Inferring gene expression from ribosomal promoter sequences, a crowdsourcing approach

Pablo Meyer, Geoffrey Siwo, Danny Zeevi, et al.

*Genome Res.* published online August 15, 2013
Access the most recent version at doi:10.1101/gr.157420.113

**Supplemental Material**
http://genome.cshlp.org/content/suppl/2013/09/24/gr.157420.113.DC1.html

**Related Content**
Mapping functional transcription factor networks from gene expression data
*Genome Res.* August, 2013 23: 1319-1328
Assemblathon 1: A competitive assessment of de novo short read assembly methods
Dent Earl, Keith Bradnam, John St. John, et al.
*Genome Res.* December, 2011 21: 2224-2241

**P<0.05**
Published online August 15, 2013 in advance of the print journal.

**Creative Commons License**
This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported), as described at http://creativecommons.org/licenses/by-nc/3.0/.

**Email Alerting Service**
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

© 2013 Meyer et al.; Published by Cold Spring Harbor Laboratory Press
Inferring gene expression from ribosomal promoter sequences, a crowdsourcing approach

Pablo Meyer,1,6 Geoffrey Siwo,2 Danny Zeevi,3 Eilon Sharon,4 Raquel Nore,1 DREAM6 Promoter Prediction Consortium,5 Eran Segal,4 and Gustavo Stolovitzky1

1IBM T.J. Watson Research Center, Yorktown Heights, New York 10598, USA; 2Eck Institute for Global Health, Department of Biological Sciences, University of Notre Dame, Notre Dame, Indiana 46556, USA; 3Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08540, USA; 4Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel

The Gene Promoter Expression Prediction challenge consisted of predicting gene expression from promoter sequences in a previously unknown experimentally generated data set. The challenge was presented to the community in the framework of the Sixth Dialogue for Reverse Engineering Assessments and Methods (DREAM6), a community effort to evaluate the status of systems biology modeling methodologies. Nucleotide-specific promoter activity was obtained by measuring fluorescent signal from promoter sequences fused upstream of a gene for yellow fluorescence protein and inserted in the same genomic site of yeast Saccharomyces cerevisiae. Twenty-one teams submitted results predicting the expression levels of 53 different promoters from yeast ribosomal protein genes. Analysis of participant predictions shows that accurate values for low-expressed and mutated promoters were difficult to obtain, although in the latter case, only when the mutation induced a large change in promoter activity compared to the wild-type sequence. As in previous DREAM challenges, we found that aggregation of participant predictions provided robust results, but did not fare better than the three best algorithms. Finally, this study not only provides a benchmark for the assessment of methods predicting activity of a specific set of promoters from their sequence, but it also shows that the top performing algorithm, which used machine-learning approaches, can be improved by the addition of biological features such as transcription factor binding sites.

[Supplemental material is available for this article.]

One of the main objectives of the Dialogue for Reverse Engineering Assessments and Methods (DREAM) (Stolovitzky et al. 2007) is to catalyze the interaction between experiment and theory in systems biology, particularly for quantitative model building. For this purpose, unpublished data is used to objectively test team predictions generated by their methods/algorithms. The evaluation of participants’ methods is blind, as inspired by the community challenges posed in CASP (Critical Assessment of Techniques for Protein Structure Prediction). CASP’s main goal is to obtain an in-depth and objective assessment of state-of-the-art techniques for protein structure prediction using a set of unpublished protein structures (Moult et al. 1995; Shortle 1995; Moult 1996). This same principle is used in DREAM where a blind benchmark is provided so predictions from different algorithms can be easily compared, thus enhancing the reliability of programs/methods used. We describe here the Gene Promoter Expression Prediction Challenge from DREAM6, identify the best performers, and discuss the main results, as well as an improvement of the top-performing algorithm. The full description of the challenge, as was presented to the participants, including the teams ranking, can be found at the DREAM website (http://the-dream-project.org).

Gene Promoter Expression Prediction challenge

The level at which genes are transcribed is determined in large part by the DNA sequence upstream of the gene, known as the promoter region. Although widely studied, we are still far from a quantitative and predictive understanding of how transcriptional regulation is encoded in cis-regulatory elements of gene promoters (Kaplan et al. 2009; Sharon et al. 2012). One obstacle in the field is obtaining accurate measurements of transcription derived from different promoters. Fusion of promoters to fluorescent reporters can be used to determine the relative contribution of transcription to the resulting mRNA levels, since they provide measurements of promoter activity independent of the sequence of the associated transcript (Kalir et al. 2001). To further address this, an experimental system was designed to measure the transcription derived from different promoters, all of which are inserted into the same genomic location upstream of a reporter gene—a yellow fluorescence protein gene (YFP) (Zeevi et al. 2011).

