METHOD ARTICLE

Prediction of fine-tuned promoter activity from DNA sequence [version 1; referees: 1 approved, 2 approved with reservations]

Geoffrey Siwo[1,2,4-6], Andrew Rider[1,3,4], Asako Tan[1,2,7], Richard Pinapati[1,2,4], Scott Emrich[1,3,4], Nitesh Chawla[1,3,4], Michael Ferdig[1,2,4]

[1]Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN, USA
[2]Department of Biological Sciences, University of Notre Dame, Notre Dame, IN, USA
[3]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA
[4]Interdisciplinary Center for Network Science and Applications (iCeNSA), University of Notre Dame, Notre Dame, IN, USA
[5]IBM TJ Watson Research Center, NY, USA
[6]IBM Research-Africa, Johannesberg, South Africa
[7]Epicentre, Madison, WI, USA

v1 **First published:** 11 Feb 2016, **5**:158 (doi: 10.12688/f1000research.7485.1)
Latest published: 11 Feb 2016, **5**:158 (doi: 10.12688/f1000research.7485.1)

Abstract

The quantitative prediction of transcriptional activity of genes using promoter sequence is fundamental to the engineering of biological systems for industrial purposes and understanding the natural variation in gene expression. To catalyze the development of new algorithms for this purpose, the Dialogue on Reverse Engineering Assessment and Methods (DREAM) organized a community challenge seeking predictive models of promoter activity given normalized promoter activity data for 90 ribosomal protein promoters driving expression of a fluorescent reporter gene. By developing an unbiased modeling approach that performs an iterative search for predictive DNA sequence features using the frequencies of various k-mers, inferred DNA mechanical properties and spatial positions of promoter sequences, we achieved the best performer status in this challenge. The specific predictive features used in the model included the frequency of the nucleotide G, the length of polymeric tracts of T and TA, the frequencies of 6 distinct trinucleotides and 12 tetranucleotides, and the predicted protein deformability of the DNA sequence. Our method accurately predicted the activity of 20 natural variants of ribosomal protein promoters (Spearman correlation r = 0.73) as compared to 33 laboratory-mutated variants of the promoters (r = 0.57) in a test set that was hidden from participants. Notably, our model differed substantially from the rest in 2 main ways: i) it did not explicitly utilize transcription factor binding information implying that subtle DNA sequence features are highly associated with gene expression, and ii) it was entirely based on features extracted exclusively from the 100 bp region upstream from the translational start site demonstrating that this region encodes much of the overall promoter activity. The findings from this study have important implications for the engineering of predictable gene expression systems and the evolution of gene expression in naturally occurring biological systems.

Open Peer Review

Referee Status: ✓ ? ?

	Invited Referees 1	2	3
version 1 published 11 Feb 2016	✓ report	? report	? report

1 **Paul Pavlidis**, University of British Columbia Canada

2 **Jianhua Ruan**, The University of Texas at San Antonio USA

3 **Jan Grau**, Martin Luther University of Halle-Wittenberg Germany

Discuss this article

Comments (0)

This article is included in the DREAM Challenges channel.

Corresponding author: Geoffrey Siwo (siwomolbio@gmail.com)

How to cite this article: Siwo G, Rider A, Tan A *et al.* **Prediction of fine-tuned promoter activity from DNA sequence [version 1; referees: 1 approved, 2 approved with reservations]** *F1000Research* 2016, **5**:158 (doi: 10.12688/f1000research.7485.1)

Grant information: The author(s) declared that no grants were involved in supporting this work.

Competing interests: The authors declare that they have no competing interests.

First published: 11 Feb 2016, **5**:158 (doi: 10.12688/f1000research.7485.1)

Introduction

Transcription is a fundamental step in the decoding of information encoded in DNA into phenotypes. Therefore, knowledge of transcriptional regulation is crucial for understanding the natural variation of gene expression[1–5] and for the accurate engineering of predictable gene expression systems[6–8]. While transcriptional regulation is one of the most highly studied areas in biology, the ability to quantitatively predict gene expression from DNA sequence remains inadequate[9,10]. Knowledge of transcription factors and their cognate binding sites continues to grow and has enhanced our ability to make qualitative predictions about gene expression. For example, a number of transcription factors are now well known to be involved in differentiation of stem cells into specific cell types, leading to potentially clinically useful applications such as induced pluripotent stem cells[11]. Inspite of this progress, only limited quantitative predictions of gene expression are possible[6–8,12,13]. Knowledge that promoter sequences of genes encode both qualitative (e.g. when to switch a gene on and off) and quantitative properties (e.g. precise levels and noise) of gene expression is implied by the heritable nature of these attributes[1–3,14]. It is becoming increasingly clear that while transcription factors are critical in gene regulation, regulatory outputs are ultimately determined by co-operation between regulators in complex circuits[15–17] and with chromatin states[18–21]. In particular, transcription factors compete for DNA binding sites with nucleosomes[22,23]. The information for nucleosome binding is largely encoded in the DNA sequence[24–27], even though *in vivo* nucleosome occupancy is highly dynamic[25,28,29]. Quantitative models of gene expression, therefore, benefit from the integration of nucleosome and transcription factor binding data[10,23,30].

A key barrier to quantitative modeling of gene expression using promoter sequence has been the lack of experimental methods for accurately measuring transcript levels. DNA microarrays and RNA-seq are the most widely-used systems for measuring transcript abundance, but this measurement can reflect many effects including promoter sequence, genomic position of a gene and post-transcriptional regulation of mRNA levels by processes like mRNA degradation. In addition, microarray and RNA-seq can be affected by systematic biases arising from sequence dependent hybridization kinetics[31] and sequence dependent read-depth coverage[32], respectively. To overcome these limitations, approaches based on promoters fused to fluorescent reporters have been developed to generate direct, real-time measurement of promoter activity with high accuracy[33]. This has been applied in large libraries of synthetic bacterial promoters thereby generating new insights on combinatorial cis-regulation[8]. It was not until recently that the first large-scale library of naturally occurring promoters of any eukaryote fused to yellow fluorescent protein (YFP) became available[30]. 110 yeast ribosomal protein (RP) promoters were fused to YFP and integrated into a different strain at a fixed genomic location, hence alleviating both post-translational and genomic context related effects[30]. Consequently, this data set is very well poised for the computational modeling of the relationship between promoter sequence and transcription activity of a eukaryotic promoter.

