Kernel methods applied to DNA sequence classification

Haozhe SUN

March 26, 2019

1 Abstract

The goal of this data challenge is to predict whether a DNA sequence region is binding site to a specific transcription factor (TF). More specifically, we work on three datasets corresponding to 3 different TFs.

First, I made some experiments by assuming the *uniform hypothesis* that these 3 different TFs can be treated equally. However, the experiments showed that this hypothesis is not valid because I got better results by assuming the *specific hypothesis* that different TFs correspond to different distributions of DNA sequences. I implemented a general scalable framework for machine learning algorithms using kernel methods from scratch. The implemented algorithms include kernel logistic regression, kernel SVM classifier and kernel 2-SVM classifier, applied to spectrum kernel and Local Alignment (LA) kernel [1]. These kernels can also be added up to combine multiple kernels. Then, ensemble techniques are used to improve the predictions. I have also studied some other kernels. [2] presents a way to use physicochemical features of DNA, [3] and [4] incorporate local DNA shape properties into the prediction models as it is known that TF binding is mediated in part by the shape of the DNA binding site.

2 Algorithms

I implemented a general scalable framework for machine learning algorithms using kernel methods. The implementation is said to be scalable because it provides uniform API to different classifiers and kernels, thus new methods can be added easily. The implemented classifiers include kernel logistic regression, kernel SVM classifier and kernel 2-SVM classifier. Kernel logistic regression is implemented by iteratively solving a weighted kernel ridge regression (WKRR), the WKRR is solved by Formula 1. The implemented kernels include spectrum kernel and Local Alignment (LA) kernel [1]. It turns out that spectrum kernel is much faster to compute than LA kernel.

$$\alpha = W^{\frac{1}{2}}(W^{\frac{1}{2}}KW^{\frac{1}{2}} + n\lambda I)^{-1}W^{\frac{1}{2}}Y \tag{1}$$

3 Uniform hypothesis

The *uniform hypothesis* consists in assuming that the DNA sequences are equally distributed in terms of classification into binding sites to different TFs. This hypothesis is useful if we do not have the prior knowledge that a fixed number of datasets corresponding to different TFs are provided. More specifically, I concatenated 6000 training samples without reordering, the last 1000 are used for validation whereas the first 5000 are used for training. Figure 1 and Figure 2 show the validation accuracy under this hypothesis.

Some details of experiments deserve to be mentioned. According to my experiments with kernel logistic regression, when λ is small (10^{-6}, 10^{-7}), training of the model with small values (especially 1, 2, 6) of k will be extremely time-consuming (several hours versus several minutes) and is thus prohibitive to use in practice. Luckily, according to other experiments, it is not likely

