Abstract
Chemical cross-linking of proteins coupled with mass spectrometry (CXMS) is becoming a widely used approach studying protein-protein interaction because of its high-throughput property. Proteins are cross-linked before being digested, which produces cross-linked peptides and linear peptides. We are interested in cross-linked peptides because they provide information about protein-protein interaction. Most cross-linkers cannot be broken in dissociation, so half of dissociated ions contain subsequences from both peptide chains. In order to generate correct theoretical spectra, all possible peptide-peptide combinations should be built. The time complexity of querying an experimental spectrum is $O(dN^2N/m)$, where $N$ is the number of linear peptides in a database, $m$ is the number of precursor bins, and $d$ is a variable related to similarity measure. The space complexity of building a cross-linked peptides’ database is $O(lN^2N)$, where $l$ is the average length of a peptide chain. When $N$ is large, none of existing methods can handle an exhaustive search.

In this seminar we will introduce a novel cross-linked peptides identification method. During database searching, we use tree structure to achieve the logarithmic time complexity. Our method is an approximate method, which means that it cannot guarantee to find the most similar theoretical spectrum. For each experimental spectrum, we chose top 3 as candidates to estimate FDR. Significant PSMs (peptide-spectrum match) were reported as results. Experiments with synthetic peptides showed that our methods could find true positive without any additional false positive. Experiments with 26S proteasome samples showed that the probability of missing one of 3 most similar theoretical spectra was 0.2.