To study a set of promoters that share many regulatory elements and thus are suitable for computational learning, data pertaining to promoters of most of the ribosomal protein (RP) genes in yeast Saccharomyces cerevisiae grown in a rich medium condition was obtained (Zeevi et al. 2011). Although ribosomal promoters may not capture generic promoter features, the challenge presented sought to model RP promoters to address questions left unanswered by successful genome-wide models (Beer and Tavazoie 2004; Gertz and Cohen 2009; Irie et al. 2011), such as what are the mechanisms behind the equimolar expression of the RP genes.
despite their varying copy numbers and how the information for fine-tuned expression is encoded in promoter regions. Also, understanding the basis of fine-tuned regulation of highly homologous promoters could provide clues to engineer promoter libraries of desired activity, starting from a parent promoter sequence.

The promoter regions for the *S. cerevisiae* RP genes were defined as the sequence immediately upstream of the ribosomal protein coding region beginning at the translation start site (TSS) and continuing 1200 bp or until reaching another upstream gene's coding sequence, selecting whichever came first. This removes a source of variability between strains derived from post-transcriptional regulation related to the coding and 3' untranslated regions. Each promoter was linked to a *URA3* selection marker (Linhiz et al. 2008) and inserted into the same fixed location in the yeast genome (Gietz and Schiestl 2007) of a master strain that contained the *yfp* gene (see Fig. 1A). In addition to 110 natural RP promoter strains, we constructed 33 strains with site-specific mutated RP promoters using similar methods (Gietz and Schiestl 2007; Linhiz et al. 2008).

The strains containing the different RP derived promoters were synchronized and grown, and their YFP fluorescence was recorded in a plate reader. The transcription initiated by each

---

**Figure 1.** Overview of the experimental system and results. (A) Illustration of the master strain into which we integrated all the tested promoters. At a fixed chromosomal location, the master strain contains a gene that encodes a red fluorescent protein (mCherry), followed by the promoter for *TEF2*, and a gene that encodes for a yellow fluorescent protein (YFP). Every tested promoter is integrated into this strain, together with a selection marker, between the *TEF2* promoter and the YFP gene. (B) Strains with different promoters have highly similar growth rates. Shown is the growth of 71 different promoter strains, measured as OD (optical density). Measurements were obtained from a single 96-well plate, with glucose-rich media and a small number of cells from each strain inserted into each well at time zero. The exponential growth phase is indicated (vertical dashed gray line). (C) Same as B, but where the measurements correspond to mCherry intensity. Note the small variability in the intensity of mCherry, which is driven by the same control promoter across the different strains. (D) Same as C, but where the measurements correspond to YFP intensity. Note the large variability in the intensity of YFP, which is driven by a different promoter in each strain. (Adapted with permission from Zeevi et al. (2011)). (E) Black line shows the scores from different participating teams plotted in descending order, and red line shows scores of aggregated teams starting with the score obtained from averaging the prediction results of the two best-performing teams, followed by the three best-performing teams, and so on until all 21 teams are included. The stand-alone dot represents the post-hoc model combining SVM and biological features.
promoter was measured by its promoter activity, defined as the average YFP fluorescence during the exponential growth phase divided by the average optical density (OD) during that time period (see Fig. 1B,D). Hence, promoter activity represents the average rate of YFP production from each promoter, per cell per second, during the exponential phase of growth (Zeevi et al. 2011). As a control for the experimental error, a red fluorescent protein (mCherry) was driven by a control promoter, identical in all strains (see Fig. 1C). Several tests were performed to gauge the accuracy and sensitivity of the system. The results showed that growth curves of all strains were nearly identical, YFP levels of independent clones of the same promoter sequence were indistinguishable from those of replicate measurements of the same clone, signals measured in the YFP wavelength were not affected by the presence of the mCherry protein, and no correlation was found between the YFP and mCherry promoter activities across the different RP promoter strains. Finally, the average difference between any two mCherry strains was 5%, and when using replicate measurements, the relative error in the estimated YFP promoter activity of an RP promoter is 2%, indicating that it is possible to distinguish between any two promoters whose activities differ by as little as 8% (Zeevi et al. 2011).

The challenge

The challenge consisted of predicting the promoter activity derived from a given RP promoter sequence. Participants were provided with a training set of 90 natural RP promoters (see Supplemental Table S1) for which both the promoter sequence and activity were known and a test set of 53 promoters (see Supplemental Table S2) for which only the promoter sequence was given. The test set was divided into two subsets. The first subset had 20 natural RP promoters. The second subset contained 33 promoters that are similar to natural RP promoters but have some mutations in their sequence. These mutations can be separated into six types: mutations of TATA boxes (Basehoar et al. 2004), of binding sites for Pho1 and Sp1—known transcriptional regulators of RP genes (Badis et al. 2008; Zhu et al. 2009), mutations to nucleosome disfavoring sequences, random mutations that occurred unintentionally while creating the natural promoters, and finally, sequences mutated intentionally with additional random mutations (see Table 1). The goal was to predict as accurately as possible the promoter activity of the 53 promoters in the test set using the 90 promoters for training.