To provide a fair assessment of the relationship between promoter sequence and quantitative transcript levels, the Dialogue for Reverse Engineering Assessments and Methods (DREAM) organized an open community challenge in 2011 (details of the challenge as well as an overview of participating teams is provided in reference 34), inviting participants to address this question using promoter activities of the RP promoter library that was not yet published[30]. Participants were provided with the activities of 90 promoters and their corresponding promoter sequences and challenged to predict the activity of 53 promoters whose activities were known only to the organizers of the challenge (Figure 1A). After a period of three months, the challenge organizers independently assessed the performance of models from 21 teams using four different statistical tests. Our team, Fighting Irish Systems Team (FIrST), attained the best performance status on the basis of a combined score by the DREAM consortium in predicting the activities of these 53 promoters (Spearman correlation between predicted and actual activities r = 0.65, P = 0.002). Our approach was built upon three key

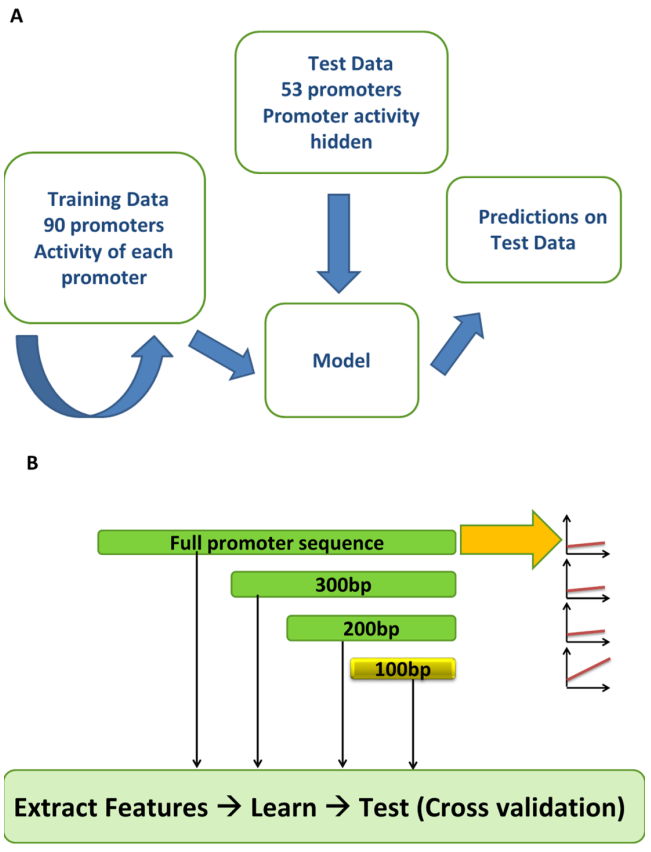

Figure 1. Summary of the DREAM6 gene expression challenge. (**A**) Training data consisted of DNA sequences for 90 yeast RP promoters whose activities were experimentally determined[30,34]. DNA sequences for blinded test set of 53 promoters whose activity was hidden also experimentally determined but withheld from the challenge participants was also provided. (**B**) Outline for strategy of modeling promoter activity. Each promoter was segmented into 100 bp non-overlapping windows with the full promoter regarded as a separate window. For each window, DNA sequence features were extracted and feature selection using a linear regression wrapper performed prior to machine learning. Performance of machine learning models trained on each window was determined in 5- and 10-fold cross-validations using Pearson correlation.

propositions: i) transcription factor binding and nucleosome binding, as well as other regulatory signals are encoded in DNA[9,10,12,27], ii) if i) is true, then explicit prior knowledge of transcription factor and nucleosome binding is not a mandatory prerequisite for prediction of promoter activity if training data is available. That is, an unbiased approach that explores the associations between DNA sequence patterns and promoter activity should be able to rediscover patterns that relate to the observed activity. To do this, we used machine learning methods to iteratively explore the association between promoter activity and DNA sequence patterns in 100 bp windows of promoter sequence. We considered sequence patterns such as k-mers (k = 1 to k = 5), homopolymer stretches, nucleosome binding and three mechanical properties of DNA (bendability[35], deformability[36] and stiffness[37]). Based on iterative exploration of different machine learning models, we established that a support vector machine (SVM) was the most predictive of promoter activity based on specific sequence patterns in the 100 bp upstream of the translation start site (TrSS). Our model outperformed those which applied transcription factor binding sites of known RP promoters[34], implying that other sequence patterns besides transcription factor binding sites can help in fine-tuning gene expression. Indeed, among the predictive features employed by our model were poly(dT-dA) tracts that occlude nucleosomes; these have since been applied to fine-tune gene expression beyond resolutions attainable by transcription factor site mutations[38]. Our study expands the understanding of sequence patterns that could potentially be useful in engineering fine-tuned gene expression.

Methods

DREAM6 challenge data

The training data composed of DNA sequence for 90 yeast RP promoters with known activities and a test data set of 53 promoters was downloaded from the DREAM challenge website (https://www.synapse.org//#!Synapse:syn2820426/wiki/71012). Details of promoter construction are available from Zeevi *et al.* 2011[30] and the DREAM website. Briefly, the organizers considered the promoter region as the sequence 1200 bp upstream of a gene or until the nearest gene. Each promoter was linked to a URA3 selection marker and inserted into the same fixed genomic location of a master yeast strain containing the *YFP* gene. In total, 110 natural RP promoter strains and 33 strains with synthetically mutated RP promoters were constructed. As a control for experimental variation, all these strains contained a control promoter (TEF2) driving the expression of red fluorescent protein (mCherry). The *mCherry*, *TEF2*, *URA3*, *RP* promoter and *YFP* were all a single contiguous DNA sequence arranged in that order. Measurements of the *mCherry* expression levels and replicates of promoters had very low variation, enabling the distinction between any two promoters with activities differing by as little as ~ 8%. The promoter activity was determined as the amount of YFP fluorescence produced during the exponential growth phase, divided by the integral of the OD during the same period. The promoter activity measures the average amount of YFP produced from each promoter, per cell, per second during the exponential phase.

Feature extraction

Each promoter sequence was divided into 100 bp non-overlapping windows. The full promoter sequence was considered as another window. To extract information from each of the windows, we considered the frequencies of specific sequences in k-mers (k = 1 to 5), length of homopolymeric stretches DNA, mechanical properties (deformability, bendability and stiffness) and nucleosome binding. K-mer counts were performed using custom scripts. DNA mechanical properties were computed using workflows constructed in the Taverna Workbench version 2.2.0[53] and BioMoby web-services (accessed in August 2011) imported from the Molecular Modeling and Bioinformatics Group, Barcelona, Spain[54]. Bendability was estimated based on trinucleotide parameters obtained from DNase I digestion and nucleosome binding data[35]. Deformability was based on parameters from the analysis of protein-DNA crystallography structures[36]. Bending stiffness was based on bending free energy using the near-neighbor model[37]. Nucleosome binding was based on trinucleotide preferences[55].

Feature selection

For each window, feature selection was performed using a linear regression wrapper in the WEKA machine learning toolkit version 3.4[56] to select feature combinations that are most predictive of promoter activity. Performance of feature combinations was tested using 5- and 10-fold cross validation.

