PEAKS

Peptide *de novo* Sequencing Result Validation

Bioinformatics Solutions Inc.

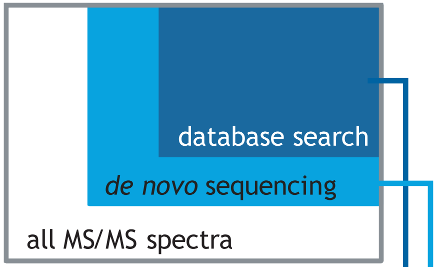
Lian Yang[1], Baozhen Shan[1], Bin Ma[2]
[1] Bioinformatics Solutions Inc, Waterloo, ON
[2] University of Waterloo, Waterloo, ON

Overview

We present a statistical method to determine a local confidence score threshold for automatic *de novo* sequencing result filtration.

Introduction

De novo sequencing is essential for complete proteomics analysis. As a supplement to protein database search, *de novo* sequencing interprets the large number of high quality spectra that do not match any database peptides, help characterize PTMs and amino acid mutations.



database search

de novo sequencing

all MS/MS spectra

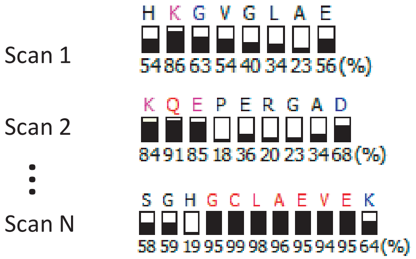
Spectra assigned to database peptides
1. Peptides within the database
2. Peptides with expected PTMs
FDR control using target/decoy DB

Spectra only interpreted by de novo
1. Peptides with unexpected PTMs
2. Peptides with mutations
3. Peptides from contaminates
4. Incomplete database
No established method for validation

The speed and accuracy of automatic *de novo* sequencing has improved significantly over the past 10 years. PEAKS software, for example, can perform *de novo* sequencing at a speed of 15 spectra per second on a desktop computer, matching the typical throughput of today's mass spectrometer.

Meanwhile, more residues are sequenced correctly thanks to the improvements in the *de novo* sequencing algorithm and also the use of high-resolution mass spectrometers with accurate mass measurements.

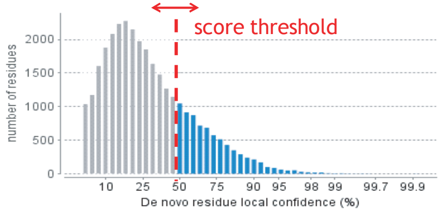
However, *de novo* sequencing often generates partially correct sequences due to ambiguities mostly caused by incomplete fragmentation. It is essential to have a *local confidence score* assigned to individual residues indicating how likely a residue is correctly sequenced.



Local Confidence Score Assigned to Individual Residues

While protein database search results are filtered using a target/decoy approach, there is no established method to filter out low confidence residues in *de novo* sequencing results.

In this research, a statistical method is proposed to determine a threshold on local confidence score by utilizing score distributions of *de novo* residues validated by database peptides.

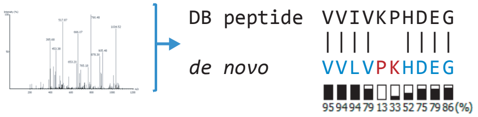


Distribution of Local Confidence Scores of Residues in *de novo* Sequeincg Result

Method

PEAKS software computes a residue local confidence score by combining multiple scoring features for the amino acid residues in a *de novo* sequence.

