Comments to the Editor

I evaluated this study in terms of novelty, technical quality and relevance. The methods applied are very sound but I have major concerns on the other topics.

My major concerns can be summed in the major comments 1-5. First, the method is not compared to state-of-the-art competitors under the same evaluation criteria, which makes hard to assess the significance of the proposed scheme. Second, the negative examples need to be extended to include a more realistic scenario. Third, the relevance of the paper is impacted by restricting the framework to human organism only. Fourth, experimental validation on cell-specific data was not performed although it is fairly easy to make such analysis with the data and technology available today. Last, statistical significance of performance was not presented, which makes performance assessment impossible.

Furthermore, the print quality of figures and the mathematical formalization need to be extensively reviewed. In a print version of the paper it is almost impossible to visualize the figures.

Finally, I would recommend the rejection based on the fact that the results are too preliminary and speculative, given the major critiques presented.

Comments to Author

In this study the authors used a novel framework to identify promoter regions in humans. In this framework, DNA features such as CG content, DNA rigidity, Kullback-Leibler divergence on k-mer frequencies and CpG islands are used with multiple support vector machines (SVMs). Training and testing are performed on data from well-known repositories and the performance is evaluated using sensitivity, specificity and averaged conditional probability. The methodology is sound, however I have some major points regarding this study.

a. Major

1. Although the authors state multiple competing methods in Section I, they are not compared with the proposed algorithm. The comparison with previous state-of-the-art methods is an essential part of methodological studies.

2. In this study, it was considered exons, introns and 3'UTR as the "negative (non-promoter) data sets". Other negative datasets are absolutely necessary. The reason for this is because on a non-controlled genome-wide application of such algorithm, distal regulatory modules would pose a big problem, since they share many DNA features, especially with regard to nucleotide frequencies (they are also bound by transcription factors). Furthermore, control background sequences such as generated by 1st-order Markov models or simply random nucleotide words with same length and nucleotide distribution as the promoter sequences are absolutely necessary. If these issues are not considered I wonder if the accuracies shown are overly optimistic since promoter regions have known differences when compared to the genomic subset of regions evaluated as negative data set.

3. The study focused on human. However, it was previously shown that such task can be generalized for eukaryotes (Abeel 2008). The inclusion of further experiments on different organisms would increase the relevance of the paper. Alternatively, showing that such framework can be applied on a generalized manner would be enough.

4. Recently, a major problem is the identification of cell-specific regulatory regions. In this sense I believe a robust promoter prediction algorithm, especially in eukaryotes, should consider alternative promoters. Many repositories like ENCODE (ENCODE Project Consortium 2012) provide a comprehensive data collection of Pol2 binding. This would result in additional experiments that would really verify the strength of the approach proposed.

5. Many claims are made about performance being higher/lower than other performance, but no statistical significance is shown. Statistical tests to address the significance of performance increase/decrease are absolutely necessary. A simple t-test is a good exploratory test. However, when performances over multiple datasets or parameter considerations are being compared, a more sophisticated approach such as the Friedman-Nemenyi test is required.

6. The evaluation of Kullback-Leibler divergence in an optimization problem can be computationally intensive as the number k (from k-mer) gets higher. Do the authors have an estimate of that? It was used k-mers of length 4, maybe an appropriate parameter tuning experiment should be performed in order to find an interesting k value which improves the performance over k-1 but in which k+1 does not significantly outperform k. I believe a more efficient strategy such as estimating the k-mer distribution with dynamic programming is a better way to address this issue.

7. SVM parameter tuning using cross-validation is generally a good approach. However, different kernel functions should be tested such as linear, polynomial, sigmoid, hyperbolic tangent. A good contribution would be to state the reason to which RBF (or other) kernel function did well in this particular problem.

8. Since the algorithm was proposed for only one organism. I would expect to see the results on different gene data sets such as ensembl, refseq and UCSC. It would be very interesting to observe the prediction power variation between different set of genes.

9. What was the threshold used in Eq.4. Why was that threshold selected?

10. It is not clear why the authors selected a trinucleotide model to calculate rigidity.

b. Minor

1. None of the figures are in vectorial format. The (print) quality of all figures is very poor.

2. The mathematical notation is not very well defined. For instance, the rigidity values are written with both nucleotide and index subscripts. Also, the Eq. 3 seems to be redundant given Eq.2. Mathematical notation should be improved for clarity and correctness.

3. Other DNA features have also been considered very important such as DNA denaturation and energy-related features (Gan 2012). These features should be included for a more comprehensive evaluation.

4. It was claimed that SVM is the best current algorithm for promoter prediction. Is there a reference for that? Is there a case study in which different algorithms were tested under the same conditions?

5. Provide a citation for the first sentence in section II.A.1 (page 2).

6. It seems to me that some terms like "phosphodiesterase-linked cytosine and guanine" could be simplified since it is not a biological paper. The example given could be written as "C followed by G on the same DNA strand".

7. The definition given for CpG island is one of many available in the literature. A direct source citation would be preferred.

8. Page 5, Section III, Paragraph 1: Should the website be stated in the "References" section and cited?

9. Results are shown in 3D and 2D bar plots and tables. I believe it would increase clarity if the result plots were standardized.

10. Page 6, Section III.D, Paragraph 2: The statistical interpretation of results in this paragraph is very simplistic. I suggest the re-writing of this paragraph.

11. Fig. 4 have repeating labels (such as "Sn") that could be removed.

12. References are not standardized. If they were generated automatically, it is important to have them manually verified.

13. I would suggest using the latex command $^\prime$ to denote "prime", as in 3’UTR (in the case Latex was used to create the document)

14. The manuscript has many grammar/clarity problems. Please review the text for grammar/clarity issues. Some examples:

- Page 1, Section I, Paragraph 1: "Thus it is a very important task that how to quickly(...)"

- The word "document" is used throughout the text as synonym of "string" or "word". I believe "string" or "word" is preferred.

- Page 2, Section II.A.1, Paragraph 2: "(...)rigidity parameter of each nucleotide at position i(...)" should be "(...)rigidity parameter of each nucleotide starting at position i(...)".

- Page 2, Section II.A.2, Paragraph 2: "(...)large number of zero in frequency(...)".

- Page 2, Section II.A.2, Paragraph 3: "(...)the frequency of n-mers in there kinds of(...)" should be "(...)the frequency of n-mers in three kinds of(...)".

- Page 4, Section II.C.2.a, Paragraph 4: Change "exton" by "exon". Also the same sentence in which this error happens appears twice, next to each other, in the manuscript.

- Page 6, Section III.C, Paragraph 2: "Then we have we get"

- Page 6, Section III.C, Paragraph 3: "feaures"

References:

[Abeel 2008] Generic eukaryotic core promoter prediction using structural features of DNA. Genome research, 18(2), 310-323.

[ENCODE Project Consortium 2012] An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414), 57-74.

[Gan 2012] A comparison study on feature selection of DNA structural properties for promoter prediction. BMC Bioinformatics, 13(1), 4+.