Analysis and validation of proteomic data generated by tandem mass spectrometry

Alexey I Nesvizhskii, Olga Vitek & Ruedi Aebersold

Supplementary Notes

1. Database search parameters
2. Sources of false identifications

Supplementary Notes

1 DATABASE SEARCH PARAMETERS

Types of fragment ions. The theoretical fragment ion spectrum for each candidate peptide is calculated using common peptide dissociation rules specific to each type of mass spectrometer (most commonly b- and y-ions). Default settings are typically provided, and can be modified by the user.

Monoisotopic versus average mass. Depending on the resolution of the instrument, the measured mass can be closer to the monoisotopic peptide mass (high-resolution instruments) or to the averaged mass—that is, the weighted average of all isotopes (low-resolution instruments such as ion traps).

Peptide ion charge state. When the exact charge state of a multiply charged ion cannot be easily determined (low mass-resolution data), the MS/MS spectrum is searched against the database repeatedly—for example, once assuming 2+ and once assuming 3+ charge state—and the higher-scoring assignment selected.

Parent ion mass tolerance. The search is limited to database peptides with calculated mass within a certain range of the measured peptide mass, typically 2–3 Da for ion-trap, 0.2–0.5 Da for time-of-flight, and 5–10 p.p.m. for Fourier-transform instruments. As an option, one can specify a fairly large mass tolerance, and then use mass accuracy as an additional discriminative parameter at the validation stage.

Enzymatic digestion constraint. Most proteolytic enzymes cleave proteins specifically after certain residues in the protein sequence: for example, after arginine and lysine residues in the case of trypsin. The digestion enzyme can be specified to restrict the search—for example, to tryptic peptides with less than two missed cleavages only, or with nonspecific cleavage at not more than one of the peptide termini.

Modifications. Searches can be performed allowing one or more static (all occurrences of the residue are modified: for example, cysteine in ICAT experiments) or variable (the residue may or may not be modified: for example, phosphorylation or methionine oxidation) modifications.

2 SOURCES OF FALSE IDENTIFICATIONS

Simplified scoring scheme. All scoring schemes are based on a simplified representation of the peptide fragmentation process. In reality, the abundance of a fragment ion depends on various factors, such as the amino acids located on either side of the fragmented peptide bond.

Contaminant and low quality spectra. Spectra acquired on chemical contaminants introduced to the sample during sample preparation, or low quality peptide spectra (poor fragmentation, high signal-to-noise ratio) are prevalent in many datasets.

Fragmentation of multiple peptide ions. Spectra resulting from simultaneous fragmentation of two or more different peptide ions having similar m/z values that are concurrently isolated in the collision cell.

Presence of homologous peptides. When the database contains several peptides with a similar molecular weight that have a high degree of sequence homology, an incorrect (homologous) peptide can score higher than the correct one. Furthermore, several amino acid combinations (for example, aspartic acid and asparagine, or glutamic acid, glutamine and lysine) cannot be resolved using low mass-accuracy instruments, and two of the amino acids, isoleucine and leucine, have identical masses.

Incorrectly determined charge state or peptide mass. Peptide mass can also be measured inaccurately (or determined using incorrect charge state) and be outside of the mass tolerance window used. Even with high mass-accuracy instruments, the measured mass could be shifted by 1 or by 2 Da from the calculated monoisotopic peptide mass owing to selection of the first or the second isotope for fragmentation.

Restricted database search. A spectrum will not be identified if it is produced by a peptide containing an unspecified modification, resulting from an unexpected protein cleavage, or containing more than the allowed number of missed cleavages.

Sequence variants and new peptides. If peptide sequence is not in the searched database, it will not be identified using the standard database search approach.