Supplementary Material for Poly(A) motif prediction using spectral latent features from human DNA sequences

Bo Xie[1], Boris R. Jankovic[2], Vladimir B. Bajic[2], Le Song[1,*], and Xin Gao[2,3,*]
[*]All correspondence should be addressed to lsong@cc.gatech.edu and xin.gao@kaust.edu.sa.
[1]College of Computing, Georgia Institute of Technology, Atlanta, GA, 30332, USA.
[2]Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia.
[3]Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia.

S1 Comparison with String Kernels on Additional Datasets

We compared the proposed method with other string kernels on two other datasets, i.e., transcription start site (TSS) prediction [Ohler *et al.*(2002)Ohler, Liao, Niemann, and Rubin] and splice site prediction [Rätsch *et al.*(2006)Rätsch, Sonnenburg, and Schäfer]. The TSS dataset was extracted from the Drosophila genome [Ohler *et al.*(2002)Ohler, Liao, Niemann, and Rubin]. It contains 1,864 positive samples and 4,658 negative samples. Each sample (a nt sequence) contains 300 nt which includes 250 nt upstream and 50 nt downstream of a TSS. We conducted a five-fold cross-validation on the TSS dataset to test the performance of the k-spectrum (SPE) kernel, the weighted degree (WD) kernel and our method. The average performance over the five folds is reported in Table S1. The splice site dataset was taken from [Rätsch *et al.*(2006)Rätsch, Sonnenburg, and Schäfer], which contains 8,842 positive samples and 253,579 negative samples. Each sample contains 141 nt centered around a AG dimer. We followed the same training/validation/test set partition as [Rätsch *et al.*(2006)Rätsch, Sonnenburg, and Schäfer], which results in 100,000 samples for training, 100,000 samples for tuning the parameters and 62,421 samples for testing. Table S1 shows the performance of different string kernels on this dataset.

For the TSS dataset, our method is consistently the best one among the three methods in terms of the error rate, the false positive rate and the false negative rate. For the splice site dataset, our method is comparable with WD. The reason that we did not have a significant advantage over WD is that this dataset was extremely unbalanced. As mentioned above, the number of the negative sequences was over 28 times more than that of the positive sequences. A single negative HMM model might not be rich enough to account for all the sequences variations. Also, the negative sequences were randomly generated, so the patterns were averaged out in fitting a single negative HMM. This "averaging" effect reduced our discriminative power.