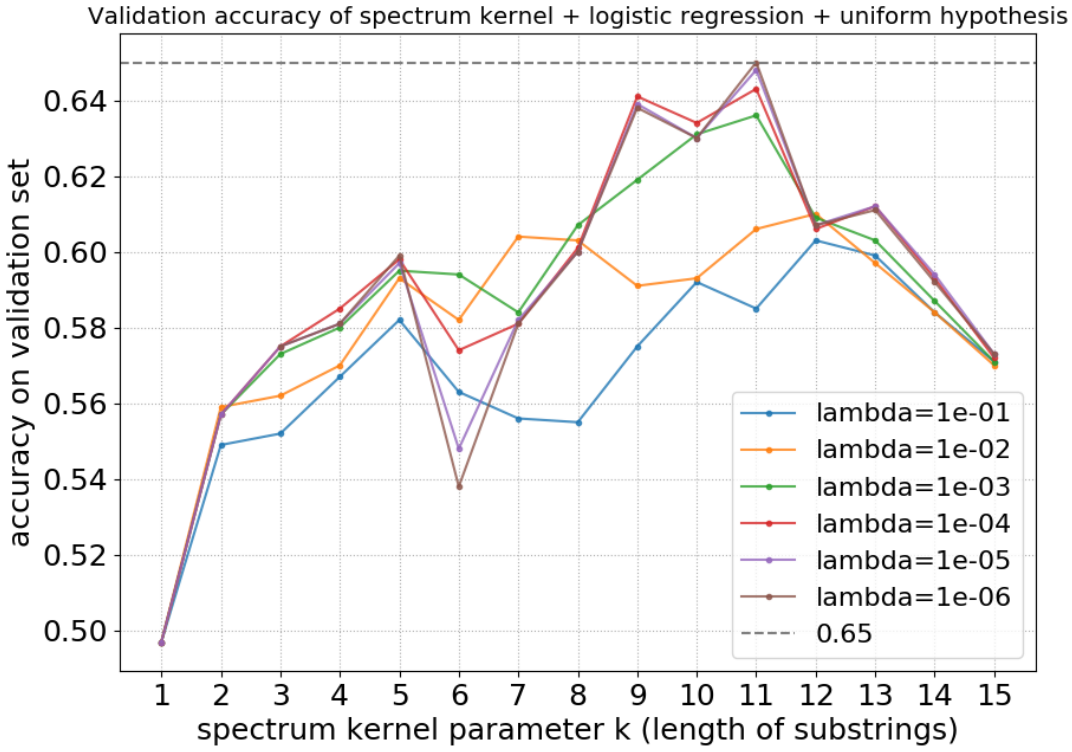

Figure 1: validation accuracy: spectrum kernel, logistic regression, uniform hypothesis.

that we can obtain good validation accuracy with those small values of k, so that is not a problem. The fact that the model is extremely time-consuming to train with small λ and small k could be explained by the small eigenvalues in the matrix $(W^{\frac{1}{2}}KW^{\frac{1}{2}} + n\lambda I)$ in Formula 1. $\lambda = 10^{-7}$ and $\lambda = 10^{-8}$ have almost the same performance as $\lambda = 10^{-6}$ on validation set. For the readability of the Figure 1, I do not plot the curve of $\lambda = 10^{-7}$, $\lambda = 10^{-8}$. When λ is too small (e.g. less than 10^{-6}), the influence of λ becomes negligible. $\lambda = 0$ will cause singular matrix error when training. Experiments with C-SVM have similar effects when C becomes too large.

Later, it will be shown that *uniform hypothesis* is not appropriate as experiments with *specific hypothese* yield better results. As the result, the validation accuracy with *uniform hypothesis* is indeed biased as all the validation set comes from the third datasets.

4 Specific hypothesis

The *specific hypothesis* consists in assuming that the DNA sequences are not equally distributed in terms of classification into binding sites to different TFs. Each of the 3 datasets is treated independently. The validation set is the last 340 samples of the 2000 training set corresponding to each TF. Figure 3, Figure 4, Figure 5, Figure 6, Figure 7 and Figure 8 show the validation accuracy under this hypothesis.

5 Ensemble

I use a simple ensemble technique to improve the prediction results. By aggregating the results of several candidate classifiers for each specific dataset, candidate classifiers are selected according to performance on validation sets, the aggregation is done by majority votes.

