Maximum Mean Discrepancy as a Similarity Metric between experimental and theoretical spectra in proteomics

Nicola De Simone*[1], Thomas Burger[2], Christophe Bruley[3], and Guido Uguzzoni[4]

[1]Etude de la dynamique des protéomes – Laboratoire Biosciences et bioingénierie pour la santé – France
[2]CEA Grenoble (BIG, Biologie à grande Echelle, EDyP) – INSERM U1038, Université Grenoble Alpes – 17 rue des Martyrs 38054 Grenoble Cedex 9, France
[3]Etude de la dynamique des protéomes – Laboratoire Biosciences et bioingénierie pour la santé – France
[4]Genetics and Chemogenomics – Laboratoire Biosciences et bioingénierie pour la santé – France

Abstract

Proteomics is the field that studies the proteome, the full set of proteins expressed by an organism. The most powerful method for proteome analysis relies on mass spectrometry (MS). Some amino acid sub-chains of the proteins studied, termed peptides, are ionized and fragmented in the mass spectrometer. Then, the instrument measures the mass-to-charge ratio (m/z) of the peptides and of their fragments. The masses and intensities of the fragments are returned as an experimental fragmentation spectrum, which is used to identify the peptide.
A typical identification workflow involves the systematic comparison of each experimental spectrum with all the theoretical fragmentation spectra of a reference database (which is in silico derived from the reference genome of the organism analyzed). The discrepancies between the two spectra are quantified by a score, and the best of all score for each experimental spectrum is used to match the spectrum to a peptide and then infer the protein (or gene) which originated it.

Consequently, the quality of the scoring function as a metric to quantify peptide-to-spectrum matches is paramount. Most search engines rely on scoring functions defined on R_+^N $\times R\_+^N, where N is the number of bins using to discretize the spectra. Such vectorizations are computationally yef ficien$

Keywords: Maximum Mean Discrepancy (MMD), Kernel based distance, Spectrum similarity, Scoring functions, Mass spectrometry

*Speaker