

Improving de novo sequencing accuracy for Ion Trap data in PEAKS software

Denis Yuen, Bin Ma, Iain Rogers

Introduction

De novo sequencing from MS/MS data is a well used method for sequencing peptides from organisms of unknown sequences, directly from their MS/MS spectra, or identifying peptides that vary from their database equivalents by some modification or mutation.

De novo sequencing programs typically require scoring functions that evaluates the fitness between a peptide sequence and the spectrum. Ma et al demonstrated that two scoring functions, used together, can improve de novo sequencing accuracy [1], but relative importance of each scoring function was not thoroughly evaluated.

In this work, the optimal weighting between multiple de novo sequencing score components is trained on a large dataset, and is demonstrated to provide a significant accuracy improvement in PEAKS Studio².

Methods

Powell's conjugate gradient algorithm, with a subset of the Wysocki data and corresponding correct answers for input, was used to train weighting constants for components of the original PEAKS score and new independent scores. De novo sequencing was run on all data before and after the improvements, and results were compared against the correct answers to gauge performance.

Agreement between the known, correct sequence and the top de novo sequence result for a given spectrum was computed automatically. For this comparison, a number of metrics were used. Number of correct de novo sequences, and percentage of residues correct are easily computed, but don't give a clear picture of how many errors are critical, and how many are due to simple transpositions, or equal mass substitutions. Since Leucine is equal in mass to Isoleucine, N=GG, TL=LT and Q=AG, any downstream analysis of de novo sequencing results must be tolerant of these, and other common substitutions. As such, a performance metric that considers equal mass substitutions should be used. For this purpose we adopted the Relative Sequence Distance (RSD) calculation proposed by Pevstov et. Al., setting the length of allowable substitutions to 1. RSD indicates the number incorrect residues, discounting those that can be corrected by making one equal mass substitution, expressed as a proportion of the peptide's length - a number from 0 to 1, with 0 being the best. Finally, each measure was summed or averaged (as appropriate) across the whole data set to illustrate the net benefit of the new scoring framework.

Data

For any machine-learning or algorithm training, a large dataset is necessary. As such, data comprising 28311 MS/MS spectra collected by Wysocki et. Al, on an LCQ mass spectrometer, and with sequences confirmed by accurate mass were used to train the scoring framework.

To avoid overtraining for a particular sample or instrument, performance was appraised on several other Ion Trap data sets:

- The "OPD dataset", is 280 spectra with known sequences selected³ from the Open Proteomics databank⁴.
- The "ISB dataset", comprising 2759 spectra is that subset of Keller et. Al's data having known sequences⁵.
- The "OrbiLTQ dataset" was collected by Scigelova and Woffendin⁶, and contains 158 spectra for which correct sequences are known. Carboxymethylated Cysteine residues are expected.
- The "Amgen dataset" was collected by Johnson et. Al⁷, on an LCQ mass spectrometer, and comprises 144 spectra for which correct sequences are known. Oxidated Methionine and Carbamidomethylated Cysteine residues are expected.

Correct answers were determined in each dataset by Sequest and/or PEAKS Protein ID search matching to the proteins known to be in the sample, and subsequent manual or accurate mass confirmation.

