Influence of Sequence Length in Promoter Prediction Performance*

Sávio G. Carvalho, Renata Guerra-Sá, and Luiz H. de C. Merschmann

Federal University of Ouro Preto (UFOP), Ouro Preto/MG, Brazil
saviogcarvalho@yahoo.com.br,
{rguerra,luizhenrique}@iceb.ufop.br

Abstract. The advent of rapid evolution on sequencing capacity of new genomes has evidenced the need for data analysis automation aiming at speeding up the genomic annotation process and reducing its cost. Given that one important step for functional genomic annotation is the promoter identification, several studies have been taken in order to propose computational approaches to predict promoters. Different classifiers and characteristics of the promoter sequences have been used to deal with this prediction problem. However, several works in literature have addressed the promoter prediction problem using datasets containing sequences of 250 nucleotides or more. As the sequence length defines the amount of dataset attributes, even considering a limited number of properties to characterize the sequences, datasets with a high number of attributes are generated for training classifiers. Once high-dimensional datasets can degrade the classifiers predictive performance or even require an infesible processing time, predicting promoters by training classifiers from datasets with a reduced number of attributes, it is essential to obtain good predictive performance with low computational cost. To the best of our knowledge, there is no work in literature that verified in a sistematic way the relation between the sequences length and the predictive performance of classifiers. Thus, in this work, sixteen datasets composed of different sized sequences are built and evaluated using the SVM and k-NN classifiers. The experimental results show that several datasets composed of shorter sequences acheived better predictive performance when compared with datasets composed of longer sequences and consumed a significantly shorter processing time.

1 Introduction

Over recent years, advances in technology have allowed an acceleration of new genomes sequencing [9], evidencing the increasing demand for data analysis automation and for improving procedures previously used [2]. This has encouraged studying and implementing several computational techniques and creating new tools to enable processing of large amounts of genomic data.

One of the first steps for functional genomic annotation is promoter identification. Promoters are regions responsible for signaling and controlling the exact

* This research was partially supported by CNPq, FAPEMIG and UFOP.

S. Campos (Ed.): BSB 2014, LNBI 8826, pp. 41–48, 2014.

position where the transcription mechanism initiates, called TSS (*Transcription Start Site*). The capability for detecting them in their different forms will make it possible to understand how, where and when transcription occurs, in addition to providing clarification on the interaction network and regulation of the transcription mechanism [8,9].

The identification of promoter sequences in genomes can be seen as a classification problem, where, given the features of a genomic sequence, it would be classified as promoter or non-promoter. Therefore, several computational approaches to predict promoters have been proposed using different classification techniques and different types of information extracted from sequences. Nevertheless, further progress is needed to improve them [14,1,6,9].

Much of the complexity of promoter prediction problem is due to their diverse nature, which makes it difficult to identify them [12,8,10]. Therefore, a crucial step for prediction success is to discover features of promoter sequences that are relevant to differentiate them from non-promoter sequences.

In the search for relevant features to distinguish between promoter and non-promoter sequences, several properties of sequences have been tested for their predictive capability. According to [14], a prediction strategy can use three types of features: structural, based on signs and based on context. Several studies have shown that in order to build accurate models to predict or describe genomic processes, the structural properties of the DNA molecules must be considered [11]. Thus, the structural properties have been widely used in recent years [14] and have also been adopted for this work.

Despite the large amount of work involving promoter prediction [12,8,1,2,6,7,9], to the best of our knowledge, none of them verified in a systematic way the relation between the length of sequences used for training classification models and their predictive performance. Thus, the aim of this work is to evaluate, through the application of classification techniques, the effect of the sequence length in discrimination between promoters and non-promoters.

The importance of this evaluation is due to the fact that, considering the structural properties, the longer the sequences used to compose datasets used for training classifiers, the greater the amount of attributes. The problem is that high-dimensional datasets, that is, with great number of attributes, make the classification a more complex process, and the result may be an increase in classifiers training time and a reduction of their predictive perfomance.