Results and analysis

The challenge was scored in four different ways using criteria that considered the "distance" between measured and predicted values or differences in rank between measured and predicted values. The first metric consists of a Pearson correlation between the predicted and measured promoter activity. The second metric is a normalized sum of squared differences. The third is the Spearman rank correlation, which is essentially the Pearson correlation between the ranks, and the fourth metric is a normalized sum of the squared difference in ranks. These metrics were then combined into a score (see Methods, Eqs. 1–5).

As shown in Figure 1E and Table 2, out of 21 participating teams, team FIRST was the best performer, with a score of 1.88, followed by team c4lab with 1.55, in a close race for the second place with the third team, which was then followed by a monotonic decrease in the participants’ scores. When a series of aggregated teams are formed by averaging the predicted promoter activity values of the best N teams, the score of the aggregated best 15 teams becomes 1.49, close to that of the second-best performing team (c4lab) (see Fig. 1E). Scores for the remaining aggregated teams are also observed to be above the fourth ranked team, showing that blending community predictions produces robust results (see Supplemental Material, DREAM6 Participants Predictions file).

We analyzed whether some participants were better at predicting specific promoters but could not find any correlation between overall team ranking and the number of promoters a team predicted best. Also, when predicting single promoters, the overall highly ranked methods did not rank first more often than lower ranked ones but fared well across all promoters.

In order to investigate whether some promoters were harder to predict, we calculated the average distance $d^2$ over all participants for promoter i from the promoter's predicted value to its measured value (see Eq. 6, Methods). As seen in Figure 2A, where promoters are ordered by increasing $d^2$, five promoters out of the 53 stand out for being predicted with less accuracy. We next divided the promoters based on $d^2$ into two groups consisting of the best 30 predictions (green dots, Fig. 2A) and the 23 worst predictions (red dots, Fig. 2A) and plotted the Pearson correlation of each of the participating teams for these two groups of promoters (Fig. 2B). For all teams, the Pearson correlation clearly separated the best-predicted and worst-predicted promoters as defined by $d^2$ showing that, for all participants, promoters could be consistently divided into two groups, one of which was harder to predict than the other.

To identify the source of the difficulty in predicting the expression values of these 23 promoters, we explored the possibility of this list being enriched for mutant promoters. Wild-type promoters were found to be distributed equally between the worst-predicted promoters (10 empty dots on red side of Fig. 2A) and best-predicted promoters (10 empty dots on green side of Fig. 2A). A Fisher test shows no statistical significance for mutant or wild-type promoter enrichment. We next used measure $x_i$ (see Eq. 7, Methods) to evaluate whether promoter activity was correlated to the difficulty of predicting its value. Figure 2C, showing how $x_i$ varies for each promoter, reflects that participants' performance is anti-correlated with promoter activity, with a Pearson correlation of 0.836. Participants' prediction accuracy can be divided into three groups according to their promoter activity $j_i$ values between 1 and 3 ($<x_i> = 0.25$ $0.73$ for i such that $1 > j_i > 3$; 18 promoters)—which fared significantly better than the following two groups: $j_i$ values less than 1 ($<x_i> = 3.02$ 1.10; for i such that $1 > j_i > 3$; 18 promoters)—which fared significantly better than the following two groups: $j_i$ values higher than 3 ($<x_i> = 1.48$ 0.51 for i such that $j_i > 3$; 7 promoters, t-test $p < 1.75$ 10^{-5}$). Both observations are independent of whether the promoters contain mutations (Fig. 2C, full and empty dots).

As we could not find clear differences between mutant and wild-type promoters when using the $d^2$ measure, we calculated a different type of distance $d^2$ to compare participant predictions and measurements (see Eq. 8, Methods). As shown in Figure 3A, $d^2$ clearly distinguishes wild-type promoters (mean value of $d^2$ is $1.62$ 0.22) from mutant promoters (mean value of $d^2$ is $2.23$ 0.41, t-test $p < 8.3$ 10^{-8}). In order to understand the differences in $d^2$ for the various mutant promoters, we formed six groups according to the nature of their mutations. In Figure 3B, the different groups of mutations were ordered according to the associated $d^2$ mean value. Participants' predictions fared better for mutations typically inducing small changes in promoter expression (low $d^2$ in Fig. 3B), such as random mutations. Conversely, sequence mutations known
Table 1. Information on the promoter sequence mutations

<table>
<thead>
<tr>
<th>Promoter</th>
<th>Random Mutations</th>
<th>Fh1 site deletions</th>
<th>Sfp1 site deletions</th>
<th>TATA box deletions</th>
<th>Nucleosome disfavoring sequences deletions</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPL13A_M1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPS12_M1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPS14B_M1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPL39_M1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPS10_M1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPL10_M2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPL8A_M1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPL21B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPL13B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPL4_M1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPL15A_M1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPL4A_M2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPL4_M2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPL15B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPL3_M1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPL11A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For every promoter, locations of TATA boxes (pink circles), and of binding sites for Rap1 (red), Fh1 (green), and Sfp1 (blue) are shown. In addition, shown is the per-base pair nucleosome occupancy of every promoter (occupancy is shown in a white to black scale, with white corresponding to no occupancy and black to full occupancy), predicted using a computational model of nucleosome sequence preferences (Kaplan et al. 2009). Also shown is a matrix (left) summary of the number of factor sites that appear in every RP promoter (counts for Rap1 are only shown for the 400 bp upstream of the TrSS; for Fh1 and Sfp1, 300 bp; and for TATA, 200 bp), along with a column representing whether the corresponding RP gene exists in a single copy in the yeast genome (first column, black) and whether it is an essential gene (second column, gray). The length of each native promoter is indicated (cyan vertical line) if it is shorter than 600 bp.