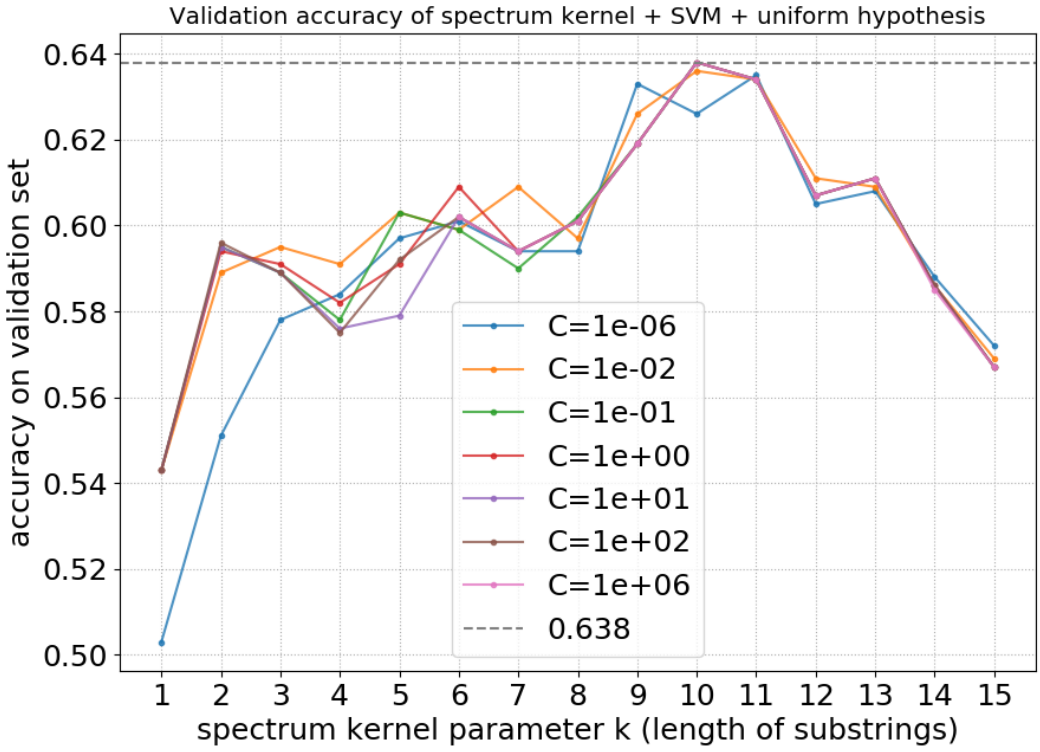

Figure 2: validation accuracy: spectrum kernel, C-SVM, uniform hypothesis.

6 Final results

My 2 final submissions are obtained from the following settings:

- Submission 1
  - dataset 0: kernel logistic regression $\lambda = 10^{-4}$, spectrum kernel $k = 9$
  - dataset 1: kernel logistic regression $\lambda = 10^{-4}$, spectrum kernel $k = 9$
  - dataset 2: kernel logistic regression $\lambda = 10^{-4}$, spectrum kernel $k = 9$
- Submission 2
  - dataset 0, ensemble:
    - kernel logistic regression $\lambda = 10^{-4}$, spectrum kernel $k = 9$
    - kernel logistic regression $\lambda = 10^{-2}$, spectrum kernel $k = 12$
    - kernel SVM $C = 1$, spectrum kernel $k = 12$
  - dataset 1, ensemble:
    - kernel logistic regression $\lambda = 10^{-4}$, spectrum kernel $k = 9$
    - kernel logistic regression $\lambda = 10^{-3}$, spectrum kernel $k = 9$
    - kernel SVM $C = 10^{-2}$, spectrum kernel $k = 8$
  - dataset 2, ensemble:
    - kernel logistic regression $\lambda = 10^{-4}$, spectrum kernel $k = 9$
    - kernel logistic regression $\lambda = 10^{-4}$, spectrum kernel $k = 8$
    - kernel SVM $C = 10^{-2}$, spectrum kernel $k = 12$

With these settings, I achieved the accuracy 0.71800 on the public leader board.