Due to the amount of data available and the attention it has received from the scientific community in recent decades [14], the genome chosen to be studied in this work was *Homo sapiens*. The experiments were conducted using a well known and reliable promoter database which is publicly available on the web.

2 Our Approach

For the studies conducted in this work, promoter and non-promoter sequences derived from human genome were used for datasets construction.

Promoters were obtained from a set of sequences available in the DBTSS database [13], version 8.0. DBTSS, which has already been used in several other works [6,8,9,2], is a set of approximately 98,000 experimentally validated promoter sequences with active TSS, where each sequence has 1201 bp (base pairs).

Non-promoters correspond to several genomic sequences that were extracted randomly from intergenic regions and from introns and exons [6]. The criteria for obtaining these sequences require that the region is at a minimum distance of 1000 nucleotides from the positions demarcated on CAGE database, indicating transcription regions, and at a minimum distance of 1000 nucleotides from the positions demarcated on RefSeq that has information denoting the beginning of genes. Thus, the selection of false non-promoter sequences is avoided. CAGE and RefSeq databases were obtained from *pppBenchmark* tool [16] website[1].

Due to computational cost to process a sequence dataset, only part of the sequences available at DBTSS database was used in the composition of the datasets of this study. Thus, a total of 7000 different promoter sequences were chosen randomly, avoiding the inclusion of noisy sequences. In addition, other 7000 non-promoter sequences complete the datasets.

Therefore, all datasets used in this work are composed of the same 14000 sequences. However, the length of sequences varies from one dataset to another. For example, the dataset called 250-50 consists of sequences represented by 301 nucleotides. For promoter sequences, this size is the sum of the number of nucleotides positioned upstream and downstream of TSS (in addition to TSS itself), that is, in the example there are 250 nucleotides upstream and 50 nucleotides downstream of TSS. Therefore, for the same dataset, TSS is always located at the same position in all promoter sequences. Since non-promoter sequences do not have TSS, their length is simply given by their number of nucleotides. Thus, in 250-50 dataset, non-promoter sequences are also composed of 301 nucleotides.

Each dataset sequence is characterized by a set consisting of 13 structural properties [11], named: A-philicity, base stack energy, B-DNA, bendability, DNA-bending stiffness, disrupt energy, DNA denaturation, free energy, nucleosome positioning, propeller twist, protein deformation, protein-DNA twist and Z-DNA. These properties, which have already been subject of other studies in literature [7,1,9], are physico-chemical and conformational properties.

Since the structural properties may be determined by local interactions among neighboring nucleotides in a sequence [11], they are represented by tables where each possible nucleotide combination is associated with a value that represents its contribution to a particular structural property. As an example, Table 1 presents the mapped values of oligonucleotides for the stacking energy structural property.

Using these 13 structural properties, each nucleotide sequence (promoters and non-promoters) is converted into a numerical vector that characterizes it. Figure 1 illustrates the conversion of a sequence to two structural properties (protein deformation and nucleosome positioning). As it can be observed, the numerical vector of each property (structural profile) is obtained from scanning the sequence of nucleotides where, depending on the structural property, each vector

[1] Available at `http://bioinformatics.psb.ugent.be/webtools/pppbenchmark/`

Table 1. Mapped values of oligonucleotides for *base stack energy* property [3]

Oligonucleotides	Value (kcal/mole)	Oligonucleotides	Value (kcal/mole)
AA	-5.37	GA	-9.81
AC	-10.51	GC	-14.59
AG	-6.78	GG	-8.26
AT	-6.57	GT	-10.51
CA	-6.57	TA	-3.82
CC	-8.26	TC	-9.81
CG	-9.69	TG	-6.57
CT	-6.78	TT	-5.37

value is obtained considering sequences of dinucleotides (protein deformability) or trinucleotides (nucleosome positioning).

