Improving protein coverage by de novo sequence homology searching with SPIDER

Weijie Yang, Denis Yuen, Bin Ma, Iain Rogers

Bioinformatics Solutions Inc.

Introduction

Database search of tandem-MS spectra has been a well used technique for protein identification. But several proteomics problems require more coverage and more scrutinous results than this technique can provide. Sequence homology searching based on peptide de novo sequences allow us to identify peptides that are not present in a database. This approach, when coupled with standard search techniques means we can better explain the data and improve coverage on the identified proteins. Alternatively, we can better explain peptides from organisms that are not present in any database[1].

In this work we build and evaluate a workflow involving PEAKS auto de novo sequencing[2] and SPIDER[3], a unique tool for peptide sequence tag based homology searching.

Data

A mixture of six standard proteins, digested with Trypsin, was analysed by LC-MS-MS using an LTQ Orbitrap mass spectrometer. Survey scans and MS/MS scans were recorded by Orbitrap with RP of 30000 and 7500 respectively. Fragmentation was completed in the C-trap for high energy CID. This method yielded the best data for de novo sequencing[4]. 638 MS/MS spectra were collected in total.

Approach

De novo sequencing was performed on all 638 spectra, using the PEAKS auto de novo tool inside PEAKS Studio 4.5.

The inChorus meta search protein ID tool was used to launch the "Sequest" (BioWorks 3.3.1) and "PEAKS Protein ID" (PEAKS 4.5) algorithms to identify the proteins. The inChorus tool checks for consensus between two search engines and extra hits, leading to more coverage and more confidence. For this sample, an average of 10% increase in protein sequence coverage is attained by combining the two search engines (see figure 1). This initial database search was able to identify the six proteins known to be in the sample, thereby explaining 220 of the spectra.

The initial inChorus search did not consider post-translational modifications. We wanted to quantify how many missed hits could be explained by post translational modifications. So a second pass search was conducted using the PEAKS Protein ID search engine, which is very flexible and robust with regards to setting post-translational modifications. The database used was constructed from the sequences of the proteins known to be in the sample, plus the reverse of their sequences. A score threshold was set, for the returned peptides, to exclude any hits to the reverse portions of the database. As such, an additional 53 spectra were explained.

	BioWorks 3.3	PEAKS + Bioworks
ADH1	45.2%	50.43%
BGAL	37.4%	55.23%
TRFE	52.3%	63.78%
ALBU	59.0%	70.18%
LYSC	60.5%	65.99%
CYT	42.3%	51.92%

Figure1: Sequence coverage obtained for each of the six proteins known to be in the sample, obtained by searching against the Swiss-Prot database. Coverage numbers were calculated after filtering with BioWorks.The BioWorks only search achieved a remarkably low false positive rate (FPR) using the following filters:

Charge state	1	2	3	4
XCorr cut-off	1.0	1.8	2.5	3.4

The PEAKS search achieved a zero false positive rate by setting a peptide score cutt-off at 50%.

Additional hits obtained by using PEAKS + BioWorks are often the result of 'recovering' a low confidence hit that one search engine would have rejected, but that can be accepted when both search engines agree. Further coverage is obtained by accepting high scoring hits that either search engine found exclusively.

After exhausting the possibilities for conventional database searching, 335 spectra remained to be explained. Visual inspection confirmed that many of these were of good quality. Additionally, reasonable de novo sequences had been assigned for a large number of the spectra. Since these peptides came from protiens known to be in the sample, but remained unexplained, their sequences may have been modified; they may have mutated from the sequence in the database.

A sequence tag based peptide homology search tool, SPIDER, was used in an attempt to reconcile the de novo sequences with the sequences of the six proteins. The six proteins' sequences and reversed sequences were again used as the reference database for SPIDER. De novo sequences, derived by 'PEAKS auto de novo' from data that could not be explained by the inChorus search or the second pass PTM search, were submitted to SPIDER for sequence tag homology searching. A further 227 spectra were explained in this way. Of these, 120 can be considered trustworthy.

Since SPIDER's score is computed only from the alignment between de novo sequence and database homologue, it does not reflect a confidence in the correctness of a hit, rather it is a measure of the goodness-of-fit. As such, it is useful to plot the distribution of scores for both random hits (matches to the reverse database) and good hits (matches to the real sequences). This allows us to judge probability of correctness for a given SPIDER score (figure2).