to induce large changes by lowering promoter expression, such as mutations to the TATA box, were the worst-predicted group (high $d^0$ in Fig. 3B). As there is not enough data to extract a statistical measure of the differences between groups of promoters, we decided to follow up on the previous observation and compare the $d^0$ value for each mutant promoter to the relative promoter activity difference induced by the mutations. As shown in Figure 3C, $d^0$ grows exponentially with increasing differences between wild-type and mutant promoter expression. Hence, prediction accuracy for mutant promoters worsened when mutations induced higher changes on expression.

Improving promoter expression prediction by adding biological features

As shown in Figure 1B, scores of aggregated teams were observed to be robustly above the fourth-ranked team but did not fare better than the three best-performing teams. As the best-performing models of this challenge did not include biological features such as the binding sites for Fh1 and Sfp1, known transcriptional regulators of RP transcription factors, we decided to try to improve model performance by including biological features in the best-performer algorithm of team FIRST. To do this, we modified a recently published mechanistically motivated model that takes into account the competition between transcription factors and nucleosomes for DNA binding sites in the regulation of gene expression (Zeevi et al. 2011) (Eqs. 9 and 10; see Methods). The score for this model based on $C_p$, the Pearson correlation between predicted and observed activity, was 0.49 (see Eq. 1, Methods). We then combined this model with that of the best-performing team, FIRST, in two ways. In the first approach, we averaged the predicted activity of each promoter by team FIRST and the mechanistic model. The correlation between the predicted and actual activities
Table 2. Scores from different teams ranked in descending order