Machine learning model exploration

Three models implemented in the WEKA toolkit[56] were considered: SVM regression using sequential minimal optimization (SMO), linear regression and regression trees. Models were trained using 66% of the data and tested using 34%, and included only the features that were selected as important by the linear regression wrapper. Performance was determined using Pearson correlation between model predictions and actual promoter activities computed in R version 2.11.1. The SVM model was selected for refinement based on high performance compared to the other models.

Application of SVM model to DREAM6 test set

Promoter activities were not available to the participants of the challenge. We applied the ensemble of 501 SVMs built from 500 different training/test sets in which 80% of the data was used in training and 20% in testing and a single SVM validated by 66% training set and 34% testing sets. Each SVM model utilized the 24 features selected by a linear regression wrapper as most predictive of promoter activity. To predict activities of the DREAM6 test set, the 24 features were extracted from the upstream 100 bp sequence for each promoter. Predictions were then made using each of the SVM models and averaged to obtain the final predictions.

Validation of model by DREAM6 consortium

Predictions from the SVM ensemble were submitted through the DREAM website to the organizers for a blinded evaluation on the test set. The DREAM organizers used four statistics and corresponding *P*-values to evaluate the performance on the test set[34]. Details of the equations used for these statistics have been published separately by the DREAM6 Promoter Prediction Consortium[34].

1. Pearson correlation between predicted and observed activities for each model submitted: To generate a *P*-value for observing a Pearson correlation coefficient of the same magnitude or smaller than that of a given participant, a null distribution was generated by randomly sampling predictions from other teams and repeating this 10,000 times[34].

2. Spearman correlation for participant between ranks of the predicted and actual ranks of promoter activities: A *P*-value was then generated using a null distribution obtained from randomly sampling the predictions made by the other participants. The process was repeated 10,000 times[34].

3. Chi-square distance metric measuring the distance between predicted and actual promoter activities: To generate a *P*-value for observing a chi-square distance metric of the same magnitude or smaller than that of a given model submission, a null distribution was generated by randomly sampling predictions from other teams and repeating this 10,000 times[34].

4. A rank distance metric measuring the difference in ranks between predicted ranks and actual ranks of promoter activities. A *P*-value was generated from a null distribution obtained by randomly sampling predicted ranks from other teams, repeating this 10,000 times.

The overall score was defined as the product of the four *P*-values[34]. All these scores were computed using R version 2.11.1.

Results

Dataset 1. Raw data for 'Prediction of fine-tuned promoter activity from DNA sequence', Siwo *et al.* 2016

http://dx.doi.org/10.5256/f1000research.7485.d113516

README.txt contains a description of the files.

Promoter activity is highly predictable using the 100 bp upstream region from TrSS

The challenge organizers provided DNA sequences and promoter activities - the average rate of YFP production from each promoter, per cell per second, during the exponential phase - for 90 RP promoters (training set) and another set of 53 promoters whose activity was withheld from participants (test set)[30]. We first partitioned the promoter sequences into 100 bp non-overlapping windows, extracted specific DNA features from each window and considered the full promoter sequence as its own window (Figure 1B). The features considered were k-mers (k = 1 to 5), length of homopolymeric stretches, nucleosome positioning and DNA mechanical properties (bendability, deformability and stiffness). For each window, we performed feature selection using a linear regression wrapper, then explored three different machine learning methods (SVM, linear regression and regression trees) to learn the association between features in the window and promoter activity (Figure 1B). The performance in each window was assessed by Pearson correlation using 5- and 10-fold cross-validations on the training data. We observed very poor correlation ($r \ll 0.5$) between predicted and actual promoter activities except when using the window comprising 100 bp from the TrSS. Therefore, we focused the SVM model on this window using 23 features (Table 1) selected by the linear regression wrapper. A test of this model on 1000 randomized splits of the data (66% training and 34% testing sets) gave an average Pearson correlation of 0.78. The performance of machine learning models can be biased by the training/test data set used. Therefore, to reduce this bias, we obtained an additional 500 SVM models trained on randomly sampled sets of 80% of the data and validated on the remaining 20%. In the DREAM test set (activities for this set were withheld from participants), we used the SVM models to make predictions for each promoter. For each promoter, the predicted activity was the average of predictions across all the ensemble of SVMs based only on the 100 bp upstream of the TrSS. These predicted activities were then submitted to the DREAM consortium for evaluation[34].

Table 1. DNA sequence features predictive of promoter activity.

DNA feature	Description
Mononucleotides	Frequency of G
Dinucleotides	Frequency of GT
Trinucleotides	Frequency of 6 trinucleotides
Tetranucleotides	Frequency of 12 tetranucleotides
T-tracts	Length of T-tracts
TA-tracts	Length of TA-tracts
DNA deformability	Negatively correlated to activity

A total of 21 teams participated in the challenge (https://www.synapse.org//#!Synapse:syn2820426/wiki/71013). Predictions from our team had a Spearman correlation of 0.65 ($P = 0.002$, Figure 2A) to the actual activities, Pearson correlation of 0.65 ($P = 0.003$), chi-squared (χ^2) distance metric of 52.62 ($P = 0.508$) and R^2 statistic measuring the difference in ranks between predicted and actual promoter activities of 35.85 ($P = 0.004$). The *P*-values were generated from the probability of obtaining a comparable or lower performance using a null distribution in which predictions were made by randomly choosing an activity for each promoter amongst all the 21 participating teams. A combined score based on the negative logarithm (base 10) of the geometric mean of the *P*-values for all the 4 scores ranked our team first[34] (Figure 2B), with significant *P*-value in three out of four of the statistical tests used for evaluation. Further, although we were not ranked first in the χ^2 distance metric, our model performed the most consistently across the multiple assessment metrics, suggesting a robustness of the method. A detailed comparison of the teams was published previously by the DREAM consortium[34].

Biological significance of selected features

The final SVM models utilized only 23 features consisting of the frequencies of the mononucleotide G, dinucleotide GT, 6 different trinucleotides, 12 different tetranucleotides, length of poly(dT) and poly(dA-dT) tracts (Table 1). The relative importance of these features based on weights for the SVM models is provided (see Data availability). The feature with the highest weight was the frequency of the mononucleotide G, correlating negatively with promoter activity. For many of these features there was no clear link to underlying mechanisms of gene regulation. However, it is possible that some of the k-mers may be implicitly linked to transcription factor binding sites. That is, the combination of different k-mer

