PromoterPredict: Sequence-based modelling of *Escherichia coli* σ⁷⁰ promoter strength yields logarithmic dependence between promoter strength and sequence

Ramit Bharanikumar [1] , **Keshav Aditya R Premkumar** [2] , **Ashok Palaniappan** [Corresp.] [3]

[1] Biotechnology, Sri Venkateswara College of Engineering, Sriperumbudur, Tamil Nadu, India

[2] Computer Science and Engineering, Sri Venkateswara College of Engineering, Sriperumbudur, Tamil Nadu, India

[3] School of Chemical and BioTechnology, SASTRA Deemed University, Thanjavur, Tamil Nadu 613401, India

Corresponding Author: Ashok Palaniappan
Email address: apalania@scbt.sastra.edu

We present PromoterPredict, a dynamic multiple regression approach to predict the strength of *Escherichia coli* promoters binding the σ⁷⁰ factor of RNA polymerase. σ⁷⁰ promoters are ubiquitously used in recombinant DNA technology, but characterizing their strength is demanding in terms of both time and money. Using a well-characterized set of promoters, we trained a multivariate linear regression model and found that the log of the promoter strength is significantly linearly associated with a weighted sum of the –10 and –35 sequence profile scores. It was found that the two regions contributed almost equally to the promoter strength. PromoterPredict accepts –10 and –35 hexamer sequences and returns the predicted promoter strength. It is capable of dynamic learning from user-supplied data to refine the model construction and yield more confident estimates of promoter strength. PromoterPredict is available as both a web service (https://promoterpredict.com) and standalone tool (https://github.com/PromoterPredict). Our work presents an intuitive generalization applicable to modelling the strength of other promoter classes.

1 **PromoterPredict: sequence-based modelling of *Escherichia coli* σ70**
2 **promoter strength yields logarithmic dependence between promoter**
3 **strength and sequence**

5 Ramit Bharanikumar[1], Keshav Aditya R. Premkumar[2] and Ashok Palaniappan[3*]
6 [1]Department of Biotechnology, Sri Venkateswara College of Engineering
7 (autonomous),Tamil Nadu 602117. India
8 [2]Dept of Computer Science and Engineering, Sri Venkateswara College of Engineering
9 (autonomous), Tamil Nadu 602117. India
10 [3]School of Chemical and Biotechnology, SASTRA Deemed University, Thanjavur, Tamil
11 Nadu 613401. India
12 [*]To whom correspondence should be addressed (apalania@scbt.sastra.edu)

14 **Abstract**: We present PromoterPredict, a dynamic multiple regression approach to
15 predict the strength of *Escherichia coli* promoters binding the σ70 factor of RNA
16 polymerase. σ70 promoters are ubiquitously used in recombinant DNA technology, but
17 characterizing their strength is demanding in terms of both time and money. Using a
18 well-characterized set of promoters, we trained a multivariate linear regression model
19 and found that the log of the promoter strength is significantly linearly associated with a
20 weighted sum of the −10 and −35 sequence profile scores. It was found that the two
21 regions contributed almost equally to the promoter strength. PromoterPredict accepts −
22 10 and −35 hexamer sequences and returns the predicted promoter strength. It is
23 capable of dynamic learning from user-supplied data to refine the model construction
24 and yield more confident estimates of promoter strength. PromoterPredict is available
25 as both a web service (https://promoterpredict.com) and standalone tool
26 (https://github.com/PromoterPredict). Our work presents an intuitive generalization
27 applicable to modelling the strength of other promoter classes.

28 **Availability**: Open source code and a standalone executable with both dynamic model-
29 building and prediction are available (under GNU General Public License 3.0) at
30 https://github.com/PromoterPredict, and require Python 2.7 or greater.
31 PromoterPredict is also available as a web service at https://promoterpredict.com.

32 **Contact**: apalania@scbt.sastra.edu

INTRODUCTION

33
34

35 The primary *E. coli* promoter-specificity factor and the one widely used in recombinant
36 DNA technology is the σ^{70} factor. Promoters recognized by σ^{70}-containing RNA
37 polymerase are called core promoters and share the following features: two conserved
38 hexamer sequences, separated by a non-specific spacer of ideally 17 nucleotides. The two
39 hexamers are located ~10 bp and ~35 bp upstream of the transcription start site, and
40 are called the −10 and −35 sequences respectively (Paget and Helmann, 2003;
41 Kadonaga, 2012). Promoters with −10 and −35 sequences matching the consensus motif
42 of the hexamers are typically stronger, meaning they initiate more transcripts per unit
43 time than promoters with less canonical −10 and −35 regions. It is known that the
44 conserved hexamer regions are vital for recognizing and optimizing the interactions
45 between DNA and the RNA polymerase (Hook-Barnard *et al.*, 2006; Feklistov and
46 Darst, 2011; Basu *et al.*, 2014).

