Integrating de novo sequencing and database search to improve peptide identification

Lei Xin,[1] Zefeng Zhang,[1] Lian Yang,[1,2] Baozhen Shan[1]

[1] Bioinformatics Solutions Inc, Waterloo, ON, Canada;[2] University of Waterloo, Waterloo, ON, Canada

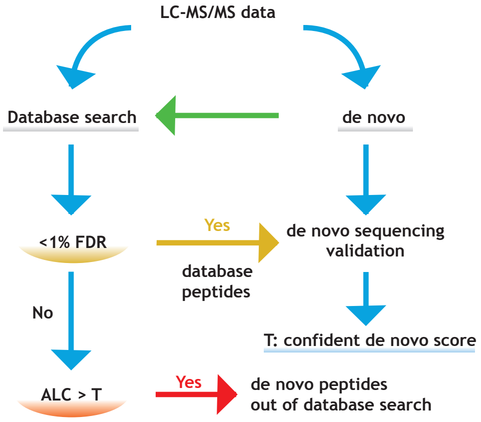
Overview

Purpose: To improve peptide identification
Methods: Integrating de novo sequencing and database search approaches
Results: 6% more peptide identification

Introduction

A key step in shotgun proteomics is peptide identification. There are two complementary approaches for the analysis of LC-MS/MS spectra: database search and de novo sequencing. A protein sequence database search is prioritized for database peptides and modified peptides, when a database is available. De novo sequencing is the only option for novel or homolog peptides which are not in the database. Unlike target-decoy approach for database search, there lacks a validation approach for peptide de novo sequencing. Here we describe a workflow integrating database search and de novo sequencing, in which database peptides are used to validate de novo peptides. Thus, the accuracy of de novo peptides can be estimated. The workflow maximized the peptide identification.

Figure 1. Workflow of integration of de novo and database search



LC-MS/MS data

Database search

de novo

<1% FDR

Yes

database peptides

de novo sequencing validation

No

T: confident de novo score

ALC > T

Yes

de novo peptides out of database search

Method

A local confidence score was assigned to each residue of the de novo peptides to indicate how likely a residue is correctly sequenced.

Let T1 be the set of MS/MS spectra. Perform de novo sequencing and database search with T1.

Let T2 be the set of the spectra identified by database search with 1% of FDR. A de novo peptide in T2 was validated with the database peptide at residue level. The local confidence score distributions were plotted for de novo residues that agree/disagree with database residues.

For the de novo peptides in T3 = T1–T2, their score distributions of correct and incorrect residues were estimated with validated distributions in Step 3.

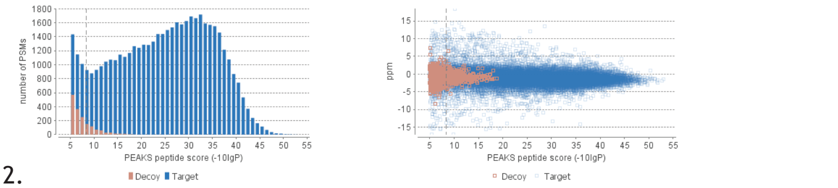
Results

The LC-MS/MS data set from yeast on an Orbitrap instrument [1] was used to demonstrate the workflow. It contains 35821 MS1 scans and 66479 MS/MS scans.

Figure 2. PSM score distribution

(a) Distribution of PEAKS peptide score

(b)Scatterplot of PEAKS peptide score versus precursor mass error



1. Perform database search with PEAKS 7 against SWISS-PROT. 43349 of 66479 spectra (65.2%) were identified with 1% false discovery rate. The target hits and decoy hits were shown in Figure 2.

2. Local confidence score was used to estimate the correctness of de novo sequence at residue level. Confident database hit was used to validate the de novo sequence of the same spectrum as shown in Figure 3. The assignment of amino acid in de novo sequence is correct if consistent with the one in database peptide, incorrect otherwise.

3. A local confidence score threshold can be determined to filter low confidence residues. Average local confidence (65%) was used to filter de novo sequences with less than 10% of residue error rate. 4428 spectra (6.7%) were identified with de novo sequences. The confident de novo peptides were exported along with confident peptides by database search.