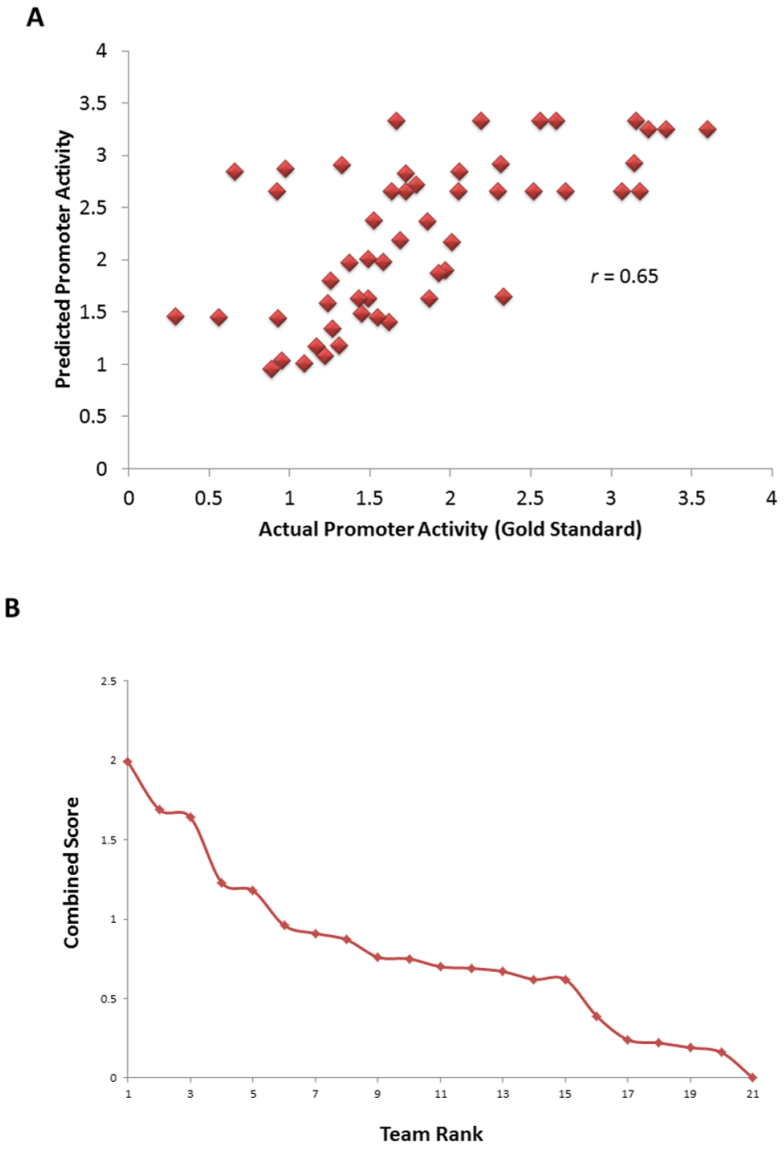

Figure 2. Performance of the SVM model on validation test set by the DREAM consortium. (**A**) Correlation between predicted activity by the SVM model and actual promoter activity of 53 promoters whose activity was not available to participants. (**B**) Performance of team FIrST relative to other 20 teams based on a combined score.

features could capture the binding motifs of specific transcription factors. For example the second most important feature in the SVM was the tetranucleotide ACCC which also occurs in the *Rap1* binding site motif[39]. In addition, frequencies of different k-mers could impact the DNA mechanical structure[40]. Among the features identified by the SVM model were poly(dT) and poly(dT-dA) tracts which influence the rigidity of DNA[24,26], thereby directly impacting nucleosome binding. Furthermore, insertion of poly(dT-dA) sequences into promoters can be used to regulate gene expression to a finer degree and at more gradual intervals than could be attained by transcription factor binding site mutations[38]. Some transcription factors are also highly dependent on the ability of DNA to bend[41–43]. In particular, TATA binding protein (TBP), which binds to the TATA box, is important for regulating the activity of RP promoters[42,44,45]. Another directly biologically relevant feature identified by the SVM was the deformability of DNA[36,46]. Promoters of low activity had more deformable DNA than those of high activity (Figure 3, $P = 0.008$). This was particularly evident at 40 to 60 bp from the TrSS when comparing the top 20 promoters with the highest versus those with the lowest activity (Figure 3).

Finally, some of the features may affect mRNA stability, especially given their potential location downstream of the transcription start sites (TSS). Besides sequence features in the 5'UTR that are close to the TSS could affect transcription, translation and mRNA stability.

Figure 3. Relationship between protein deformability of promoters and activity. Among the top 20 promoters with extreme activities (high and low), significant deviation in deformability occurs at the -40 to -60 bp region from the TrSS (T-test $P = 0.008$).

Error profile of SVM promoter activity model

Understanding the biases in prediction accuracy could provide biological insights into promoter classes and allow for refinement of models. Therefore, we investigated relationships between the nature of the test promoters and the magnitude of prediction error made by our model. Among the 53 test promoters provided by the DREAM challenge, 20 were natural yeast RP promoters while 33 were variants of these promoters with specific synthetic mutations introduced. These mutations included changes in the binding sites of the TBP, *Rap1, Fhl* and *Sfp1*, as well as introduction of nucleosome disfavoring sequences and random mutations. At the time of the challenge, participants were not aware of these mutations. The performance of our model on the set of natural promoters was much higher (Pearson correlation r= 0.73, $P = 0.0003$) compared to that for the mutated promoters (Pearson correlation r= 0.57, $P = 0.0005$). The prediction error was significantly less for natural promoters versus the mutated promoters (Student's t-test, $P = 0.01$, Figure 4A). This could partly be due to the composition of the training set, which contained only natural promoters. Similar poor performance was also observed in the models obtained from other teams[34]. In addition, most of the synthetic mutations were introduced at promoter locations residing outside of the 100 bp region from the TrSS and could not therefore be detected by our model. We also examined the correlation between the observed promoter activity and the prediction error. Promoters of low activity had larger prediction error (Pearson correlation between promoter activity and prediction error r= -0.31, $P = 0.02$, Figure 4B). Notably, natural promoters had slightly lower activity compared to synthetic promoters ($P = 0.02$) so the correlation between activity and prediction error may be a consequence of the low predictability of synthetic promoters. Thus, future models may benefit from data on activities of mutated promoters, which could enable a more accurate modeling of the impact of mutation on specific transcription factor binding sites.

Discussion

The quantitative modeling of gene expression has the potential to enhance our understanding of how gene regulation is fine-tuned in natural populations and has implications for the design of predictable gene expression systems. The DREAM6 challenge data set for promoter activity prediction was a unique opportunity to evaluate the predictability of gene expression from its promoter sequence. Given that all promoters were derived from natural yeast RP promoters that are expressed in the exponential phase[30], the challenge posed was more targeted towards DNA sequence patterns that fine-tune gene expression rather than simply determine the 'on/off' expression status. RP transcription regulation occurs in a highly coordinated manner and is critical for growth, allowing cells to adjust their protein synthesis capacity to physiological needs[47,48]. This is especially crucial as RP gene expression accounts for 50% of transcripts produced by RNA polymerase II[49] and their dysregulation leads to reduced fitness[47,48]. The yeast genome contains 137 RP genes, of which 19 encode a unique RP and 59 are duplicated. The proper functioning of ribosomes requires that all the ribosome components be expressed in equimolar concentrations[50] while simultaneously remaining responsive to physiological needs[51,52]. This is potentially challenging given the copy-number differences between the RP genes because high copy number genes generally show increased expression. The regulatory mechanisms underlying this fine-tuned regulation are not known. By accurately predicting the activity of the RP genes using the promoter sequences, we demonstrate that a considerable amount of this information is encoded in the DNA sequence.