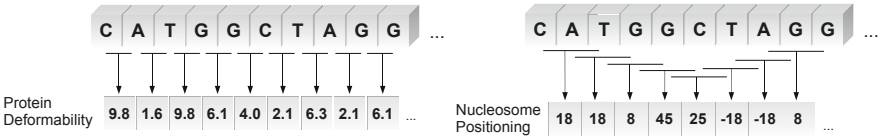


Fig. 1. Conversion of a sequence to two structural properties

Considering the conversion schema previously mentioned, in order to show the capability of the structural properties to discriminate promoter from non-promoter sequences, Figure 2 illustrates, for two structural properties, the average structural profile of promoter and non-promoter sequences of the 250-50 dataset. In this figure, TSS is located at the 0 position.

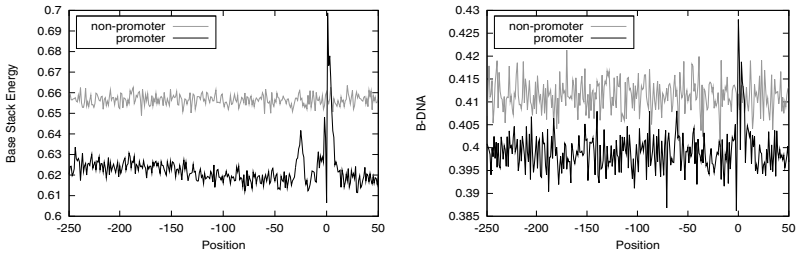


Fig. 2. Structural profiles for the *250-50* dataset

The complete characterization of a sequence is given by a single numerical vector resulting from the junction of the vectors representing each of the 13 structural properties considered in this work. The size of the resultant vector of these junctions corresponds to the number of predictor attributes of the datasets

used for classifiers training. In addition to these predictor attributes, each sequence has a value for the class attribute, which indicates whether that sequence is promoter or non-promoter. As an example, the largest dataset used in our experiments, the 250-50 one, results in a set of 3898 predictor attributes. Table 2 shows the number of predictor attributes for each dataset used in this work.

Table 2. Predictor attributes for each dataset

Dataset	Number of atributes	Dataset	Number of atributes
10-1	141	20-50	908
10-3	167	30-50	1038
10-5	193	40-50	1168
10-10	258	50-50	1298
10-20	388	100-50	1948
10-30	518	150-50	2598
10-40	648	200-50	3248
10-50	778	250-50	3898

As it can be observed in Table 2, the length of sequences used to compose the dataset defines the amount of their attributes. Several studies in literature have addressed the problem of promoter prediction using datasets containing sequences of 250 nucleotides or more [12,2,8,9]. Although a limited amount of features is being used in characterization of sequences, high-dimensional datasets are generated for classifiers training.

The problem with high-dimensional datasets, that is, with high number of attributes, is that they make classification a more complex process, often consuming an infeasible time for training classifiers and degrading their predictive performance.

Therefore, to predict promoters by training classifiers from datasets with a reduced number of attributes, it is essential to obtain good predictive performance with low computational cost. This way, the objective of the experiments conducted in this work is to evaluate the impact of the sequence length variation on the classifiers performance.

3 Computational Experiments

3.1 Classifiers and Experimental Setup

SVM (Support Vector Machine) and k-NN (k-Nearest Neighbours) classifiers, usually adopted in data mining works, were chosen to evaluate the impact of the sequence length variation on the performance of predictive models. Experiments were conducted using the caret package (short for *classification and regression training*) in R [15], which is a programming language and an environment widely used in statistical and graphics computation for data analysis.

k-NN classifier's idea is very simple. Each dataset instance is described by an n-dimensional vector, where k corresponds to the number of predictor attributes.

To classify a new instance (an instance whose class is unknown), the classifier uses distance metrics to determine the k training instances that are more similar to the instance to be classified. Then, the most frequent class among the similar k instances is attributed to the new instance. In k-NN, the k value is an input parameter.

