

Application Forum

Search for the Undiscovered Peptide

Using de novo sequencing and sequence tag homology search to improve protein characterization

Bin Ma, Iain Rogers

Bioinformatics Solutions Inc., Waterloo, ON, Canada

Because protein sequence databases will never be complete, cannot account for mutations between individuals and contain gene prediction errors, a large number of peptides cannot be identified using conventional MS/MS data search techniques – a hidden problem in protein identification and characterization research.

A new tool, SPIDER^{1*} is used to discover hidden peptides. Using a de novo sequence and a homologous sequence from the database, SPIDER reconstructs the real peptide, highlighting mutations and allowing for de novo sequencing error.

PROOF OF CONCEPT

Bovine Serum Albumin was analyzed on a Waters Q-TOF mass spectrometer. PEAKS^{*} software was used to generate de novo sequences for 28 spectra². These were the input for SPIDER searching against a Human database. SPIDER returned reasonable homologous peptides from Human Albumin for all spectra. The algorithm's proposed 'reconstructed' sequence was correct (i.e. matched exactly to Bovine Albumin) in 24 cases.

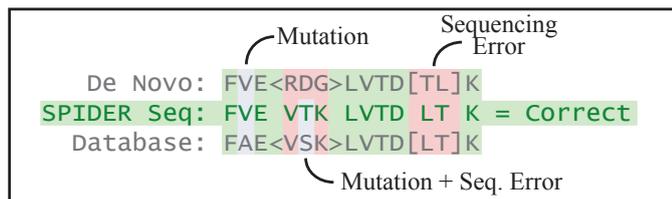


Figure 1: Illustrating how SPIDER reconstructs the correct sequence from a de novo sequence and a homologous peptide.

HOW MUCH ARE WE MISSING?

A tryptic digest of six known proteins was analysed by an LTQ Orbitrap mass spectrometer. Database search, using PEAKS Protein ID^{*} and Mascot[†] together, identified the proteins and explained (i.e. matched to the known proteins) 220 of the 692 spectra. Searching the identified proteins again, with a few variable modifications (Acetylation, Deamidation, Oxidation) turned on, explained a further 53 spectra.

De novo sequences, derived by PEAKS auto de novo^{*} from data that could not be explained otherwise, were submitted to SPIDER for sequence tag homology searching. To avoid false positives, only hits scoring higher than the best random hit (to the reverse sequences) were kept. As such, the new SPIDER tool explained another 120 spectra (Figure 2).

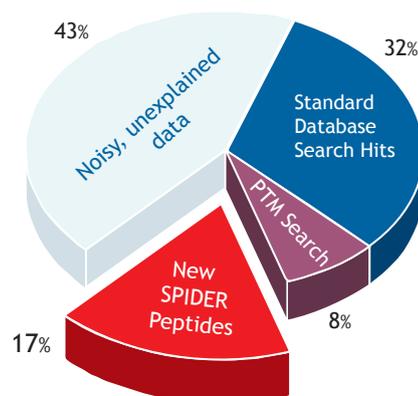


Figure 2: Peptide identification performance of various complementary techniques. SPIDER helps find a significant number of new peptides.

CONCLUSION

SPIDER allows us to find homologous peptides in the database and reconstruct the real peptide sequence from them. Since peptide sequence variations are common, use of this technique provides for significantly better understanding of our samples.

REFERENCES

1. Y. Han, B. Ma, and K. Zhang. 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. Rapid Communications in Mass Spectrometry, 17(20):2337-2342 2003.

[†] Mascot[™] is available from Matrix Science Ltd., UK.

^{*} SPIDER, PEAKS auto de novo, and PEAKS Protein ID, are available as part of the PEAKS Studio software from www.bioinfor.com.

Download this sheet, data, results and/or a demo of SPIDER from:
www.bioinfor.com/biotechniques2007.php