It is intriguing that our model did not explicitly use transcription factor binding site information and focused only on the 100 bp upstream region. Some of the features identified by our model may influence transcription factor binding or nucleosomes indirectly, and could even affect mRNA translation. Transcription factors are critical for gene regulation. Their empirically identified binding sites are 6 to 8 bp, theoretically putting an upper bound on the level of regulatory flexibility that can be attained by mutating positions at these sites[30,38]. Cooperation between transcription factors or competition among them[15–17], and with nucleosomes[23], provides an additional mechanism for fine-tuned gene expression. RP promoters with high activity have not only more nucleosome disfavoring sequences but also characteristic spatial organization of the binding sites for *Rap1*, *Sfp1* and *Fhl1*[30]. The low performance of our model on synthetic promoters containing targeted mutations in transcription factor binding sites and nucleosome disfavoring sequences reinforces the importance of these factors. Consistent with this, the combination of our model and the mechanistically driven model involving transcription factors and nucleosome binding[30] was more predictive of promoter activity[34]. Our findings have implications for understanding the fine-tuned regulation of RP genes and engineering desirable activity in synthetic promoters.

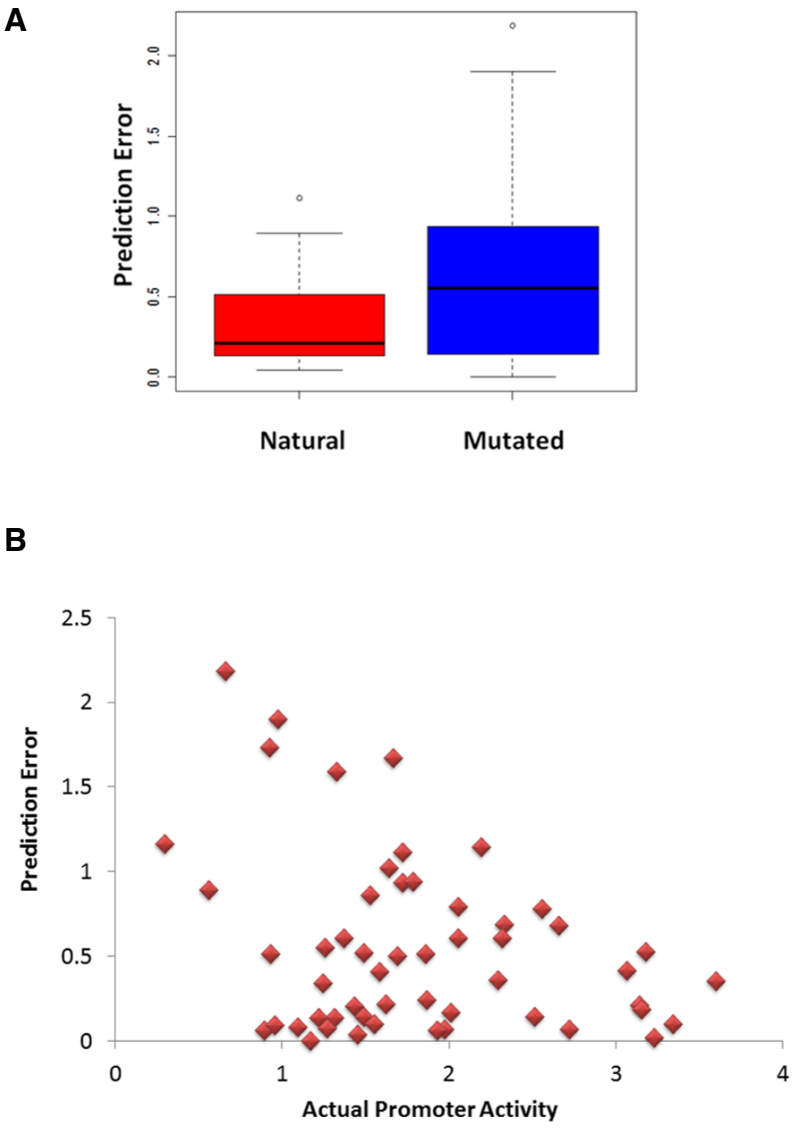

Figure 4. Dependence of prediction error on promoter class or activity. (**A**) Natural promoters had a lower prediction error compared to synthetically mutated promoters. (**B**) Prediction error is negatively correlated to promoter activity.

Data availability

F1000Research: Dataset 1. Raw data for 'Prediction of fine-tuned promoter activity from DNA sequence', Siwo *et al.* 2016, 10.5256/f1000research.7485.d113516[57]

Author contributions

GHS, RSP, AT, AKR conceived the methods and performed the analysis. All authors wrote the manuscript.

Competing interests

The authors declare that they have no competing interests.

Grant information

The author(s) declared that no grants were involved in supporting this work.

Acknowledgements

This work would not have been possible without the pre-publication provision of data to the DREAM challenge by Dr. Eran Segal and his group at the Weizmann Institute of Science, Israel, and the curation of the challenge by the DREAM committee: Drs. Gustavo Stolovitzky, Pablo Meyer and Rachel Norel at IBM Research, USA. We are grateful to the DREAM6 Promoter Prediction Consortium for the rigorous evaluation of the models.

References

1. Schadt EE, Monks SA, Drake TA, *et al.*: **Genetics of gene expression surveyed in maize, mouse and man.** *Nature.* 2003; **422**(6929): 297–302.
PubMed Abstract | Publisher Full Text

2. Tirosh I, Reikhav S, Sigal N, *et al.*: **Chromatin regulators as capacitors of interspecies variations in gene expression.** *Mol Syst Biol.* 2010; **6**: 435.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Tirosh I, Weinberger A, Carmi M, *et al.*: **A genetic signature of interspecies variations in gene expression.** *Nat Genet.* 2006; **38**(7): 830–834.
PubMed Abstract | Publisher Full Text

4. Field Y, Fondufe-Mittendorf Y, Moore IK, *et al.*: **Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization.** *Nat Genet.* 2009; **41**(4): 438–445.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Gonzales JM, Patel JJ, Ponmee N, *et al.*: **Regulatory hotspots in the malaria parasite genome dictate transcriptional variation.** *PLoS Biol.* 2008; **6**(9): e238.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Ellis T, Wang X, Collins JJ: **Diversity-based, model-guided construction of synthetic gene networks with predicted functions.** *Nat Biotechnol.* 2009; **27**(5): 465–471.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Gertz J, Cohen BA: **Environment-specific combinatorial *cis*-regulation in synthetic promoters.** *Mol Syst Biol.* 2009; **5**: 244.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Gertz J, Siggia ED, Cohen BA: **Analysis of combinatorial *cis*-regulation in synthetic and genomic promoters.** *Nature.* 2009; **457**(7226): 215–218.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Kim HD, Shay T, O'Shea EK, *et al.*: **Transcriptional regulatory circuits: predicting numbers from alphabets.** *Science.* 2009; **325**(5939): 429–432.
PubMed Abstract | Free Full Text

