KERNELS FOR PROTEIN HOMOLOGY DETECTION

J. Dylan Spalding

School of Engineering, Computer Science and Mathematics,
University of Exeter

Submitted by J. Dylan Spalding to the
University of Exeter
as a thesis for the degree of
Doctor of Philosophy in Computer Science
October 2009.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other university.

_____ Signature.

J. Dylan Spalding

Abstract

Determining protein sequence similarity is an important task for protein classification and homology detection, which is typically performed using sequence alignment algorithms. Fast and accurate alignment-free kernel based classifiers exist, that treat protein sequences as a "bag of words". Kernels implicitly map the sequences to a high dimensional feature space, and can be thought of as an inner product between two vectors in that space. This allows an algorithm that can be expressed purely in terms of inner products to be 'kernelised', where the algorithm implicitly operates in the kernel's feature space.

A weighted string kernel, where the weighting is derived using probabilistic methods, is implemented using a binary data representation, and the results reported. Alternative forms of data representation, such as Ising and frequency forms, are implemented and the results discussed. These results are then used to inform the development of a variety of novel kernels for protein sequence comparison.

Alternative forms of classifier are investigated, such as nearest neighbour, support vector machines, and multiple kernel learning. A kernelized Gaussian classifier is derived and tested, which is informative as it returns a score related to the probability of a sequence belonging to a particular classification. Support vector machines are tested with the introduced kernels, and the results compared to alternate classifiers. As similarity can be thought of as having different components, such as composition and position, multiple kernel learning is investigated with the novel kernels developed here.

The results show that a support vector machine, using either single or multiple kernels, is the best classifier for remote protein homology detection out of all the classifiers tested in this thesis.

Contents

List of Tables

List of Figures

Notation and Abbreviations

λ	Eigenvalue or wildcard penalty
$\bar{\bar{\mathbf{\Lambda}}}$	Matrix of singular values from SVD
$\bar{\bar{\mathbf{\Sigma}}}$	Covariance matrix
$\bar{\bar{\boldsymbol{U}}}$	$m \times m$ unitary matrix from SVD of a $m \times n$ matrix
$\bar{\bar{\boldsymbol{V}}}$	$n \times n$ unitary matrix from SVD of a $m \times n$ matrix
$\bar{\bar{\boldsymbol{X}}}$	Matrix X
$\bar{\bar{\boldsymbol{X}}}^{+}$	Pseudo-inverse of matrix X
μ	Arithmetic mean
ν	Eigenvector
$\boldsymbol{X}$	Vector X
$\boldsymbol{X}^{\top}$	Vector X transpose
k	Length of the k-mer
k-mer	Contiguous oligomer, or subsequence, of length k
kNN	k nearest neighbour classifier
$Tr\bar{\bar{\boldsymbol{X}}}$	Trace of matrix X
AROC	Area under Receiver Operating Characteristic curve
GMM	Gaussian Mixture Model
kGMM	kernelised Gaussian Mixture Model
kPCA	kernel Principal Component Analysis
NN	Nearest neighbour classifier
PCA	Principal Component Analysis
PSST	Probabilistic Sequence Search Tool
ROC	Receiver Operating Characteristic
SVD	Singular Value Decomposition
wild_n	Wildcard kernel with n wildcards