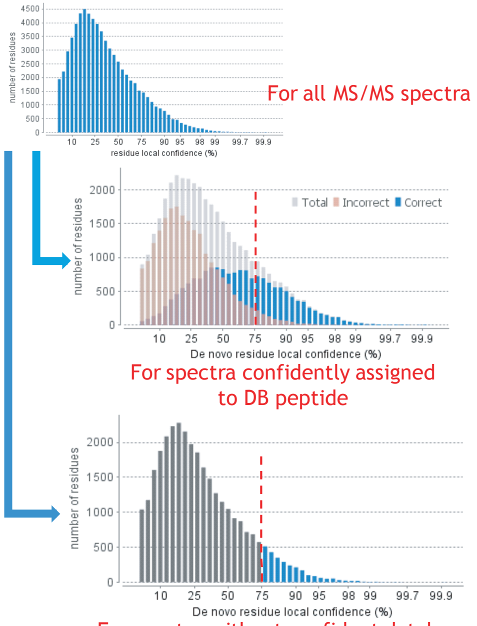
In proteomics analysis, after protein database search is performed, a *de novo* sequence can be validated when the MS/MS spectrum is also confidently assigned to a database peptide.



Validate *de novo* Residues using DB Peptides

By plotting the score distributions for *de novo* residues that agree/disagree with database peptides, a score threshold *T* can be determined to give a desired residue error rate for residues above the threshold. The threshold *T* is then applied to filter the *de novo* sequencing results on the spectra without a confident database peptide assignment.

Dataset 1 (Ion Trap/ETD)



For all MS/MS spectra

For spectra confidently assigned to DB peptide

For spectra without confident database peptide assignment

Local Confidence Score Distributions

Result

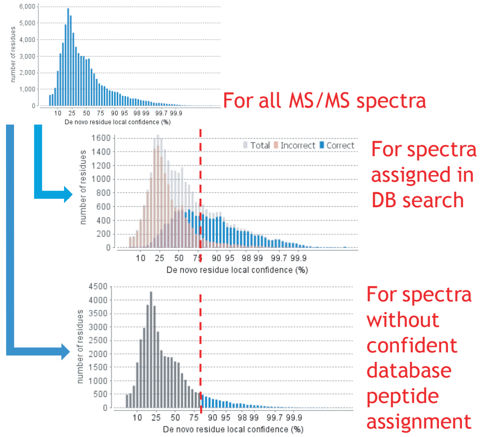
Three proteomics datasets were used in the evaluation. The three testing data sets contain 8031, 5152, 58159 MS/MS spectra, acquired from Ion Trap/ETD, Ion Trap/CID and Orbitrap/HCD, respectively.

De novo sequencing and protein database search were performed on each data set. Database peptide assignments with PSM FDR <1% are considered to be confident.

Local confidence score is calculated for every residue in each *de novo* sequence. *De novo* residues on spectra with confident database peptide assignments are validated.

The local confidence score distributions are plotted for residues which agree/disagree with the database peptide and also for residues only interpreted by *de novo* sequencing. Local confidence thresholds are automatically determined to have a residue error rate at 15%.

Dataset 2 (Ion Trap/CID)

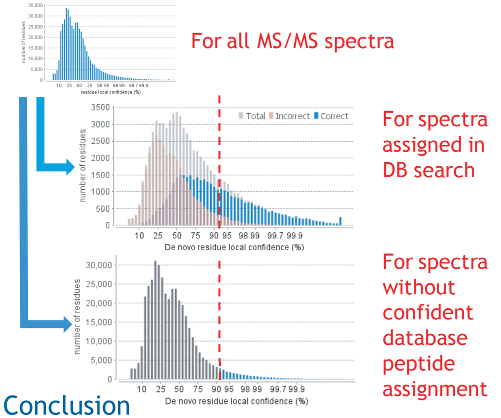


For all MS/MS spectra

For spectra assigned in DB search

For spectra without confident database peptide assignment

Dataset 3 (Orbitrap/CID)



For all MS/MS spectra

For spectra assigned in DB search

For spectra without confident database peptide assignment

Conclusion

PEAKS local confidence score separates the correct and incorrect *de novo* residues, and roughly represents the chance for a *de novo* residue to be correct.

The proposed method provides a guideline to automatically set a threshold on a local confidence score, which highlights confident residues in *de novo* sequencing results.