10. Segal E, Widom J: **From DNA sequence to transcriptional behaviour: a quantitative approach.** *Nat Rev Genet.* 2009; **10**(7): 443–456.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Takahashi K, Yamanaka S: **Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.** *Cell.* 2006; **126**(4): 663–676.
PubMed Abstract | Publisher Full Text

12. Kim HD, O'Shea EK: **A quantitative model of transcription factor-activated gene expression.** *Nat Struct Mol Biol.* 2008; **15**(11): 1192–1198.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Irie T, Park SJ, Yamashita R, *et al.*: **Predicting promoter activities of primary human DNA sequences.** *Nucleic Acids Res.* 2011; **39**(11): e75.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Cookson W, Liang L, Abecasis G, *et al.*: **Mapping complex disease traits with global gene expression.** *Nat Rev Genet.* 2009; **10**(3): 184–194.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Karczewski KJ, Tatonetti NP, Landt SG, *et al.*: **Cooperative transcription factor associations discovered using regulatory variation.** *Proc Natl Acad Sci U S A.* 2011; **108**(32): 13353–13358.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Mjolsness E: **On cooperative quasi-equilibrium models of transcriptional regulation.** *J Bioinform Comput Biol.* 2007; **5**(2B): 467–490.
PubMed Abstract | Publisher Full Text

17. Das D, Banerjee N, Zhang MQ: **Interacting models of cooperative gene regulation.** *Proc Natl Acad Sci U S A.* 2004; **101**(46): 16234–16239.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Lam FH, Steger DJ, O'Shea EK: **Chromatin decouples promoter threshold from dynamic range.** *Nature.* 2008; **453**(7192): 246–250.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Mirny LA: **Nucleosome-mediated cooperativity between transcription factors.** *Proc Natl Acad Sci U S A.* 2010; **107**(52): 22534–22539.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Li XY, Thomas S, Sabo PJ, *et al.*: **The role of chromatin accessibility in directing the widespread, overlapping patterns of *Drosophila* transcription factor binding.** *Genome Biol.* 2011; **12**(4): R34.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Choi JK, Kim YJ: **Intrinsic variability of gene expression encoded in nucleosome positioning sequences.** *Nat Genet.* 2009; **41**(4): 498–503.
PubMed Abstract | Publisher Full Text

22. Lidor Nili E, Field Y, Lubling Y, *et al.*: **p53 binds preferentially to genomic regions with high DNA-encoded nucleosome occupancy.** *Genome Res.* 2010; **20**(10): 1361–1368.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Raveh-Sadka T, Levo M, Segal E: **Incorporating nucleosomes into thermodynamic models of transcription regulation.** *Genome Res.* 2009; **19**(8): 1480–1496.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Segal E, Widom J: **Poly(dA:dT) tracts: major determinants of nucleosome organization.** *Curr Opin Struct Biol.* 2009; **19**(1): 65–71.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Kaplan N, Moore IK, Fondufe-Mittendorf Y, *et al.*: **The DNA-encoded nucleosome organization of a eukaryotic genome.** *Nature.* 2009; **458**(7236): 362–366.
PubMed Abstract | Publisher Full Text | Free Full Text

26. van der Heijden T, van Vugt JJ, Logie C, *et al.*: **Sequence-based prediction of single nucleosome positioning and genome-wide nucleosome occupancy.** *Proc Natl Acad Sci U S A.* 2012; **109**(38): E2514–22.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Segal E, Widom J: **What controls nucleosome positions?** *Trends Genet.* 2009; **25**(8): 335–343.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Lee CK, Shibata Y, Rao B, *et al.*: **Evidence for nucleosome depletion at active regulatory regions genome-wide.** *Nat Genet.* 2004; **36**(8): 900–905.
PubMed Abstract | Publisher Full Text

29. Shivaswamy S, Bhinge A, Zhao Y, *et al.*: **Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation.** *PLoS Biol.* 2008; **6**(3): e65.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Zeevi D, Sharon E, Lotan-Pompan M, *et al.*: **Compensation for differences in gene copy number among yeast ribosomal proteins is encoded within their promoters.** *Genome Res.* 2011; **21**(12): 2114–2128.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Yang YH, Dudoit S, Luu P, *et al.*: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.** *Nucleic Acids Res.* 2002; **30**(4): e15.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Oshlack A, Wakefield MJ: **Transcript length bias in RNA-seq data confounds systems biology.** *Biol Direct.* 2009; **4**: 14.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Kalir S, McClure J, Pabbaraju K, *et al.*: **Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria.** *Science.* 2001; **292**(5524): 2080–2083.
PubMed Abstract | Publisher Full Text

34. Meyer P, Siwo G, Zeevi D, *et al.*: **Inferring gene expression from ribosomal promoter sequences, a crowdsourcing approach.** *Genome Res.* 2013; **23**(11): 1928–1937.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Brukner I, Sánchez R, Suck D, *et al.*: **Trinucleotide models for DNA bending propensity: comparison of models based on DNaseI digestion and nucleosome packaging data.** *J Biomol Struct Dyn.* 1995; **13**(2): 309–317.
PubMed Abstract | Publisher Full Text

36. Olson WK, Gorin AA, Lu XJ, *et al.*: **DNA sequence-dependent deformability deduced from protein-DNA crystal complexes.** *Proc Natl Acad Sci U S A.* 1998; **95**(19): 11163–11168.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Sivolob AV, Khrapunov SN: **Translational positioning of nucleosomes on DNA: the role of sequence-dependent isotropic DNA bending stiffness.** *J Mol Biol.* 1995; **247**(5): 918–931.
PubMed Abstract | Publisher Full Text

38. Raveh-Sadka T, Levo M, Shabi U, *et al.*: **Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast.** *Nat Genet.* 2012; **44**(7): 743–750.
PubMed Abstract | Publisher Full Text

39. Lascaris RF, Mager WH, Planta RJ: **DNA-binding requirements of the yeast protein Rap1p as selected *in silico* from ribosomal protein gene promoter sequences.** *Bioinformatics.* 1999; **15**(4): 267–277.
PubMed Abstract | Publisher Full Text

40. Packer MJ, Dauncey MP, Hunter CA: **Sequence-dependent DNA structure: tetranucleotide conformational maps.** *J Mol Biol.* 2000; **295**(1): 85–103.
PubMed Abstract | Publisher Full Text