Sequence Variation

Many peptides identified by SPIDER could not have been identified by traditional database searching because their sequences vary from the ones in the database as a result of some mutation. Figure 3 below shows some examples.

```
De Novo:  [EA]YVEVTK          De Novo:  TPW(CarboxymethC)NDGR
                |||||                     |           |    ||||
Real:     [AE]YVEVTK          Real:     TPW(CarboxymethC)NDGR
           || +|||||                      |           |    ||||
Database: [AE]FVEVTK          Database: RWW(CarboxymethC)NDGR

De Novo:  VFPSED[LG]R         De Novo:  SAGWNL<WD->K
          ||||||    |                   ||||||      |
Real:     VFPSED[GL]R         Real:     SAGWNI<VMA>K
          |||||+ ||  |                   ||||||    |  |
Database: VFPSEN[GL]R         Database: SAGWNI<PMA>K

De Novo:  VALD(CarboxymethC)QLAQVAER
          ||||          |    ||||||||
Real:     VALD(CarboxymethC)QLAQVAER
          + |+          |    ||||||||
Database: IGLN(CarboxymethC)QLAQVAER
```

Figure 3: Examples of sequence variates as found by the SPIDER search algorithm. Green denotes an exact match, light blue a homology mutation, and red a 'non-critical' de novo sequencing error corrected by comparison to the homologue.

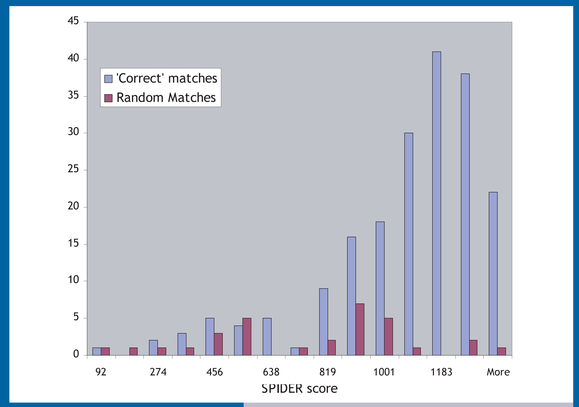

Figure 2: Illustrating the distribution of SPIDER scores for random matches, and hits to the known sequences.

Results and Conclusions

While the standard search algorithms, Sequest and PEAKS Protein ID, were able to identify all six of the proteins in the sample, there remained considerable portions of those proteins' sequences unaccounted for, and a large portion of the data (335 of 638) unexplained.

Using the workflow involving PEAKS auto de novo and SPIDER, a significant portion of the data can be explained (Figure 4). By explaining more of the high quality data, we offer a better explanation, and a better understanding of protein samples. Such a workflow can be very well employed in protein characterization research, but should also find a home in any proteomics analysis.

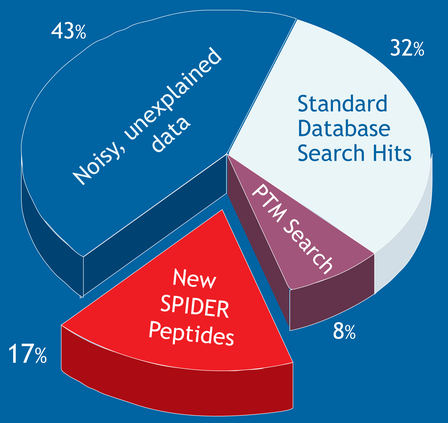

Figure 4: The proportion of the data explained by the various components of the workflow involving SPIDER.

References

2. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A. Lajoie, G., PEAKS: Powerful Software for Peptide De Novo Sequencing by MS/MS. (Rapid Communications in Mass Spectrometry, 17(20):2337-2342. 2003).

4. Iain Rogers, Michaela Scigelova, Gary Woffendin, Optimizing Data Acquisition for Automated de novo Sequencing, ASMS Poster 2007.

8. Pevtsov, S., Fedulova, I., Mirzaei, H., Buck, C., Zhang, X., Performance Evaluation of Existing De Novo Sequencing Algorithms (Journal of Proteome Research, 2006).