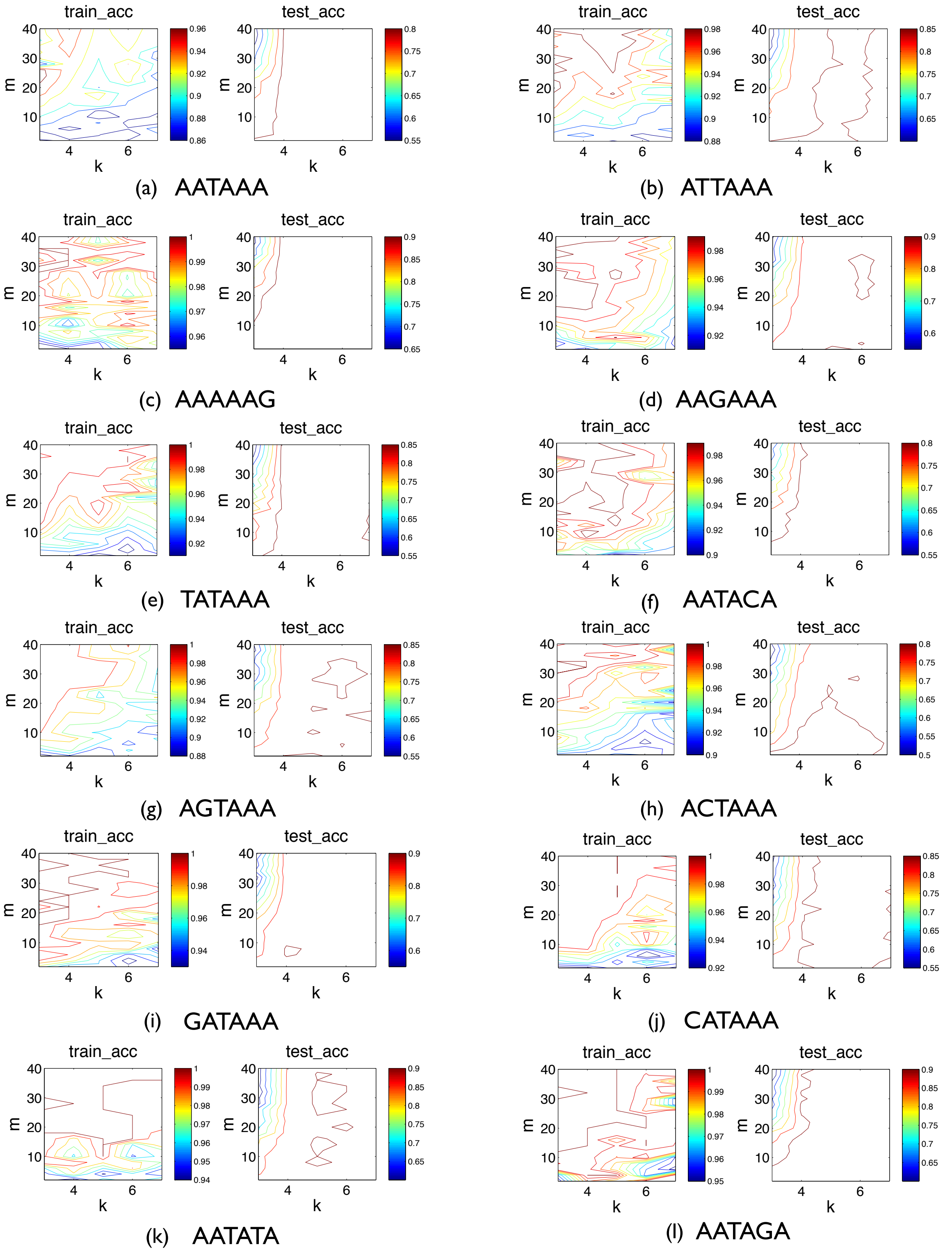
Table S1. Comparison of the performance of our method (HMM) with SPE and WD on the TSS dataset and the splice site dataset. "Rel" denotes the relative improvement of HMM with respect to SPE. The lowest error rate for each motif variant is indicated in bold. All the values in the table are in percentiles. For splice site, SPE did not finish within 24 hours, so we do not report its performance here.

Dataset	Error Rate				False Negative Rate				False Positive Rate			
	SPE	WD	HMM	Rel	SPE	WD	HMM	Rel	SPE	WD	HMM	Rel
TSS	13.99	12.87	**11.19**	13.06	29.40	24.30	**22.91**	5.74	7.83	8.30	**6.49**	21.74
Splice site	-	**1.03**	1.12	-8.84	-	**18.98**	23.02	-21.27	-	0.39	**0.34**	12.71

References

Ohler *et al.*(2002)Ohler, Liao, Niemann, and Rubin. Ohler, U., Liao, G.-c., Niemann, H., and Rubin, G. M. (2002). Computational analysis of core promoters in the drosophila genome. *Genome Biol*, **3**(12), RESEARCH0087.

Rätsch *et al.*(2006)Rätsch, Sonnenburg, and Schäfer. Rätsch, G., Sonnenburg, S., and Schäfer, C. (2006). Learning interpretable svms for biological sequence classification. *BMC Bioinformatics*, **7 Suppl 1**, S9.



(a) AATAAA

(b) ATTAAA

(c) AAAAAG

(d) AAGAAA

(e) TATAAA

(f) AATACA

(g) AGTAAA

(h) ACTAAA

(i) GATAAA

(j) CATAAA

(k) AATATA

(l) AATAGA

Figure S1. Grid search of parameters k (the number of nucleotides to combine into a mega-observation) and m (the number of hidden states) for training and testing accuracies of the 12 variants.