41. Laurens N, Rusling DA, Pernstich C, *et al.*: **DNA looping by FokI: the impact of twisting and bending rigidity on protein-induced looping dynamics.** *Nucleic Acids Res.* 2012; **40**(11): 4988–4997.
PubMed Abstract | Publisher Full Text | Free Full Text

42. Starr DB, Hoopes BC, Hawley DK: **DNA bending is an important component of site-specific recognition by the TATA binding protein.** *J Mol Biol.* 1995; **250**(4): 434–446.
PubMed Abstract | Publisher Full Text

43. Vijayan V, Zuzow R, O'Shea EK: **Oscillations in supercoiling drive circadian gene expression in cyanobacteria.** *Proc Natl Acad Sci U S A.* 2009; **106**(52): 22564–22568.
PubMed Abstract | Publisher Full Text | Free Full Text

44. Parvin JD, McCormick RJ, Sharp PA, *et al.*: **Pre-bending of a promoter sequence enhances affinity for the TATA-binding factor.** *Nature.* 1995; **373**(6516): 724–727.
PubMed Abstract | Publisher Full Text

45. Bosio MC, Negri R, Dieci G: **Promoter architectures in the yeast ribosomal expression program.** *Transcription.* 2011; **2**(2): 71–77.
PubMed Abstract | Publisher Full Text | Free Full Text

46. Yonetani Y, Kono H: **Sequence dependencies of DNA deformability and hydration in the minor groove.** *Biophys J.* 2009; **97**(4): 1138–1147.
PubMed Abstract | Publisher Full Text | Free Full Text

47. Li B, Vilardell J, Warner JR: **An RNA structure involved in feedback regulation**

of splicing and of translation is critical for biological fitness. *Proc Natl Acad Sci U S A.* 1996; **93**(4): 1596–1600.
PubMed Abstract | Publisher Full Text | Free Full Text

48. Deutschbauer AM, Jaramillo DF, Proctor M, *et al.*: **Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast.** *Genetics.* 2005; **169**(4): 1915–1925.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Warner JR: **The economics of ribosome biosynthesis in yeast.** *Trends Biochem Sci.* 1999; **24**(11): 437–440.
PubMed Abstract | Publisher Full Text

50. Spahn CM, Beckmann R, Eswar N, *et al.*: **Structure of the 80S ribosome from *Saccharomyces cerevisiae*--tRNA-ribosome and subunit-subunit interactions.** *Cell.* 2001; **107**(3): 373–386.
PubMed Abstract | Publisher Full Text

51. Ju Q, Warner JR: **Ribosome synthesis during the growth cycle of *Saccharomyces cerevisiae*.** *Yeast.* 1994; **10**(2): 151–157.
PubMed Abstract | Publisher Full Text

52. Causton HC, Ren B, Koh SS, *et al.*: **Remodeling of yeast genome expression in response to environmental changes.** *Mol Biol Cell.* 2001; **12**(2): 323–337.
PubMed Abstract | Publisher Full Text | Free Full Text

53. Oinn T, Addis M, Ferris J, *et al.*: **Taverna: a tool for the composition and enactment of bioinformatics workflows.** *Bioinformatics.* 2004; **20**(17): 3045–3054.
PubMed Abstract | Publisher Full Text

54. Goñi JR, Fenollosa C, Pérez A, *et al.*: **DNAlive: a tool for the physical analysis of DNA at the genomic scale.** *Bioinformatics.* 2008; **24**(15): 1731–1732.
PubMed Abstract | Publisher Full Text

55. Satchwell SC, Drew HR, Travers AA: **Sequence periodicities in chicken nucleosome core DNA.** *J Mol Biol.* 1986; **191**(4): 659–675.
PubMed Abstract | Publisher Full Text

56. Hall M, Frank E, Holmes G, *et al.*: **The WEKA data mining software: an update.** *SIGKDD Explor.* 2009; **11**(1): 10–18.
Publisher Full Text

57. Siwo G, Rider A, Tan A, *et al.*: **Dataset 1 in: Prediction of fine-tuned promoter activity from DNA sequence.** *F1000Research.* 2016.
Data Source

Open Peer Review

Current Referee Status:   

Version 1

Referee Report 20 June 2016

doi:10.5256/f1000research.8064.r14225

Jan Grau
Institute of Computer Science, Martin Luther University of Halle-Wittenberg, Halle, Germany

The authors present FIrST, an approach for predicting promoter activity from sequence, which won one of the DREAM6 challenges. FIrST is using only simple sequence features in a limited range (100 bp) upstream of the translation start site for making its predictions, which distinguishes it from several other approaches in this field.
Prediction results are convincing and the method appears to be sound. However, currently the method is not described detailed enough. In addition, I have a few further major and minor concerns regarding the current version of the manuscript:

Major comments:

1. In the list of features described in section "Feature extraction", some seem redundant to me. For instance, the trinucleotide parameters for bendability are just computed from the k-mers for k=3. Also nucleosome binding prediction was based on trinucleotide preference. Please explain why it may be useful to also include those 3-mer-derived features in addition to the 3-mers themselves.

2. The description of methods in section "Machine learning model exploration" is too coarse. Please provide more detail on the SVMs, linear regression, and regression trees employed. It also remains unclear if the scales of features are normalized somehow, before their values are provided to the SVM.

3. No details are given on the selected 3-mers and 4-mers (Table 1). Please provide a list of the specific k-mers selected by FIrST. It may also be reasonable to discuss potential biological reasons for their importance (as partly covered for TATA-boxes on page 6).

4. Considering Fig. 3, I wondered if the difference in deformability may be related to transcription initiation. Or, stated differently, might we observe an ever clearer signal if all sequences (and their deformability profiles) would be aligned by the transcription start site (TSS) instead of the TrSS? One idea in the same direction, which could contribute to the novelty of the manuscript, would be to evaluate similar profiles (of sequences aligned to TSS or TrSS) for all features found to be informative by FIrST. For instance, one could expect to see something like general fluctuations of G/C content, or the TATA-box in 4-mer profiles as a spike approx. 35 bp before the TSS. From my perspective, this might improve the novelty of the manuscripts and the interpretation of features.

Minor comments:

1. The data from the DREAM6 challenge only consider a special subset of genes (ribosomal genes) and only in yeast. It is unclear if the features derived by the authors' method would also be informative for higher eukaryotes. I understand that this question cannot be finally answered from the DREAM6 data, but the authors might comment on this issue.

2. Figure 1B remains a bit unclear. In the caption and the main text, the authors explain that they use non-overlapping 100bp sub-sequences. However, from the figure it rather seems that they consider upstream sequences of 300 bp, 200 bp and 100 bp (and the full promoter sequence) relative to the translation start site. Please clarify.

3. In section "DREAM6 challenge data" of "Methods", the authors refer to "the sequence 1200 bp upstream of a gene", where "upstream of the translation start site" (as in the remainder of the text) would be more specific.