<table>
<thead>
<tr>
<th>Rank</th>
<th>Cp</th>
<th>X2</th>
<th>Sp</th>
<th>R2</th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
<th>p4</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6475</td>
<td>0.526197</td>
<td>0.6469</td>
<td>35.852</td>
<td>0.0035</td>
<td>0.515</td>
<td>0.0011</td>
<td>0.0152</td>
<td>1.8759</td>
</tr>
<tr>
<td>2</td>
<td>0.5386</td>
<td>0.308202</td>
<td>0.4938</td>
<td>37.7716</td>
<td>0.0099</td>
<td>0.0013</td>
<td>0.162</td>
<td>0.0306</td>
<td>1.5462</td>
</tr>
<tr>
<td>Team263</td>
<td>0.5184</td>
<td>0.297988</td>
<td>0.4436</td>
<td>37.1604</td>
<td>0.1461</td>
<td>0.0007</td>
<td>0.3505</td>
<td>0.0245</td>
<td>1.5107</td>
</tr>
<tr>
<td>Team164</td>
<td>0.4925</td>
<td>0.351053</td>
<td>0.4837</td>
<td>40.5142</td>
<td>0.2243</td>
<td>0.0121</td>
<td>0.1946</td>
<td>0.07</td>
<td>1.1079</td>
</tr>
<tr>
<td>Team259</td>
<td>0.4959</td>
<td>0.313914</td>
<td>0.456</td>
<td>48.8889</td>
<td>0.2132</td>
<td>0.0019</td>
<td>0.2982</td>
<td>0.3445</td>
<td>1.0941</td>
</tr>
<tr>
<td>Team140</td>
<td>0.534</td>
<td>0.612089</td>
<td>0.5887</td>
<td>47.7112</td>
<td>0.1077</td>
<td>0.4205</td>
<td>0.1316</td>
<td>0.2694</td>
<td>0.0628</td>
</tr>
<tr>
<td>Team250</td>
<td>0.4743</td>
<td>0.407905</td>
<td>0.5326</td>
<td>50.29</td>
<td>0.291</td>
<td>0.0783</td>
<td>0.0825</td>
<td>0.4066</td>
<td>0.7791</td>
</tr>
<tr>
<td>Team284</td>
<td>0.5253</td>
<td>0.657449</td>
<td>0.4732</td>
<td>37.8666</td>
<td>0.1285</td>
<td>0.0908</td>
<td>0.2315</td>
<td>0.0317</td>
<td>0.7669</td>
</tr>
<tr>
<td>Team17</td>
<td>0.5211</td>
<td>0.690551</td>
<td>0.477</td>
<td>41.6918</td>
<td>0.1391</td>
<td>0.9477</td>
<td>0.2176</td>
<td>0.0945</td>
<td>0.6418</td>
</tr>
<tr>
<td>Team154</td>
<td>0.4457</td>
<td>0.554508</td>
<td>0.4901</td>
<td>40.4362</td>
<td>0.4071</td>
<td>0.6326</td>
<td>0.1735</td>
<td>0.0685</td>
<td>0.6286</td>
</tr>
<tr>
<td>Team21</td>
<td>0.4197</td>
<td>0.57103</td>
<td>0.5661</td>
<td>49.2351</td>
<td>0.5181</td>
<td>0.694</td>
<td>0.029</td>
<td>0.3599</td>
<td>0.6065</td>
</tr>
<tr>
<td>Team76</td>
<td>0.469</td>
<td>0.504026</td>
<td>0.4598</td>
<td>42.6173</td>
<td>0.3116</td>
<td>0.4159</td>
<td>0.2826</td>
<td>0.1176</td>
<td>0.5914</td>
</tr>
<tr>
<td>Team61</td>
<td>0.5327</td>
<td>0.554348</td>
<td>0.4214</td>
<td>46.5912</td>
<td>0.1109</td>
<td>0.632</td>
<td>0.3171</td>
<td>0.2494</td>
<td>0.564</td>
</tr>
<tr>
<td>Team187</td>
<td>0.5467</td>
<td>0.619649</td>
<td>0.5064</td>
<td>55.8043</td>
<td>0.0825</td>
<td>0.8386</td>
<td>0.1266</td>
<td>0.6453</td>
<td>0.5619</td>
</tr>
<tr>
<td>Team257</td>
<td>0.502</td>
<td>0.476773</td>
<td>0.4457</td>
<td>50.6775</td>
<td>0.1939</td>
<td>0.2976</td>
<td>0.3036</td>
<td>0.4244</td>
<td>0.5322</td>
</tr>
<tr>
<td>Team253</td>
<td>0.4648</td>
<td>0.919082</td>
<td>0.4386</td>
<td>50.7745</td>
<td>0.3984</td>
<td>0.9996</td>
<td>0.3721</td>
<td>0.4286</td>
<td>0.2993</td>
</tr>
<tr>
<td>Team264</td>
<td>0.3278</td>
<td>0.440867</td>
<td>0.2259</td>
<td>77.3447</td>
<td>0.8439</td>
<td>0.1645</td>
<td>0.9797</td>
<td>0.9916</td>
<td>0.2178</td>
</tr>
<tr>
<td>Team245</td>
<td>0.3031</td>
<td>0.464831</td>
<td>0.2973</td>
<td>70.9865</td>
<td>0.8963</td>
<td>0.2503</td>
<td>0.8949</td>
<td>0.8163</td>
<td>0.1964</td>
</tr>
<tr>
<td>Team265</td>
<td>0.3932</td>
<td>0.502789</td>
<td>0.3729</td>
<td>81.9649</td>
<td>0.6294</td>
<td>0.4104</td>
<td>0.665</td>
<td>0.9973</td>
<td>0.1916</td>
</tr>
<tr>
<td>Team176</td>
<td>0.2658</td>
<td>0.476607</td>
<td>0.2383</td>
<td>72.0592</td>
<td>0.9485</td>
<td>0.2969</td>
<td>0.9698</td>
<td>0.9737</td>
<td>0.1438</td>
</tr>
<tr>
<td>Team138</td>
<td>0.1297</td>
<td>0.888242</td>
<td>0.0928</td>
<td>106.762</td>
<td>0.9984</td>
<td>0.9994</td>
<td>0.9997</td>
<td>1</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

Only names of the two best-performing teams are indicated. Cp (see Eq. 1) indicates the Pearson metric, X2 the score based on the x² metric (see Eq. 2), Sp the score based on the Spearman metric (see Eq. 3), and R2 the score based on the rank² metric (see Eq. 4). p1, p2, p3, and p4 are the associated F-values based on the null-hypothesis generated from randomized values for the distances Cp, X2, Sp, and R2. Note that F-values become significant across the table if a less stringent null-hypothesis is applied. The last column is the final score calculated as the P-value product: \[ p = \prod_{i=1}^{3} p_i \] (see Eq. 5).

remained the same as for FIRST (0.65) (see Table 1), demonstrating the robustness of aggregating predictions even when one method has considerably lower performance. Given that the method by team FIRST did not explicitly use transcription factor binding, we reasoned that incorporating the transcription factor binding site information directly into team FIRST’s model should be complementary to the method and could reveal interactions between transcription factors and sequence context. To test this idea, we included the transcription factor binding affinities for each promoter as additional features to those used by team FIRST (see Supplementary Table S3 for details on the features). We then trained a Support Vector Machine (SVM) using the combined features from both models. The resulting model provided predictions that had a correlation of 0.67 to the actual promoter activity and a combined score of 2.6 (Cp = 0.67298; X2 = 39.79601; Sp = 0.66815; R2 = 30.75429) (see Fig. 1E and Supplementary Data, DREAM6 Participants Predictions files), presenting a significant improvement and best performance compared to all the other teams or the aggregate of their predictions.

