PROPERTIES OF AVERAGE SCORE DISTRIBUTIONS OF SEQUEST: THE PROBABILITY RATIO METHOD

Pedro Navarro[1,4], Salvador Martínez-Bartolomé[1,2], Fernando Martín-Maroto[1], Daniel López-Ferrer[3], Antonio Ramos-Fernández[2], Margarita Villar, Josefa P. García-Ruiz[4], and Jesús Vázquez[4]

[1]These authors contributed equally to this work;
[2]Present address: Centro Nacional de Biotecnología,
Universidad Autónoma de Madrid, 28049 Cantoblanco, Madrid, Spain;
[3]Present address: Biological Sciences Division and Environmental Molecular Sciences Laboratory,
Pacific Northwest National Laboratory, Richland, WA 99352;
[4]Present address: Centro de Biología Molecular Severo Ochoa,
Universidad Autónoma de Madrid, 28049 Cantoblanco, Madrid, Spain.

High throughput identification of peptides in databases from tandem mass spectrometry data is a key technique in modern proteomics. Distribution-based scores are widely used to discriminate correct peptide identifications from large datasets of identified MS/MS spectra using searching engines such as SEQUEST. In this work we study the mathematical properties of average SEQUEST score distributions by introducing the concept of spectrum quality and expressing these average distributions as compositions of single-spectrum distributions. Our analysis leads to a novel indicator, the probability ratio, which takes optimally into account the statistical information provided by the first and second best scores. The probability ratio is a non-parametric and robust indicator that makes spectra classification according to parameters such as charge state unnecessary and allows a peptide identification performance, on the basis of false discovery rates, that is better than that obtained by other empirical statistical approaches. Besides, these identification methodologies are accompanied by the use of decoy databases to estimate the number of positive assignations and calculate false discovery rates. In conjunction with target databases, decoy databases may be used separately or in the form of concatenated databases, allowing a competition strategy; depending on the method used two alternative formulations are possible to calculate error rates. We show that both separate and concatenated approaches clearly overestimate error rates and, after analyzing as a whole the joint distribution of matches obtained after performing a separate database search and applying the competition strategy, we propose a new, integrated algorithm, tested in the practice with several scores, which makes a more accurate calculation of false discovery rates.