4. With protein BLAST, 90% of the exported de novo peptides are significant (Table 1).

Figure 3. Validation of de novo local confidence score

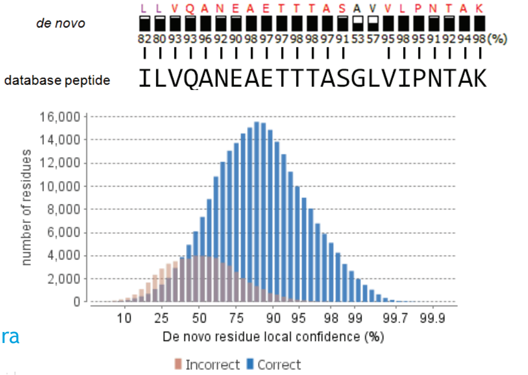
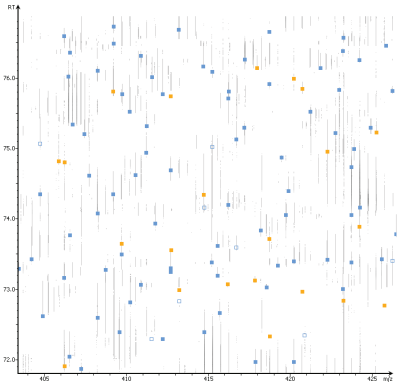


Table 1. Protein BLAST results of de novo peptides (part of the list)

Peptide	Protein	E-value	De novo Sequence	ALC	ppm
LSPVVVIGTGLAGLAAANELVNK	P32614	5.09E-06	Q(+.98)LSPVVVLGTGLAGLAAANELVNK	75	2.3
SQVYFDVEADGQPIGRVVFK	P14832	1.49E-05	EQVYFDVEADGQPLGRVVFK	87	-2.4
ATLHFVPQHEEEQVYSISGK	P11745	1.04E-04	VN(+.98)LHFVPQHEEEQVYSLSGK	69	-1.7
TDTNENFEGVSFMGK	P18562	1.11E-03	TTHLTDTNENFEGVSFMGK	81	-4.5
EENLRPAYDDQVNEDVYK	Q12211	1.16E-03	EQ(+.98)ENLRPAYDN(+.98)QVNEDVYK	88	-1.2
DHMEVFPAGSSFPSTK	P32589	1.21E-03	FC(+57.02)FN(+.98)EHMEVFPAGSSFPSTK	67	-1.5
NVLWDENNMSEYLT	P00044	1.72E-03	NVLWDENNMSQ(+.98)YLTLLM(+15.99)K	71	7.1
SAITALTPNQVNDELNK	P22203	1.91E-03	ESALTALTPNQVNDELNK	79	-1.2
QNSESIRMVLIGPPGAGK	A7THY5	2.41E-03	Q(+.98)SSESLRMVLLGPPGAGK	67	-1.2
LVLEVADHYVLDDLYAK	P32353	5.53E-03	QC(+57.02)LVLEVAN(+.98)HYVLDDLYAK	86	-2.1
LDAHLAPHPCPGK	P53751	6.14E-03	LDTHLAPHPC(+57.02)PGK	80	-0.7
DVTFLNDCVGPEVEAAVK	P00560	8.02E-03	DVTM(+15.99)LPDC(+57.02)VGPEVEAAVK	65	-3.4
LLEAFGSGTAAVVSPIK	P38891	9.40E-03	HC(+57.02)LLQ(+.98)AFGSGTAAVVSPLK	82	0.7

Figure 4. LC-MS map with MS2 spectra



🟦 MS2 spectra with database hits (65.2%)
🟨 MS2 spectra with de novo hits (6.7%)
□ MS2 spectra remain unmatched (28.1%)

Conclusions

A workflow to improves peptide identification.

References

[1] Nagaraj N. et al. System-wide perturbation analysis with nearly complete coverage of yeast proteome by single-shot ultra HPLC runs on a bench to Orbitrap. Mol Cell Proteomics 11(3):M111.013722 (2012).