**Discussion**

The scoring and analysis of submitted predictions for the DREAM6 Gene Promoter Expression Prediction challenge revealed excellent performances (see Fig. 1E and Table 2). This is, indeed, remarkable, as the data set presented a difficult learning problem due to the high homology between the promoters in the relatively small RP promoter training set—yeast only has 137ribosomal promoters—and lower dynamic range of promoter activity compared to what would be observed on a genome-wide scale. Methods with typically high accuracy in genome-wide predictions ranked 11 and 12 here (see Supplemental Table S4), indicating that the challenge posed by RP promoters is distinct and requires the development of specific methods in order to be solved.

Choosing the right scoring scheme to evaluate the challenge was essential, as participants fared differently depending on the metric used (see Table 2). The best-performing team did not get the top score for all metrics nor all promoters but was the most consistent. Also, participants had difficulties while predicting low-expressed promoters and certain mutant RP promoters. Finally, community predictions were robust to the aggregation of teams' results, and the best score of 2.6 was obtained by combining features from team FIRST’s machine learning model and a mechanistic model based on biophysical assumptions.

During their presentations at the DREAM6 conference, the best-performing teams, FIRST and c4lab, showed that mutated promoters were harder to predict than natural promoters. Team FIRST mainly used the first 100 bp of the promoter to predict promoter activity, and team c4lab used a 12-mer motif. Team FIRST tried to include features such as k-mer counts (mono, di, tri, tetra, and penta), homopolymer stretches, promoter length, DNA bendability, DNA protein deformability, DNA bending stiffness, and nucleosome binding potential. They used a machine learning SVM approach to select 12 features that can be summarized as follows: one mononucleotide G, one dinucleotide GT, six trinucleotides, 12 tetranucleotides, length of T-tracks and TA-tracks, DNA deformability (a detailed description of this model will appear in a different manuscript). Team c4lab also used different k-mer counts to finally concentrate on 12-mer motifs used in a support vector regression approach but did not find any correlation between the 12-mers and biological features such as distance to a cPRES or copy number motifs (see Supplemental Table S4 for a brief description of other participants' methods).

Neither of the best-performing teams directly used general features related to transcription factors such as TATA boxes and nucleosome occluding sequences. Actually, none of the four biological features targeted by the mutations—TATA boxes, binding sites for the transcriptional regulators Pho1 and Sfp1, and mutations to nucleosome disfavoring sequences—were detected by the participants. Since most participants did not include these features in their models, it is not surprising that many fared worse with
promoters where these sequences were mutated. Figure 3, B and C show precisely that, as mutation-induced expression changes increase, predictions become worse. One exception is team FIRST’s machine learning method that was able to identify a number of nucleosome disfavoring features, in particular TA-tracts, as being useful in predicting promoter activity.

During the DREAM6 conference discussion, an audience member proposed that the training set should have included mutated promoter sequences. However, an intended feature of the challenge was to indicate that mutated sequences were present in the test set without giving hints or providing training data on sequence changes that could affect the promoter expression level.

We expected participants to analyze the origin of these mutations and think that our strategy was correct, as Figure 2A shows that, although participants did not look for the origin of mutated promoters, these were distributed equally between the groups of best-and worst-predicted promoters. It is only when all mutated and wild-type promoters are separated into two groups that participants’ predictions for those two groups can be differentiated (Fig. 3A).

The mechanism by which Fhl1, Sfp1, Rap1, and TATA boxes contribute to the promoter expression appears to follow a simple rule, where more sites from these factors in closer proximity to the
whose activity is higher than 3 are mutated promoters, shown to be difficult to predict. Low-activity promoters are RPLA1B_Mut1, RPLISA_Mut1, RPL21B, RPL4A_Mut6, RPL11A, RPL35B, RPL39_Mut1, and RPS14B_Mut1. As the experimental setup can distinguish promoter activities separated by less than 8%, we do not think that the difficulties with predicting low promoters arise from experimental limitations while measuring lower signals. Instead, as shown in Table 1, promoters RPLA1B_Mut1, RPL21B, RPL11A, RPL35B, RPL39_Mut1, and RPS14B_Mut1 have dispersed or lack binding motifs (see also Supplemental Table S5). The other mutations present in promoters of low activity are RPL4A_Mut6 and RPLISA_Mut1, which cause an 70% decrease in promoter activity, and as discussed, participants had difficulties predicting strong mutation effects. We conclude that the difficulty participants had while predicting low-expressed promoters is, indeed, due to less information available in these promoter sequences and a less coherent organization of the different sequence features, with very few TATA boxes, FH1, Rap, 1, and Sp1 sites.

Finally, the improvement of the best-performing model, by mixing a biology-based mechanistic approach and machine learning techniques, implies that the wisdom of crowds could be tapped further by methods that directly incorporate distinct features from independent models. Simple aggregation might miss the interactions between the different features in the models selected. Estimating the relative contributions of features extracted from each model could be approached as a learning problem where the different models are reduced to being independent tools for feature selection. Once the relevant features are selected, they are integrated into a new model, and adequate parameters are learned once again. Overall, we think this study not only provides a benchmark for the assessment of methods predicting promoter activity from sequence, but it also shows that understanding the basis of fine-tuned regulation of highly homologous promoters could provide clues for engineering promoter libraries to obtain a desired promoter strength from a parent promoter sequence.