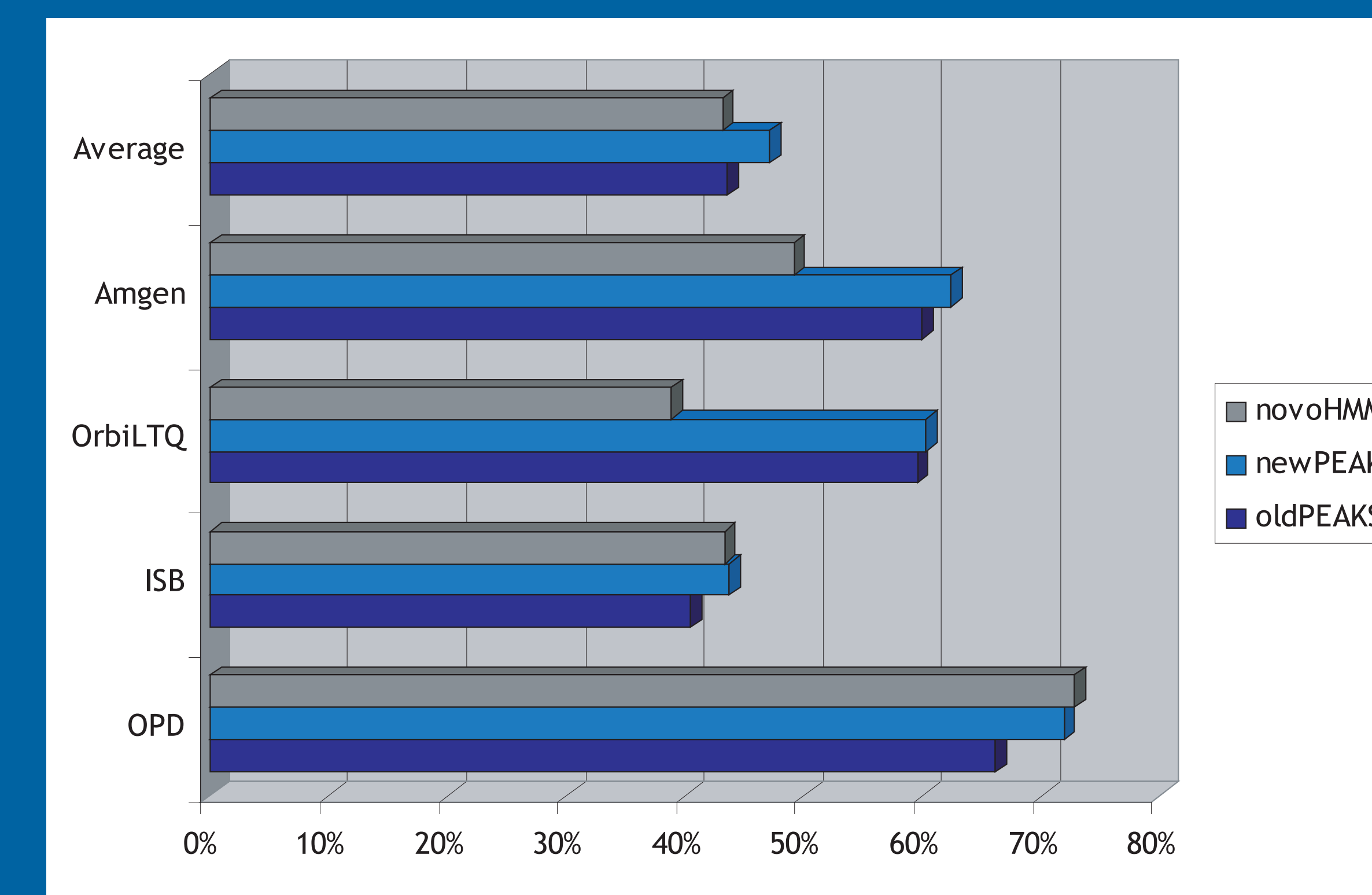


Figure 1: proportion of residues assigned correctly in the de novo sequence candidates proposed by each of the three algorithms. The topmost cluster of columns represent an average across each of the five data sets.