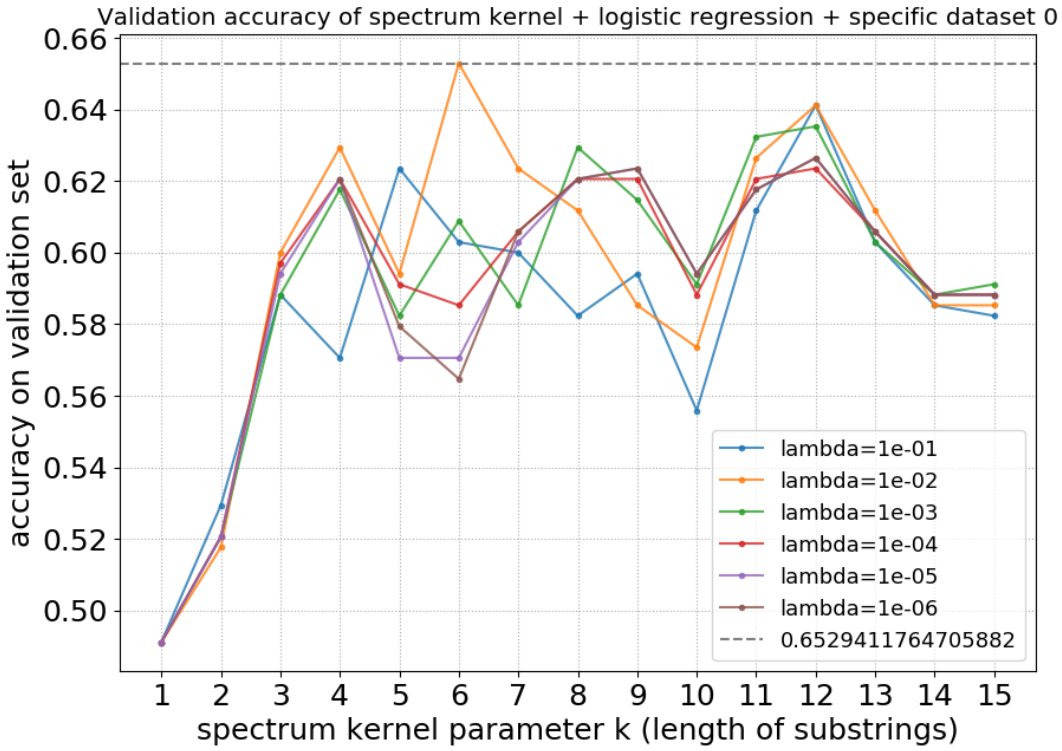

Figure 3: validation accuracy: spectrum kernel, logistic regression, specific hypothesis, dataset 0

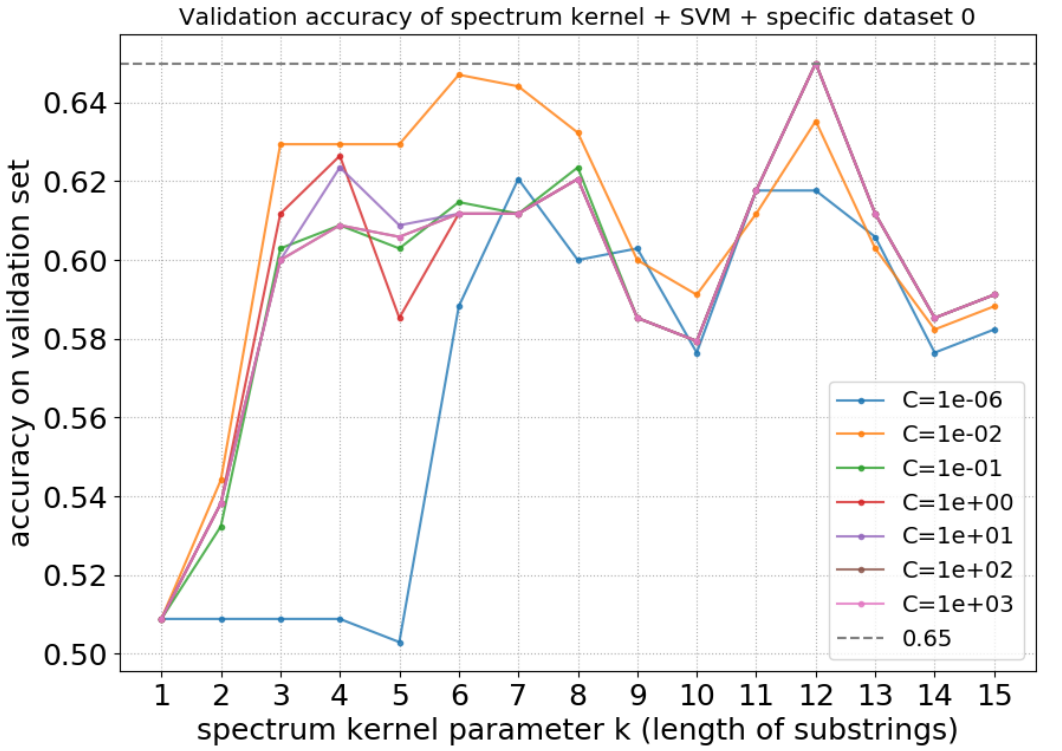

Figure 4: validation accuracy: spectrum kernel, C-SVM, specific hypothesis, dataset 0