Methods

Constructing promoter strains

A construct of ADH1 terminator–mCherry–TEF2 promoter–YFP–ADH1 terminator–NAT7 was inserted into the SGA-compatible strain Y8205 at the his3 deletion location (the construct replaced chromosome 15, at base pairs 721987–722506). The resulting strain served as a master strain for the entire library. Desired promoters were lifted by PCR from the BY4741 yeast strain. Primers contained one part matching the ends of the lifted promoters, and a constant part at their 5′ end matching the first 25 bases of the YFP gene (for reverse primers) or a linker sequence (for forward primers; see all primer sequences in Zeevi et al. 2011). Each promoter was linked to a URA3 selection marker (Linsbiz et al. 2008) and then amplified such that its genomic integration sites increased to 45/50 bp. Integration into the genome was performed by homologous recombination as described in Gietz and Schiestl (2007). All steps were performed on 96-well plates, except for growing the final clones, which was performed on six-well plates (2% agar, SCD–URA). To validate the inserted promoter sequences, the insertions were lifted from each target strain by PCR and sequenced.

Constructing promoter strains with targeted mutations

To create a mutated promoter, we amplified it in two parts which flank the desired mutation area. The left part was amplified using a reverse primer with a 35-bp tail at its 5′ end that contains the desired mutation, while the right part was amplified using a for-
ward primer that also had a similar tail. The two new parts, both containing the desired mutation in an overlapping region of 35 bp, were then connected, similar to the way in which we connected promoters to the URA3 selection marker. See Table 1 and Supplementary Table S6 for more information.

**Library measurements**

Cells were inoculated from stocks kept at 30°C into SCD (180 mL, 96-well plate) and left to grow at 30°C for 48 h, reaching complete saturation. Next, 8 mL were passed into fresh medium (180 mL) according to the desired condition (e.g., SCD, ethanol, heat shock). Measurements were carried out every 20 min using a robotic system (Tecan Freedom EVO) with a plate reader (Tecan Infinite F500). Each measurement included optical density (filter wavelengths 600 nm, bandwidth 10 nm), YFP fluorescence (excitation 500 nm, emission 540 nm, bandwidths 25/25 nm, accordingly), and mCherry fluorescence (excitation 570 nm, emission 630 nm, bandwidths 25/35 nm, accordingly). Measurements were carried out using a total of eight different conditions. In all experiments, yeast cells were grown on SC (6.9 g/L YNB, 1.6 g/L amino acids complete). Four conditions used different 2% sugar growth media: SC-glucose, SC-galactose, SC-ethanol, and SC-glycerol. The other four conditions used SC-glucose with an additional stress factor: Rapamycin (40 mg/mL), amino acid starvation (no amino acids except histidine and leucine), heat shock (39°C), and osmotic stress (750 mM KCl). Every strain was measured in three biological replicates for each condition. Most of the data analysis was performed on data from growth on SC-glucose (without stress), which was measured in five replicates.

**Scoring**

The challenge was scored in four different ways using criteria based on the “distance” between measured and predicted values or differences in rank between measured and predicted values. As we requested predictions of the expression levels from N = 53 promoter sequences, let us denote by $X_{ip}$ the predicted activity of promoter $i$ for participant $p$, and $j_1$ the measured activity of promoter $i = 1, 2, \ldots, 53$ and $p = 1, 2, \ldots, P$, where $P = 21$ is the number of teams that participated in the challenge. The score based on a Pearson metric for participant $p$ is defined by

$$C_p = \frac{\sum X_{ip} - j_1^2}{\sum X_{i}^2 - j_1^2}.$$

In order to calculate for each participant the probability of getting by chance a score at least as good, we randomly sampled the predictions across the entire set of participants. For each promoter $i = 1, 2, \ldots, 53$, we chose at random one of the $X_{ip}$ predictions, where $p = 1, 2, \ldots, P$. We thus obtained a value of $C_p$ which corresponded to one possible random choice of predictions among all the participants. By repeating the same process 100,000 times, we generated a null distribution of distances between measured and estimated values, from which a $P$-value can be estimated for $C_p$. For each participant, that $P$-value was denoted as $p_i$.

The score based on the $x^2$ metric for participant $p$ is defined by

$$d_{1}^{2} = \frac{\sum |X_{ip} - j_1|}{\sum |X_{i} - j_1|}.$$

The null hypothesis was generated in a similar way by generating $P$-values resulting from the permutation of participants' predicted values for a given promoter, and also for each participant, and that $P$-value was denoted as $p_j$.