Considering each dataset instance as a point in n-dimensional space, the basic idea of SVM is to find a hyperplane with maximum margin of separation, ie, one that provides the separation of training instances, with maximum margin, in two portions in n-dimensional space. Once the optimal hyperplane is found, the classification of a new instance is made by determining its position in relation to the separation hyperplane. Although this method was originally proposed for binary classification problems, several extensions have been proposed in literature to make it suitable for multi-class classification problems.

In order to set the algorithms parameters for the dataset used in this study, experiments were conducted by varying the parameters values C (0.25, 0.5, 1, 2, 4), *gamma*([0.1, 0.0001]), for SVM (using RBF kernel) and k (1, 3, 5, 7, 9) for k-NN. Table 3 presents the best parameter values obtained for each dataset and therefore used in our experiments to obtain the results presented here. All experiments were carried out on a Core i7-2600 @ 3.40GHz PC with 12 GBytes of RAM.

Table 3. k-NN e SVM parameters

	k-NN	SVM	
Dataset	k	C	*gamma*
10-1	9	1	3.64e-03
10-3	9	0.5	3.05e-03
10-5	9	0.5	1e-02
10-10	9	2	1e-03
10-20	9	1	1e-03
10-30	9	1	1e-03
10-40	9	1	7.84e-04
10-50	9	1	1e-03

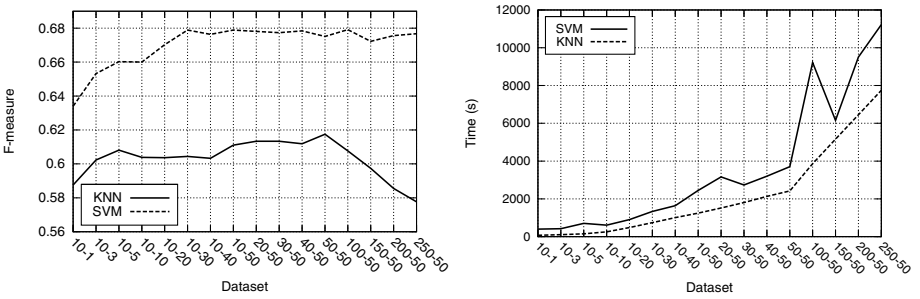
	k-NN	SVM	
Dataset	k	C	*gamma*
20-50	9	1	1e-03
30-50	9	0.5	1e-03
40-50	9	0.5	1e-03
50-50	7	0.5	1e-03
100-50	9	1	1e-03
150-50	9	0.5	1.96e-04
200-50	9	1	1.56e-04
250-50	9	1	1e-04

3.2 Experimental Results

The classifiers predictive performance was measured using k-cross-validation (k =10) and F-measure metric. For each dataset, the same test partitions were used in the evaluation of classifiers.

The results of the experiments are presented in Figure 3 graphs. Figure 3(a) graph shows the predictive performance of SVM and k-NN classifiers for each of the 16 datasets evaluated. Figure 3(b) graph shows the processing time spent in the classification process for these datasets.

As it can be seen in Figure 3(a) graph, the SVM classifier obtained better predictive performance than the k-NN one for all datasets evaluated.



(a) (b)

Fig. 3. (a) Average F-measure and (b) processing time

Yet, the most important thing to observe in Figure 3(a) graphs is that for both classifiers, the decrease in the length of sequences used in the datasets did not necessarily imply a reduction in their predictive performance. SVM performance remained relatively stable for datasets composed of sequences ranging in size from 301 (250-50) to 41 (10-30) nucleotides, presenting a marked degradation in performance only for sequences containing less than 41 nucleotides. k-NN achieved its best performance with the 50-50 dataset and, even for the dataset composed of shorter sequences (10-1), presented superior predictive performance compared with larger datasets (250-50).

Figure 3(b) graph shows that, for both classifiers, time spent for processing datasets grows exponentially with the increase of the length of sequences that compose them. It is worth noting that in many cases, a dataset composed of shorter sequences achieves superior predictive performance compared with longer sequence datasets and time spent in processing is significantly shorter than that consumed by longer sequence datasets. For example, for SVM, the 10-30 dataset presents predictive performance slightly higher than that achieved by the 250-50 dataset and time spent in processing is more than 8 times shorter than that spent by the 250-50 dataset.

