
ABSTRACT

A classical problem in computational biology has been Remote Homology Detection – to detect homology in cases of low sequence similarity. Currently, the most effective methods that solve this problem involve SVMs. A crucial component that affects their performance is the “kernel function” that quantifies the sequence similarity. This paper introduces a new set of kernels called local alignment kernels for protein sequences, to detect high-scoring local alignments.

The key motivation behind Local Alignment (LA) kernels is to extend the classical work on local alignment scoring to kernel functions. For this purpose, the authors chose to use the Smith Waterman (SW) alignment score, which provides a similarity measure between proteins by incorporating knowledge (which has been optimized over time) about protein evolution. However, it cannot be used directly by an SVM as it does not have positive definiteness. Local Alignment(LA) kernels overcome this gap by mimicking the behavior of the SW score.

However, for practical applications, LA kernels encounter the diagonal dominance problem, wherein the kernel value decreases rapidly with increasing similarity. SVM performance in such cases is affected, and is rectified by taking a logarithm of the LA kernels. When these new kernels were tested on the SCOT superfamily, they outperformed other popular SVM methods.

The accuracy of this method comes from the combination of two popularly used algorithms:

1. SVM Algorithm – which performs classification on almost any kind of data (in most real-world applications) when a kernel function is provided.
2. Smith-Waterman Scores – local alignment which have been optimized over time to quantify similarity of biological sequences to a reasonable degree of accuracy.

While the SW Score alone is a weak method to detect remote homology, the results were highly impressive when SW Scores were used as a Kernel Function, combined with an SVM.

DISCUSSION

The papers discussed in class use Mismatch Kernels with SVMs, wherein sequence similarity is measured based on shared occurrences of fixed-length patterns of data. This paper uses LA kernels based on SW score optimized through evolution for the same purpose.

A key contrast between using mismatch kernels and LA kernels is that LA kernels outperform mismatch kernels in terms of accuracy, as all possible local alignment scores are considered. However, LA kernels use Dynamic Programming, and are thus slower than mismatch kernels. A Speed-Accuracy tradeoff needs to be made depending on the application.

The approach towards kernel design in this paper is also non-typical. Usually, a vector representation of sequences is first chosen, and the inner product in these representations decides the valid kernel. In this paper, the authors start with a similarity measure which is relevant to the problem, and then transform it into a valid kernel.