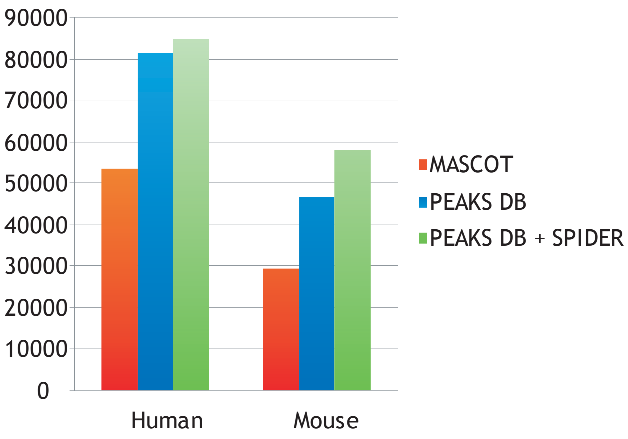

Figure 1. Identification results above 1% FDR of H. sapiens data set against H. sapiens database and H. sapiens data set against M. musculus database.

SPIDER considers point mutations when identifying PSMs. Even when point mutations are reported, care is taken to insure the precursor mass error tolerance cut off is followed.

Figure 2 shows the mass error plot of the PEAKS DB + SPIDER search of the M. musculus. The vertical line shows the 1% FDR cut off for that search. A 20 ppm mass error tolerance was set for each search. All the identifications made with PEAKS DB and SPIDER above 1% FDR were well within 20 ppm of the precursor ion mass.

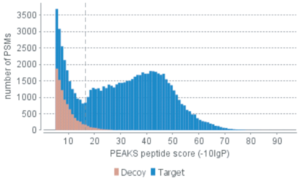

Figure 2. Mass Error Plot of H. Sapiens Data Set Against M. Musculus Database

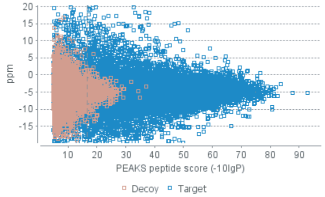

Figure 3. Score Distribution of H. Sapiens Data Set Against M. Musculus Database

Taking only unique peptide identifications into account, it is clear that the addition of SPIDER homology search to the work flow is essential when searching a homologous species database. When run against the H. sapiens database, 4.7% of the identifications were discovered by SPIDER. When run against the M. musculus database 20.8% of the identifications were discovered by SPIDER. A much higher percentage of the identifications are made by SPIDER when searching a homologous species database. This is because of SPIDER's ability to take into account point mutations between the de novo sequencing result and the database entry.

Conclusion

SPIDER homology search is able to add extra confident identifications to database search results. It is most useful when a cross species search approach must be used

References

1. Han, Y., Ma, B., and Zhang, K., SPIDER: Software for Protein Identification from Sequence Tags Containing De Novo Sequencing Error (Journal of Bioinformatics and Computational Biology, 3(3):697-716. 2005).
2. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., Lajoie, G., PEAKS: Powerful Software for Peptide Denovo Sequencing by MS/MS. (Rapid Communications in Mass Spectrometry, 17(20):2337-2342, 2002).
3. Zhang J. et al. PEAKS DB: De Novo sequencing assisted database search for sensitive and accurate peptide identification, MCP.M111.010587 (2012).
4. Ma, B., Johnson, R., De Novo Sequencing and Homology Searching. MCP.O111.014902 (2011).