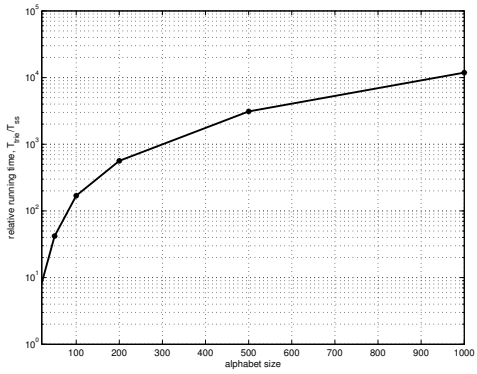

Figure 1: Relative running time T_{trie}/T_{ss} (in logarithmic scale) of the mismatch-trie and mismatch-ss as a function of the alphabet size (mismatch(5,1) kernel, $n = 10^5$)

Table 1: Running time (in seconds) for kernel computation between two strings on real data

	long protein	protein	dna	text	music
n	36672	116	570	242	6892
$\|\Sigma\|$	20	20	4	29224	1024
(5,1)-trie	1.6268	0.0212	0.0260	20398	526.8
(5,1)-ss	0.1987	0.0052	0.0054	0.0178	0.0331
time ratio	8	4	5	10^6	16,000
(5,2)-trie	31.5519	0.2918	0.4800	-	-
(5,2)-ss	0.2957	0.0067	0.0064	0.0649	0.0941
time ratio	100	44	75	-	-

the semi-supervised setting for neighborhood mismatch kernels; for example, computing a smaller neighborhood mismatch(5,2) kernel matrix for the *labeled sequences* only (2862-by-2862 matrix) using the Swiss-Prot unlabeled dataset takes $1,480$ seconds with our algorithm, whereas performing the same task with the trie-based algorithm takes about 5 days.

6.2 Empirical performance analysis

In this section we show predictive performance results for several sequence analysis tasks using our new algorithms. We consider the tasks of the multi-class music genre classification [16], with results in Table 2, and the protein remote homology (superfamily) prediction [9, 2, 18] in Table 3. We also include preliminary results for multi-class fold prediction [14, 15] in Table 4.

On the music classification task, we observe significant improvements in accuracy for larger number of mismatches. The obtained error rate (35.6%) on this dataset compares well with the state-of-the-art results based on the same signal representation in [16]. The remote protein homology detection, as evident from Table 3, clearly benefits from larger number of allowed mismatches because the remotely related proteins are likely to be separated by multiple mutations or insertions/deletions. For example, we observe improvement in the average ROC-50 score from 41.92 to 52.00 under a fully-supervised setting, and similar significant improvements in the semi-supervised settings. In particular, the result on the Swiss-Prot dataset for the $(7, 3)$-mismatch kernel is very promising and compares well with the best results of the state-of-the-art, but computationally more demanding, profile kernels [2]. The neighborhood kernels proposed by Weston et al. have already shown very promising results in [7], though slightly worse than the profile kernel. However, using our new algorithm that significantly improves the speed of the neighborhood kernels, we show that with larger number of allowed mismatches the neighborhood can perform even better than the state-of-the-art profile kernel: the (7,3)-mismatch neighborhood achieves the average ROC-50 score of 86.32, compared to 84.00 of the profile kernel on the Swiss-Prot dataset. This is an important result that addresses a main drawback of the neighborhood kernels, the running time [7, 2].

Table 2: Classification performance on music genre classification (multi-class)

Method	Error
Mismatch (5,1)	42.6±6.34
Mismatch (5,2)	35.6±4.99

Table 3: Classification performance on protein remote homology prediction

dataset	mismatch (5,1)		mismatch (5,2)		mismatch (7,3)	
	ROC	ROC50	ROC	ROC50	ROC	ROC50
SCOP (supervised)	87.75	41.92	90.67	49.09	91.31	52.00
SCOP (unlabeled)	90.93	67.20	91.42	69.35	92.27	73.29
SCOP (PDB)	97.06	80.39	97.24	81.35	97.93	84.56
SCOP (Swiss-Prot)	96.73	81.05	97.05	82.25	97.78	86.32

For multi-class protein fold recognition (Table 4), we similarly observe improvements in performance for larger numbers of allowed mismatches. The balanced error of 25% for the (7,3)-mismatch neighborhood kernel using Swiss-Prot compares well with the best error rate of 26.5% for the state-