OPEN ACCESS Freely available online

ISSN: 2153-0602
Journal of
Data Mining in
Genomics & Proteomics

Opinion Article

Evaluation of Protein Identification Using Randomized Sequence Database

Brook Searle*

Department of Genome Sciences, University of Washington, Seattle, United States

DESCRIPTION

The analysis of samples from understudied and/or undersequenced species as well as samples whose proteomes are derived from other organisms raises two significant issues. The first is if proteomic information gathered from an unusual sample type contains peptide tandem mass spectra. The second question is the availability of an appropriate protein sequence database for proteomic searches. A desirable technique for high-throughput proteomics (proteome) investigation is mass spectrometry. The most common method for identifying peptides and proteins is database searching using uninterested peptide tandem Mass Spectrometry (MS/MS) spectra. SEQUEST and Mascot are the two most often used database search algorithms for database searches of uninterested peptide MS/MS spectra. Additionally, open-source algorithms like X! Tandem and OMSSA (Open Mass Spectrometry Search Algorithm) have just been released. The fact that a significant portion of the top-scoring pep-tide matches to each spectra are false positive identifications, frequently as many as 90%, poses a significant challenge to the identification of peptides and, consequently, proteins using database searches of uninterrupted peptide MS/MS spectra. As a result, the identified peptides and therefore proteins must be screened.

This is often accomplished by either employing scoring criteria built into each database search algorithm or algorithms that analyse the outcomes of a database search programme to keep just the candidates with the highest likelihood of being accurate matches. Several statistical models, including our own earlier peptide and protein identification models and the more recent Logistic Identification Of Peptide Sequences (LIPS) model, have been developed that analyse the search results from the SEQUEST algorithm to predict whether a peptide match is correct. Typically, peptide MS/MS spectra from directly mixes of well-known standard protein samples are used to train these models.

The scoring methods used by the search algorithms Mascot, X! Tandem, and OMSSA all assess the probability of observed peptide fragment mass matches happening arbitrarily based on idealised random models.

Peptide identification using SEQUEST-based thresholds indicating the need for experiment-based estimates of FPRs. The statistical approaches described above can be used to generate estimates of the probability a candidate peptide match is correct and subsequently FPRs. However, the specific organisms under study, growth conditions, sequence databases, experimental protocols, types of instrumentation, and sample complexities of any particular experiments are unlikely to match the conditions under which these models are trained or the assumptions of idealized random peptide fragmentation models have been made. Therefore, the probabilistic estimates from these models are not likely to be reliable in the majority of situations, and study-specific (sample-specific) methods for estimating the error rates of peptide and protein identifications are vitally needed.

Recent research has demonstrated the requirement for experiment-based estimations of False Positive Rates (FPRs) for peptide identification utilising thresholds based on SEQUEST. The statistical methods mentioned above can be used to produce FPRs and estimates of the likelihood that a candidate peptide match is accurate. However, it is unlikely that the specific organisms being studied, growth conditions, sequence databases, experimental protocols, instrumentation types, and sample complexity of any given experiment will match the circumstances in which these models are trained or the presumptions of idealised random peptide fragmentation models have been made. Therefore, it is imperative to develop study-specific (sample-specific) methodologies for calculating the error rates of peptide and protein identifications as the probabilistic predictions from these models are unlikely to be trustworthy in the majority of circumstances.

Correspondence to: Brook Searle, Department of Genome Sciences, University of Washington, Seattle, United States, E-mail: bsearle@uw.as.edu

Received: 03-Jan-2023, Manuscript No. JDMGP-23-19625; **Editor assigned:** 06-Jan-2023, JDMGP-23-19625 (PQ); **Reviewed:** 20-Jan-2023, QC No. JDMGP-23-19625; **Revised:** 27-Jan-2023, Manuscript No. JDMGP-23-19625 (R); **Published:** 03-Feb-2023, DOI: 10.4172/2153-0602.23.14.274

Citation: Searle B (2023) Evaluation of Protein Identification Using Randomized Sequence Database. J Data Mining Genomics Proteomics. 14:274