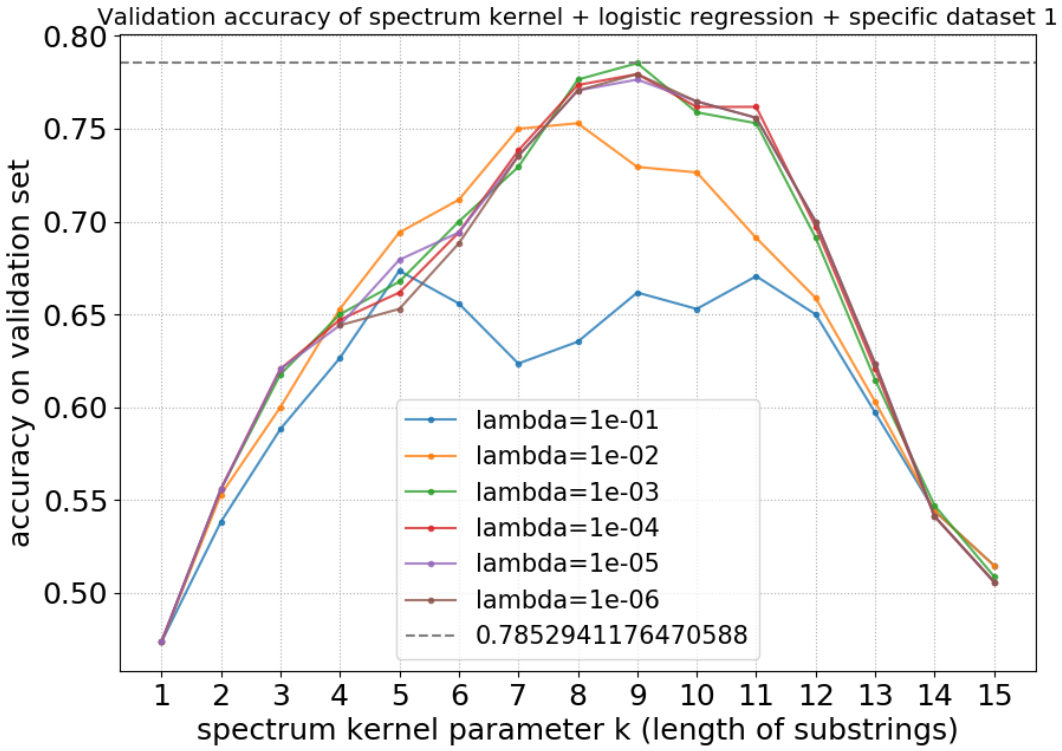

Figure 5: validation accuracy: spectrum kernel, logistic regression, specific hypothesis, dataset 1

