Abstract:

With the revolution in DNA sequencing technology, the amount of available sequence data grows at an exponential rate. An essential part of every genome project is the computational gene prediction. A major obstacle, however, is the absence of reliable and comprehensive sets of known genes to train the algorithms on.

Active learning is a special case of semi-supervised learning, where a training set is iteratively built from a small, but insufficient labelled dataset. In active learning the algorithm itself selects what example to include next, and then queries an "oracle" for the correct label. The selection criteria used are typically deterministic however, for instance by selecting the most uncertain example, resulting in a selection bias unsuitable for parameter estimation.

In this talk we present the basic ideas behind HMM-based gene prediction, and discuss a training method that combines the ideas of active learning and sequential sampling to construct an efficient, yet statistically robust training set. While this method can be extended to parameter estimation in general, we discuss the particular issues that arise when applying this to gene prediction.