4 Conclusion

Promoter prediction is a fundamental step for genome functional annotation and, therefore, several computational approaches have been proposed using different classification techniques. However, to best of our knowledge, none of them verified in a systematic way the relation between the length of sequences used for training classification models and their predictive performance. This way, experiments were conducted to analyze the impact of the sequence length variation on the classifiers performance.

In order to perform the analysis previously mentioned, 16 datasets composed of different sized sequences were generated and evaluated using the SVM and k-NN classifiers. The experimental results show that the decrease in the length

of sequences used in the composition of the datasets did not necessarily result in a reduction of the classifiers predictive performance. In addition, several bases composed of shorter sequences achieved superior predictive performance compared with datasets composed of longer sequences and consumed a significantly shorter processing time.

As future work, we plan to apply techniques for selecting attributes in datasets generated in this study aiming at reducing the datasets number of attributes and improving classifiers predictive performance.

References

1. Abeel, T., Saeys, Y., Bonnet, E., Rouzé, P., Van de Peer, Y.: Generic eukaryotic core promoter prediction using structural features of dna. Genome Research 18(2), 310–323 (2008)
2. Abeel, T., Saeys, Y., Rouzé, P., Van de Peer, Y.: Prosom: core promoter prediction based on unsupervised clustering of dna physical profiles. Bioinformatics 24(13), i24–i31 (2008)
3. Baldi, P., Brunak, S., Chauvin, Y., Pedersen, A.G.: Computational applications of dna structural scales. In: Glasgow, J.I., Littlejohn, T.G., Major, F., Lathrop, R.H., Sankoff, D., Sensen, C. (eds.) ISMB, pp. 35–42. AAAI (1998)
4. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
5. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
6. Dineen, D., Schroder, M., Higgins, D., Cunningham, P.: Ensemble approach combining multiple methods improves human transcription start site prediction. BMC Genomics 11(1), 677 (2010)
7. Florquin, K., Saeys, Y., Degroeve, S., Rouzé, P., Van de Peer, Y.: Large-scale structural analysis of the core promoter in mammalian and plant genomes. Nucleic Acids Research 33(13), 4255–4264 (2005)
8. Gan, Y., Guan, J., Zhou, S.: A pattern-based nearest neighbor search approach for promoter prediction using dna structural profiles. Bioinf. 25(16), 2006–2012 (2009)
9. Gan, Y., Guan, J., Zhou, S.: A comparison study on feature selection of dna structural properties for promoter prediction. BMC Bioinformatics 13(1), 4 (2012)
10. Grishkevich, V., Hashimshony, T., Yanai, I.: Core promoter t-blocks correlate with gene expression levels in c. elegans. Genome Research 21(5), 707–717 (2011)
11. Meysman, P., Marchal, K., Engelen, K.: DNA structural properties in the classification of genomic transcription regulation elements. Bioinformatics and Biology Insights 6, 155–168 (2012)
12. Ohler, U., Niemann, H., Liao, G.C., Rubin, G.M.: Joint modeling of dna sequence and physical properties to improve eukaryotic promoter recognition. Bioinformatics 17(suppl. 1), S199–S206 (2001)
13. Yamashita, R., Sugano, S., Suzuki, Y., Nakai, K.: Dbtss: Database of transcriptional start sites progress report in 2012. Nucleic Acids Res. 40(D1), 150–154 (2012)
14. Zeng, J., Zhu, S., Yan, H.: Towards accurate human promoter recognition: a review of currently used sequence features and classification methods. Briefings in Bioinformatics 10(5), 498–508 (2009)
15. Kuhn, M., Johnson, K.: Applied Predictive Modeling. Springer (2013)
16. Abeel, T., Van de Peer, Y., Saeys, Y.: Toward a gold standard for promoter prediction evaluation. Bioinformatics 25(12), i313–i320 (2009)