4. In section "Feature extraction", the authors explain that "each promoter sequence was divided into 100 bp non-overlapping windows", while in the previous section they explain that the full 1200 bp sequences do not extend over the nearest gene. From my understanding, this may result in some of the sequences being shorter than 1200 bp, and their length might not be dividable by 100. Please explain how such cases are handled.

5. At the end of section "Validation of model by DREAM6 consortium", the authors explain that "the overall score was defined as the product of the four P-values", whereas later they explain that $-\log_{10}$ of the geometric mean of the p-values was used as the overall measure. Although bot definition are equivalent with respect to the resulting ranking, I would suggest to provide one consistent definition of the overall score.

6. From the manuscript it did not become fully clear if the TA-tracts (also termed poly(dA-dT) tracts in some parts of the manuscript) are tracts of poly "A or T" or tracts of poly "AT"-dinucleotides.

7. In section "Error profile of SVM promoter activity model", the authors explain that natural promoters had (slightly) lower activity than synthetic promoters and that the prediction error of the SVM is lower for natural promoters. However, I did not get the idea, why this should explain that low activity genes had larger prediction errors.

8. In section "Error profile of SVM promoter activity model", the authors explain that one reason why FIrST did not perform well for synthetic promoters is that most mutations had been introduced outside the 100 bp range considered by FIrST. However, this reasoning partly contradicts the claim of the authors that most of the transcriptional activity may be explained from the sequence in that 100 bp window. If this would truly be the case, mutations outside this range should have only minor effects.

9. In the Discussion, the authors mention that TF binding motifs are 6 to 8 bp in length. While this may be true for several yeast TFs, it is not correct for eukaryotes in general and motifs may be wider than 10 bp.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Referee Report 07 April 2016

doi:10.5256/f1000research.8064.r12380

Jianhua Ruan
Department of Computer Science, The University of Texas at San Antonio, San Antonio, TX, USA

This article describes the winning method of the DREAM6 promoter activity prediction challenge. While a meta analysis of the competing methods participated in the challenge has already been published (Meyer *et al.*, 2013), this article provides more details of the winning method and some additional analysis of the predictive model, which may lead to better understanding of the predictability of gene transcription. While its contribution is undoubtable, this article should be revised to address several issues:

Major issues:

1. The 23 features utilized by the SVM model (as well as their coefficients in the model) is not provided explicitly in the main text nor in the supplement file. Table 1 in the main text shows that 6 trinucleotides and 12 tetranucleotides are important features, but it is nowhere to be found which tri- and tetra-nucleotides they are. For lengths of T or TA-tracts, the supplement file shows several different values, including mean, median and stdev. It is unclear which one is actually used by the SVM model. Similarly, supplement file shows 79 values for deformability and it is unknown which one is used.

2. In the case of ranking the features by their SVM coefficients, the authors need to clarify if the feature values were normalized prior to model building, as these features are on very different scales and if not normalized the ranking of the coefficients are not very meaningful.

3. The main conclusion in the subsection "Error profile of SVM promoter activity model" do not seem to make sense. First, promoters of low activity had larger prediction error. Then the authors stated that natural promoters had lower activity. This seems to contradict with their observation that the prediction error was significantly less for natural promoters than for mutated promoters.

Minor issues:

1. Authors only mentioned that feature selection was done in WEKA with wrapper. More details need to be given. For example, what was the selection strategy used by the wrapper, e.g., exhaustive search, greedy forward search, backward search, or other types of heuristics?

2. What is the purpose of first training 1000 SVM classifiers using 66% of data as training and 34% as testing, and then another 500 SVM classifiers using 80% as training and 20% as testing?

References

1. Meyer P, Siwo G, Zeevi D, Sharon E, Norel R, Segal E, Stolovitzky G: Inferring gene expression from ribosomal promoter sequences, a crowdsourcing approach.*Genome Res*. 2013; **23** (11): 1928-37
PubMed Abstract | Publisher Full Text

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Referee Report 02 March 2016

doi:10.5256/f1000research.8064.r12530

Paul Pavlidis
Centre for High-Throughput Biology and Department of Psychiatry, University of British Columbia, Vancouver, BC, Canada

Siwo *et al*. give a detailed report of their entry to the DREAM promoter activity prediction assessment, conducted in 2011. The paper describing the results of the assessment appeared in 2013 (Meyer *et al*.), and the entry from Siwo *et al*. ("FiRST") was the top-performer overall. Meyer *et al*. gives few details about the specific methods, mentioning only that the FiRST entry used an SVM and did not use TF binding site motif information. Here it is clarified that FiRST is a simple method that uses only part of the sequence and the most prominent features were about nucleotide content.

Because it is perhaps a little eye-opening (even embarrassing, depending on one's point of view) that the best method in the assessment is so simple, this paper is an important footnote to Meyer *et al*. but it could be fleshed out further to get at what is going on. My suggestions for revisions are to give more detail about the properties of the sequences used and the relationship to performance.

FiRST predicts from only the 100 bases of sequence upstream of the translation start (which was considered as part of the promoter by DREAM; I note this is not "upstream of the gene" as described by Siwo *et al*. in the methods section), and that their predictions were dominated by the effect of a simple measure of G content. Siwo *et al*. report that they did worse at predicting the synthetically mutated promoters (this was apparently not true overall across methods as reported by Meyer *et al*.). In Meyer *et al*., adding tf binding information to FiRST improved performance.

The authors mention this, but the most important reason that FiRST does poorly at predicting the synthetic mutations seems to be that most of the mutations (seems to be 29 out of 33, based on Table 1 of Meyer *et al*.) are not in the 100 bp window used. That is, because in most cases these synthetic sequences were (as I understand it) identical in features to other examples while having different activities, for the purposes of FiRST, they could only introduce prediction errors. In light of this fact the rest of the speculation about why performance varied in this way seems extraneous.

It would also be useful to see more detailed information on the sequences used (e.g., the G content or other features), and the prediction error in each case. How well does one predict using G content alone? This might all be reconstructed from the data supplement helpfully provided, but the authors should consider providing the analysis. It also seems reasonable to ask for more details about the performance of other sequence windows.

The main other missing piece from this paper is any discussion or evidence that the method works beyond the narrow confines of the DREAM setup. Even for the RP genes, does it make a useful prediction, that increasing the G content of RP promoters in that 100 bp window will decrease promoter

activity? I am fine with leaving this as “future work” but it would be worth mentioning.

Figure 2B is apparently the same as part of Figure 1E from Meyer *et al*., except FiRST is not marked (actually there is a small difference in the values plotted; the combined score for FiRST looks closer to 2 than the 1.87 reported and plotted in Meyer *et al*.). The authors should clearly cite Meyer *et al*. in the figure caption as the source of the data for this figure, or simply point the readers to Meyer *et al*., or else explain where the data came from if not from Meyer *et al*.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.