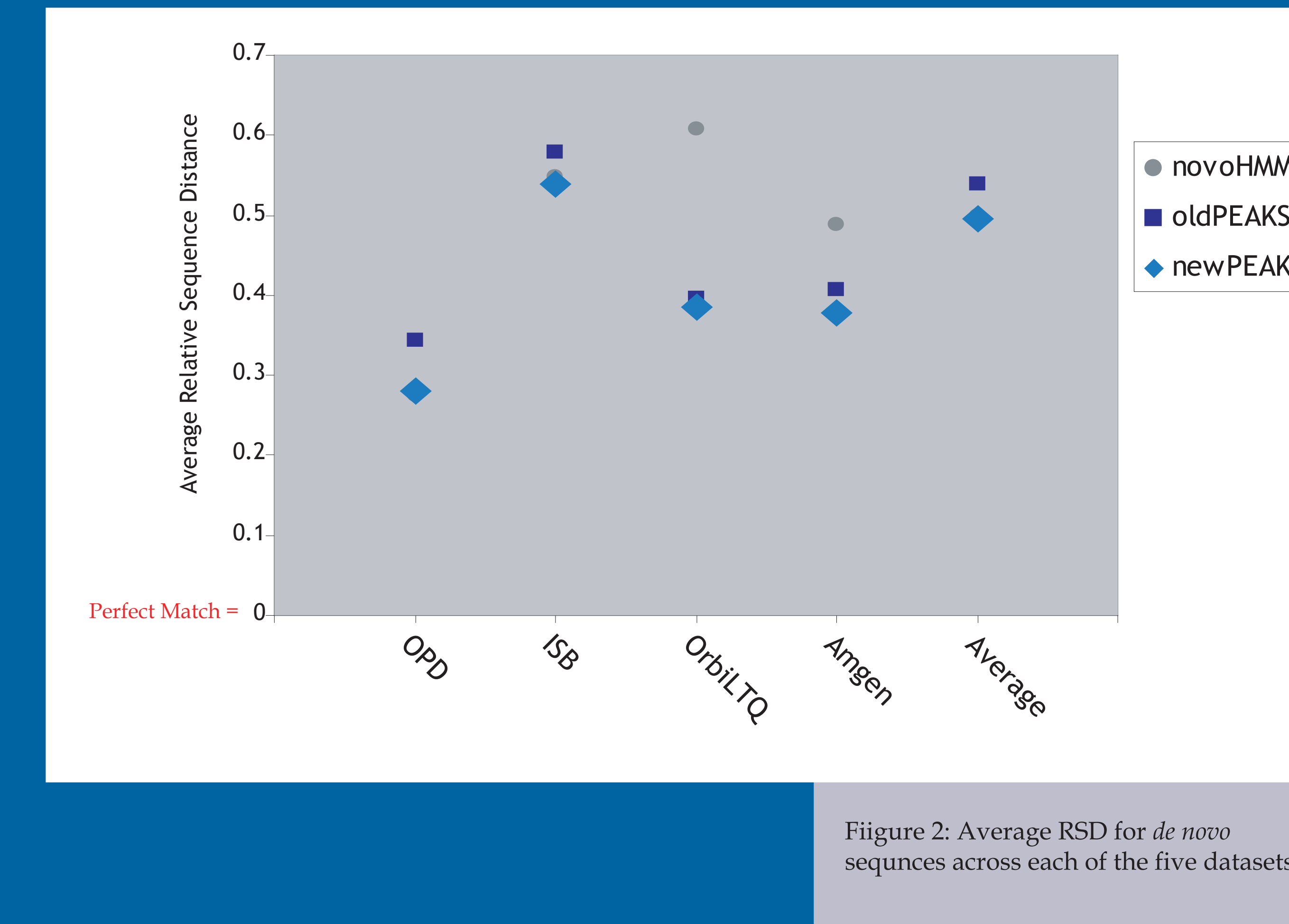


Figure 2: Average RSD for de novo sequences across each of the five datasets.

Results

Performance of the new scoring framework was weighed against the previous PEAKS auto de novo algorithm and, where possible, another de novo sequencing software.

As shown in Figure 1, the new scoring framework contributed to a improvement in the number of residues assigned correctly. Figure 2 shows this same performance benefit as a decrease in relative sequence distance. The RSD allows for equal mass substitution errors, but the improvement in RSD is proportional to the improvement in number of correctly assigned residues. From this we can conclude that the additional correct residues are found by avoiding critical errors.

Figure 3 shows a significant increase in the number of completely correct de novo sequences when using the new scoring framework.

Figure 3: number of completely correct de novo sequence candidates proposed by the three algorithms on each dataset. The far right cluster of columns represents a sum total for all five data sets.

Conclusions

Results on the total 31214 spectra from all three data sets combined show a marginal improvement in proportion of correctly assigned amino acids and average RSD, but a significant (>10%) improvement in the number of completely correct sequences. This suggests that while a number of the de novo sequences were close to being correct with the previous version of PEAKS, an improvement to the scoring algorithms can help determine the more correct of two similar candidates.

References

1. Ma, B., Lajoie, G., Improved positional confidence score in MS/MS peptide de novo sequencing. (ASMS 2006 poster MP348).
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).
3. Lijuan Mo, Debojyoti Dutta, Yunhu Wan, Ting Chen, MSNovo: a Dynamic Programming Algorithm for De Novo Peptide Sequencing via Tandem Mass Spectrometry, Jan 2007.
4. Prince, J.T., Carlson, M.W., Want, R., Lu, P., Marcotte, E.M., The need for a public proteomics repository. Nature Biotechnology 22 (2004), 471-474.
5. 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).
6. Iain Rogers, Michaela Scigelova, Gary Woffendin, Optimizing Data Acquisition for Automated de novo Sequencing, ASMS Poster 2007.
7. J. A. Taylor and R. S. Johnson (1997) "Sequence database searches via de novo peptide sequencing by tandem mass spectrometry". Rapid Comm. Mass Spec. 11:1067-1075.
8. Pevstov, S., Fedulova, I., Mirzaei, H., Buck, C., Zhang, X., Performance Evaluation of Existing De Novo Sequencing Algorithms (Journal of Proteome Research, 2006).

