

A Computational Method for Assessing Peptide-Identification Reliability in Tandem Mass Spectrometry Analysis with SEQUEST

Jane Razumovskaya^{1,3}, Victor Olman¹, Dong Xu^{1,3}, Ed Uberbacher^{1,3}, Nathan Verbermoes³, Ying Xu^{1,2,3}

¹Life Sciences Division and ²Computer Sciences and Mathematics Division, Oak Ridge National Laboratory, TN 37830-6480, USA, ³School of Genome Science and Technology, University of Tennessee, Knoxville, TN 37922, USA

Abstract

High throughput protein identification in mass spectrometry is predominantly achieved by first identifying tryptic peptides using SEQUEST and then by combining the peptide hits for protein identification. Peptide identification is typically carried out by selecting SEQUEST hits above a specified threshold, the value of which is typically chosen empirically in an attempt to separate true identifications from the false ones. These SEQUEST scores are not normalized with respect to the composition, length and other parameters of the peptides. Furthermore, there is no rigorous reliability estimate assigned to the

protein identifications derived from these scores. Hence the interpretation of SEQUEST hits generally requires human involvement, making it difficult to scale up the identification process for genome-scale applications. To overcome these limitations, we have developed a method, which combines a neural network and a statistical model, for “normalizing” SEQUEST scores, and also for providing a reliability estimate for each SEQUEST hit. This method improves the sensitivity and specificity of peptide identification compared to the standard filtering procedure used in the SEQUEST package, and provides a basis for estimating the reliability of protein identifications.