Kernel methods for DNA barcoding

Pavel P. Kuksa and Vladimir Pavlovic

Department of Computer Science
Rutgers University
{pkuksa, vladimir}@cs.rutgers.edu

We study kernel methods for species-level identification based on short DNA fragments known as barcodes and propose a framework for fast string kernel algorithms suitable for large and increasingly growing sets of barcode data. We describe a divide-and-conquer algorithm for mismatch kernel computation that improves currently known bound for the mismatch kernel algorithms. We also introduce a new framework for string kernels with feature selection that leads to algorithms that are simple and easy to implement as well as faster and more memory efficient in practice than the state-of-art suffix tree-based approaches. Crucial benefits of the proposed approach is its computational efficiency and high accuracy.

There are several types of kernels for biological sequences, including kernels derived from probabilistic models [4], k-mer string kernels [6, 7], and weighted-decomposition kernels [9]. In this work we focus on recently proposed k-mer string kernels, and their application in the DNA barcoding setting for efficient and accurate species-level identification. We report novel ways to more efficiently compute the k-mer kernels. We show that using the proposed algorithms a k-mer kernel with m mismatches for N sequences of length n can be computed $O(nN/u)$ times faster, where u is the total number of different k-mers in the N sequences. We also show that the k-mer spectrum kernel can be efficiently computed for N sequences of length n in $O(Nnk)$ time and linear space using sorting, resulting in elimination of large storage overhead and time constants associated with suffix-tree based algorithms [6].

DNA Barcoding. Biological species identification through DNA barcodes has been proposed recently in [2]. In the DNA barcoding setting, DNA sequencing of the mitochondrial region is used to obtain a relatively short sequence or a DNA barcode that is subsequently used as a marker for identification and classification of the species. Several computational methods, based on pairwise alignments [13] or statistical approaches using evolutionary distances [11], have been applied to the tasks of identification and analysis of the DNA barcode data. However, a number of challenges remain to be addressed, including the accuracy of identification, and efficiency and scalability of computational methods. Identification accuracy is the critical issue in DNA barcoding and was a topic of many previous studies [10, 8, 13, 11]. We propose an efficient approach to perform sequence identification that will provide both computationally efficient and accurate solution to the problem of multiclass sequence classification.

Species identification method. In our approach, species identification is performed by first transforming sequences (potentially of varying length) into fixed-length representations (string spectrums) and then classifying sequences into one of many established species classes. For classification, we use Support Vector Machine (SVM) classifiers [12, 14]. The use of discriminative classifiers, like SVM, to perform DNA-based species identification is motivated by their strong performance in similar tasks of protein classification and remote homology detection [6, 7, 5, 1]. As a critical component of this approach to address efficiency of computations we propose a framework for faster string kernel algorithms.

Experimental results. We demonstrate that our methods are successful in discovering identity of specimens when applied to several collections of barcodes. At the same time we are able to identify signatures (small subsets of sequence features) for many species classes. Our results show that DNA barcoding can be successfully used to identify specimens by examining a *few* sequence features, resulting in increased scalability and interpretability of current computational approaches to barcoding and species classification tasks. Our results demonstrate that string kernel-based methods for species identification provide an effective and highly accurate alternative to the traditional methods. In our series of experiments, we evaluate several feature spaces and test the performance of our classifiers on a number of publicly available barcode datasets. We compared performance of the string kernel-based method using SVM as

Topic: learning in biological systems, learning algorithms
Preference: oral/poster
First author presenting

a base learner with a number of other classification methods. In particular, we evaluated Fisher kernel method [4], PSI-BLAST, ridge regression, and nearest neighbor methods. Our experiments show order of magnitude running time improvement (Table 1) for the k-mer kernel with m mismatches (by factors of 100-200 times depending on the dataset size) (we implemented and tested our algorithms in Matlab). In our experiments we also observe that k-mer string kernels considerably improve identification accuracy compared to previously reported results of [8, 11] (for example, on *Astraptes* dataset [3] the test error rate of multi-class SVM is only 0.67% compared to 9% in [8] or 20% in [11]). Our experiments with feature selection show that even with only 10% of the k-mers remained, k-mer kernels demonstrate similar or enhanced classification performance.

Table 1: Running time comparison (k-mer kernel with m mismatches, k=5, m=1)

Barcode dataset	algorithm 1, time (s)	algorithm 2, time (s)
Dataset 1 Astraptes N=466, n=600	16	2841
Dataset 2 Hesperiidae N=2135, n=600	301	75218

Algorithm 1 = new algorithm, algorithm 2 = original mismatch algorithm
machine configuration used: 2.8Ghz CPU, 1GB RAM

References

[1] Jianlin Cheng and Pierre Baldi. A machine learning information retrieval approach to protein fold recognition. *Bioinformatics*, 22(12):1456–1463, June 2006.

[2] Paul D. N. Hebert, A. Cywinska, S.L. Ball, and J.R. deWaard. Biological identifications through dna barcodes. In *Proceedings of the Royal Society of London*, pages 313–322, 2003.

[3] Paul D. N. Hebert, Erin H. Penton, John M. Burns, Daniel H. Janzen, and Winnie Hallwachs. Ten species in one: Dna barcoding reveals cryptic species in the neotropical skipper butterfly astraptes fulgerator. In *PNAS*, volume 101, pages 14812–14817, 2004.

[4] Tommi Jaakkola, Mark Diekhans, and David Haussler. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 7(1-2):95–114, 2000.

[5] Rui Kuang, Eugene Ie, Ke Wang, Kai Wang, Mahira Siddiqi, Yoav Freund, and Christina Leslie. Profile-based string kernels for remote homology detection and motif extraction. In *CSB '04: Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB'04)*, pages 152–160, Washington, DC, USA, 2004. IEEE Computer Society.

[6] Christina S. Leslie, Eleazar Eskin, and William Stafford Noble. The spectrum kernel: A string kernel for svm protein classification. In *Pacific Symposium on Biocomputing*, pages 566–575, 2002.

[7] Christina S. Leslie, Eleazar Eskin, Jason Weston, and William Stafford Noble. Mismatch string kernels for svm protein classification. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *NIPS*, pages 1417–1424. MIT Press, 2002.

[8] Mikhail V. Matz and Rasmus Nielsen. A likelihood ratio test for species membership based on dna sequence data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1462):1969–1974, 2005.

[9] Sauro Menchetti, Fabrizio Costa, and Paolo Frasconi. Weighted decomposition kernels. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 585–592, New York, NY, USA, 2005. ACM Press.

[10] C. P. Meyer and G. Paulay. Dna barcoding: error rates based on comprehensive sampling. *PLoS Biol*, 3(12), December 2005.

[11] Rasmus Nielsen and Mikhail Matz. Statistical approaches for dna barcoding. *Systematic Biology*, 55(1):162–169, 2006.

[12] B. Schölkopf and A. J. Smola. *Learning with kernels*. MIT Press, 2002.

[13] Dirk Steinke, Miguel Vences, Walter Salzburger, and Axel Meyer. Taxi: a software tool for dna barcoding using distance methods. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1462):1975–1980, 2005.

[14] Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998.