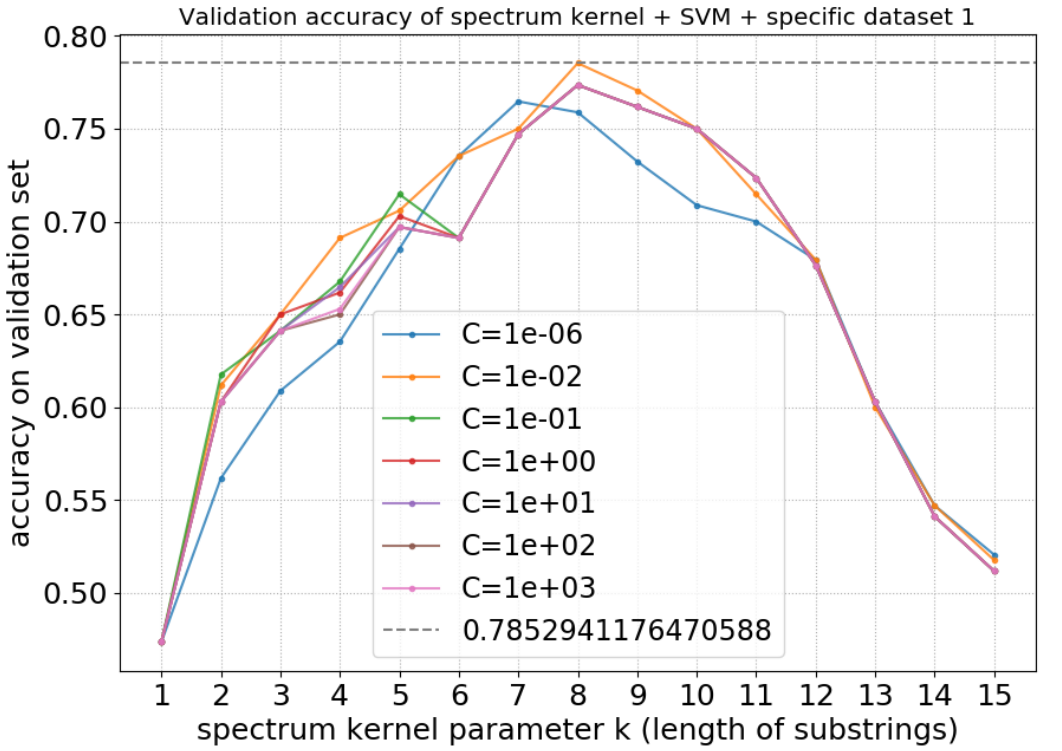

Figure 6: validation accuracy: spectrum kernel, C-SVM, specific hypothesis, dataset 1

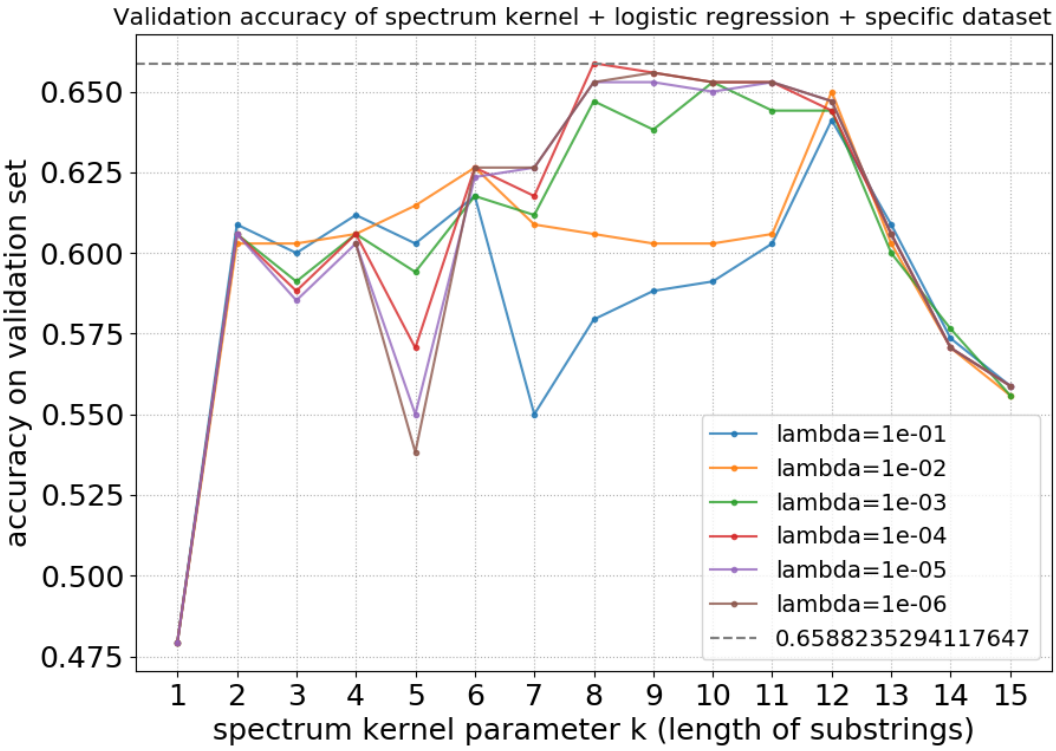

Figure 7: validation accuracy: spectrum kernel, logistic regression, specific hypothesis, dataset 2