We also defined the score by comparing the rank of predicted values to the actual rank of measured values. Let us denote by $R_{ip}$ the predicted rank of promoter $i$ for participant $p$, and $r_i$ the rank of the measured promoter $i = 1, 2, \ldots, 53$ and $p = 1, 2, \ldots, P$. Then, the score based on a Spearman metric for participant $p$ is defined by

$$S_p = \frac{\sum R_{ip} - r_i}{\sum R_{ip} - R_{i}}.$$

A null prediction was created by randomly permuting participants' predicted values for a given promoter and then ranking a given "random" participant $i$ to obtain the $R_{ip}$ ranks across the 53 different rankings of promoters, thus generating a distribution of distances between measured and estimated values, for which a $P$-value denoted as $p_j$ can be estimated for $S_p$. The score based on a rank$^2$ metric for participant $p$ is defined by

$$R_{ip}^2 = \frac{\sum R_{ip}^2 - \frac{r_i^2}{N}}{\sum R_{ip}^2 - \frac{R_{i}^2}{2}}.$$

where $j_1$ is the rank of proximity of $X_{ip}$ to $j_1, 1 < j_1 < \ldots < 53$ and $r_i$ the rank of the measured promoter $i = 1, 2, \ldots, 53$. The null hypothesis was derived from the random permutation of participants' predicted values for a given promoter and then ranking a given "random" participant. The derived $P$-value is denoted as $p_k$. The overall score was defined as a function of the product of all the $P$-values defined as

$$Score = \frac{1}{4} \log \frac{1}{P} \sum p_j.$$

**Prediction distances to promoter values**

The average distance $d_{1}^{2}$ over all participants $p$ for promoter $i$ from the promoter predicted value ($X_{ip}$) to the promoter measured value ($j_1$) is defined as

$$d_{1}^{2} = \frac{\sum |X_{ip} - j_1|^2}{\sum |j_1|^2}.$$

We also considered whether promoter activity was correlated to the difficulty to predict its value and used the following measure $x_i$ defined by

$$x_i = \frac{\sum |X_{ip} - j_1|^2}{\sum |X_{ip}|^2}.$$

We finally calculated a different type of distance $d_{1}^{2}$ to compare participant predictions and measurements, defined such that

$$d_{1}^{2} = \frac{\sum |X_{ip} - j_1|^2}{\sum |j_1|^2}.$$

**Combined model**

We considered binding sites for three transcription factors—Rap1 (Wade et al. 2004), Fhl1 (Harbison et al. 2004; Schwalder et al. 2004; Schwalder et al. 2004).
2004; Wade et al. 2004), Sfp1 (Badis et al. 2008; Zhu et al. 2009)—that have been shown to influence yeast ribosomal gene expression. Our model considered promoter activity Xp as directly proportional to the binding likelihood of each of the three transcription factors to their cognate motifs, above a specific threshold, relative to the nucleosome binding potential of the same sites:

\[ X_p \propto 1 + \frac{P(t)}{t} + \sum_{i} w_t P(t) b_S E_i \]

where \( P(t) \) is the set of all potential binding sites for transcription factor \( t \) above a certain threshold, \( w_1 \) is a coefficient measuring the relative contribution of factor \( t \) to the promoter activity determined using MATLAB's nonlinear solver, and \( P(t) = b_S E \) is the probability that transcription factor \( t \) binds its potential site at position \( i \) in promoter sequence \( S \). To determine the binding sites for the three transcription factors, we used their sequence specificities documented in position weight matrices (PWMs) (Sakai et al. 2004; Badis et al. 2008; Zhu et al. 2009). In estimating the binding threshold for each transcription factor, we explored the correlation between promoter activity and sites above each possible threshold at intervals of 0.1. For each transcription factor, we considered potential binding sites as those with an affinity above the threshold and located within known spatial localization sites: for Rap1, 400 bp upstream of the TrsS, for Hh1 and Sfp1, 300 bp upstream of the TrsS (Zeevi et al. 2011). We then modeled the probability for transcription factor binding as the weight of the configuration in which the factor is bound divided by the sum of the weight of that configuration, the weight of the configuration in which the DNA is unbound, and the weight of the configuration in which a nucleosome is bound to the site:

\[ P(t) b_S E_i \frac{A_p S E_i}{1 + A_p S E_i + A_{nuc} S E_i} \]

where 1 represents the DNA unbound configuration, \( A_p S E \) represents the affinity of transcription factor \( t \) for the binding site at position \( i \) in promoter \( S \), and \( A_{nuc} S E \) is the affinity of nucleosomes for position \( i \) in promoter \( S \).

For \( A_{nuc} S E \), we used a sequence-based nucleosome affinity model to compute the average nucleosome occupancy (Kaplan et al. 2009).

We applied \( w_1 \) coefficients obtained from a nonlinear solver trained on 90 promoters to predict promoter activities of a held-out set of 53 promoters used in the DREAM challenge.

**DREAM6 Consortium**


**Acknowledgments**

We thank Noam Kaplan from the Weizmann Institute of Science for computing the average nucleosome occupancies and Kahn Rhrissorarak for useful comments on the manuscript.

**References**


Received March 12, 2013; accepted in revised form August 14, 2013.