Fast and accurate semi-supervised protein homology detection with large uncurated sequence databases

Pai-Hsi Huang, Pavel Kuksa, Vladimir Pavlovic
Rutgers University
Piscataway, NJ 08854
pkuksa@cs.rutgers.edu

May 20, 2008

Abstract

Establishing structural and functional relationship between sequences in the presence of only the primary sequence information is a key task in biological sequence analysis. This ability is critical for tasks such as inferring the superfamily membership of unannotated proteins (remote homology detection) when no secondary or tertiary structure is available. Recent methods such as profile kernels and mismatch neighborhood kernels have shown promising results by leveraging unlabeled data and explicit modeling mutations, insertions and deletions using *mutational neighborhood*. However, the size of such neighborhood exhibit exponential dependency on the cardinality of the alphabet set which incurs expensive cost for kernel evaluation and hence hinders the use of such powerful tools. Moreover, another missing component in previous studies for large-scale semi-supervised protein homology detection is a systematic and biologically motivated approach for leveraging the unlabeled data set.

In this study, we propose a systematic and biologically motivated approach for extracting relevant information from unlabeled sequence databases. We also propose a method to remove the bias caused by overly represented sequences which are commonly seen in the unlabeled sequence databases. Combining these approaches with a class of kernels (*sparse spatial sampling kernels, SSSK*) that effectively model mutation, insertion, and deletion, we achieve fast and accurate semi-supervised protein homology detection on three large unlabeled databases. The resulting classifiers based on our proposed methods significantly outperform previously published state-of-the-art methods in performance accuracy and exhibit order-of-magnitude differences in experimental running time.