Additional Paper for ‘Machine Learning for Protein Classification’: “Mismatch
String kernels for discriminative protein classification”

Paper reference:
“Combining Pair-wise Sequence Similarity and Support Vector Machines for remote Protein Homology
Detection” Li Liao & William Stafford Noble. Proceedings of the Sixth Annual International Conference

Abstract
Understanding the structure and function of proteins has always been an important research
goal in computational biology. One standard way to understand the function of a protein is by
homology detection, the goal of which is to try to detect sequence similarities between an unknown
protein and other proteins whose functions are already better understood since typically, sequence
similarity implies functional similarity. The authors report that the current state of the art for protein
homology detection is the SVM-Fisher method of Jaakkola et al which uses a generative, profile
Hidden Markov Model (HMM) and a discriminative learning algorithm for classification namely the
Support Vector Machine (SVM). The authors improve on SVM-Fisher by proposing SVM- pair-wise
which uses a pair-wise sequence similarity algorithm such as the Smith-Waterman algorithm instead
of the HMM. The suggested method differs from the SVM-Fisher scheme in the vectorization step
that happens before the SVM training. In the SVM-Fisher scheme, the vector representation of a
protein was its gradient with respect to a Profile HMM. For SVM- pair-wise, the vector consists of a
list of pair-wise sequence similarity scores. The authors justify this vectorization change by citing
three specific reasons. Firstly, the new representation is simpler since it eliminates the need for the
profile HMM topology and parameterization as well as training via expectation maximization.
Secondly, this scheme does not require a multiple alignment of training set sequences. Thirdly, this
scheme uses the information contributed by negative examples as well, when learning the decision
boundary. The authors compare two variants of their scheme with five other protein homology
detection methods including SVM-Fisher and other BLAST based methods. The methods are
evaluated based on their ability to discover previously unseen families from the SCOP database using
all other members of the family’s superfamily for training. The SVM- Pair-wise variants outperform
the existing methods in detecting remote homologs. The authors note that the improvements obtained
are not entirely attributable just to the use of negative examples. In fact the reduction in training time
when using just the positive examples appears an appealing alternative in spite of a marginal drop in
homology detection accuracy. Use of K-nearest neighbor, another discriminative classifier instead of
the SVM does offer gains compared to PSI-BLAST but underperforms compared to the SVM based
methods, thus justifying the utility of use of the SVM algorithm.

Discussion
Both papers tackle the problem of classification of proteins into appropriate structural and
functional classes based on the sequence information. The use of mismatch kernels for discriminative
protein classification offers two potential benefits over the SVM- pair-wise approach discussed
above. Firstly, the use of kernels solves an efficiency issue since it obviates the need for vectorization
as now the vectors do not need to be explicitly computed. Secondly, the mismatch kernel suggests a
biologically well motivated way to compare sequences by allowing for mutations between sequence
patterns. The mismatch tree data structure enables efficient computation of the mismatch kernel and
linear time prediction. When the SCOP sequences are augmented by domain homologs of positive
training examples to assist HMM based methods, the mismatch kernel SVM performs competitively
with SVM-Fisher. In the absence of these domain homologs, the authors compare the mismatch
kernel SVM with SVM- pair-wise and find that the performances are on par. Another advantage that
the SVM mismatch kernel method offers is that, unlike the SVM Pair-wise method which offers no
biological insight about the sequence it classifies, it allows for extracting high scoring k-mers from a
trained SVM to look for discriminative motif regions in the positive sequence family.