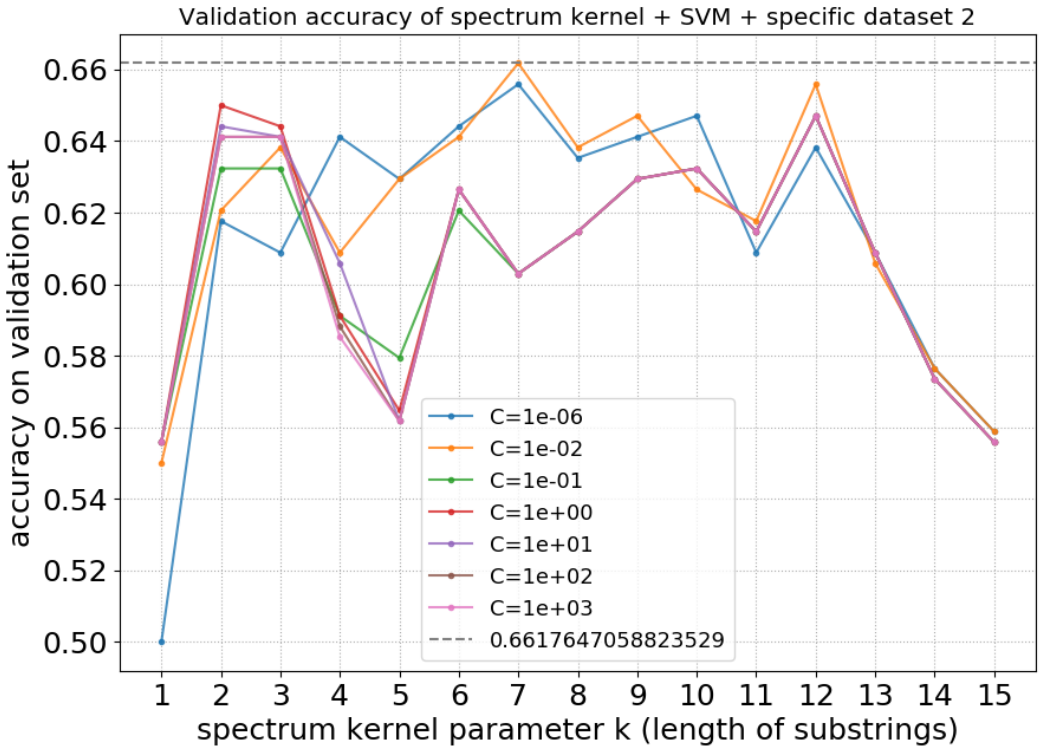

Figure 8: validation accuracy: spectrum kernel, C-SVM, specific hypothesis, dataset 2

References

[1] Hiroto Saigo, Jean-Philippe Vert, Nobuhisa Ueda, and Tatsuya Akutsu. Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11):1682–1689, 02 2004.

[2] Mark Maienschein-Cline, Aaron R. Dinner, William S. Hlavacek, and Fangping Mu. Improved predictions of transcription factor binding sites using physicochemical features of dna. In *Nucleic acids research*, 2012.

[3] Tianyin Zhou, Lin Yang, Yan Lu, Iris Dror, Ana Carolina Dantas Machado, Tahereh Ghane, Rosa Di Felice, and Remo Rohs. Dnashape: a method for the high-throughput prediction of dna structural features on a genomic scale. In *Nucleic Acids Research*, 2013.

[4] Wenxiu Ma, Lin Yang, Remo Rohs, and William Stafford Noble. Dna sequence+shape kernel enables alignment-free modeling of transcription factor binding. In *Bioinformatics*, 2017.