47 Theory has yielded a linear relationship between the total promoter score and the
48 natural log of promoter strength (Berg and von Hippel, 1987). Strength of *E. coli* σ^E RNA
49 polymerase promoters were studied by Rhodius and Mutalik (2010), who suggested
50 that a study of core (i.e., σ^{70}) promoters of housekeeping genes could be complicated by
51 the additional role of transcription activators and limited data on promoter strengths.
52 The complexity of *E. coli* σ^{70} promoter sequences has been treated from an information
53 theoretic standpoint by Shultzaberger *et al.* (2007). Many resources are available to
54 predict the location of promoters in a genomic seqeunce mainly by identifying the −10
55 and −35 regulatory sequences (for example, de Jong *et al.* (2012)), but there is no
56 (freely) available tool to predict the strength of such sequences. Here we provide a web
57 based platform as well as a standalone tool for the predictive modelling of the strength
58 of σ^{70} core promoters, with the option to dynamically include user data into the
59 predictive model.

METHODS

60
61

62 ***Generative model of promoter sequences***. A generative model of the −10 and −35
63 promoter sequences is constructed using two Position Weight Matrices (PWM_{-10} and
64 PWM_{-35}) in the following manner. The training set is drawn from the well-characterised
65 Anderson collection of 19 activator-independent promoters maintained at the Registry
66 of standard biological parts (http://parts.igem.org/Promoters/Catalog/Anderson).
67 Nucleotide-wise counts at each position of the hexamer motifs were augmented by a
68 pseudo-count prior to correct for *E. coli* GC content of 50.8% and the resulting
69 frequency matrices were converted into log-odds matrices using Biopython
70 (www.biopython.org).
71

72 ***Linear modelling of promoter strength***. Following Berg and von Hippel (1987),
73 we modelled the relationship between the promoter sequences and the *ln* of the
74 promoter strength using multiple linear regression. Each promoter sequence is scored
75 with respect to the generative models of the −10 and −35 motifs (i.e., the PWM_{-10} and
76 PWM_{-35} matrices) and the two scores obtained formed the feature space of the
77 regression modelling. The regression coefficients to be determined represent the
78 weights of the -10 and -35 regions in the regression analysis. The Anderson promoter
79 library provided promoter strengths normalized in the range 0.00 to 1.00 with respect
80 to the strongest promoter. It was noted that the normalisation step would not affect a
81 linear relationship, altering only the constant of the regression. The normalised strength
82 values were log-transformed to obtain the required response variable values. Since the
83 *ln* function rapidly descends towards − Inf with decreasing promoter strength, we
84 capped the infimum of promoter strength at 0.01 prior to log-transformation. The least-
85 squares cost function was minimized using iterative gradient descent. The model
86 parameters were assessed using t-statistics, and the overall model was assessed using F-
87 statistic and the adjusted multiple coefficient of determination given by:

88 Adj. $R^2 = 1 − \{(1-R^2)*[(n-1)/(n-m-1)]\}$ ...(1)

89 The model was validated using leave-one-out cross-validation (LOOCV) .

90 **RESULTS**

91 The conservation profile of the −35 and −10 hexamer sequences of the promoters in the
92 Anderson library was visualized using sequence logos and shown in Fig. 1. The site
93 scores of each promoter sequence were regressed on the ln of the promoter strength. A
94 summary of this process with the training data, log-transformation of the promoter
95 strength and predicted response values is presented in Table 1. The modelling process
96 converged within 10^5 iterations by tuning the gradient descent to a learning rate (α) of
97 0.015, and the following model was obtained:

98 *ln* (promoter strength) = -17.111 + 1.015*(PWM_{-35}) + 0.949*(PWM_{-10}) ...(2)

99 It is observed that the weight coefficients of the two PWM features are almost equal. We
100 derived an independent solution of the multiple regression using `R` (www.r-project.org)
101 and obtained a correlation coefficient of 0.998 between the fitted values of the two
102 models. The interval estimates of the coefficients of the regression were computed in `R`
103 using `confint(fit, level=0.95)`, and obtained the following 95% confidence
104 intervals:

105 `Intercept  :    (-24.05737951, -9.625336)`

106 `−35        :    (0.59755850, 1.377074)`

107   `-10`   `:`   `(0.07977664 , 1.814886)`

108

109   The interval estimates did not include zero, and this implied that the coefficients were
110   significant at the 0.05 level. The p-value of the PWM_{-35} coefficient was $< 10^{-4}$ and that of
111   $PWM_{-10} \approx 0.03$. The intercept was significant at a p-value $\approx 10^{-4}$. The F-statistic of the
112   overall regression was significant at a p-value $< 10^{-4}$ and adj. R^2 was ≈ 0.65. The plane of
113   best fit corresponding to the above model is visualized in Fig. 2.

114   The model was then cross-validated using a 19-fold LOOCV (similar to jack-knife).
115   Cross-validation yielded a correlation coefficient of 0.75 (Table 2).

116   **DISCUSSION**

117   In addition to the independent contributions of -35 and -10 sites to promoter strength,
118   we were interested in exploring if any interactions between them could contribute to
119   promoter strength. To this end, we examined the following model in `R`:

120   `lm(logStrength ~ PWM35 * PWM10)`

121   where `PWM35` and `PWM10` represent the corrresponding site scores. This model
122   resulted in a lower adj. R^2 value than that without any interactions. All the four p-values
123   of the regression parameters (intercept, `PWM35`, `PWM10` and interaction) were not
124   significant. The F-statistic was also not significant, thus discounting any interaction
125   between the sites in the present dataset. On this basis, the null hypothesis of absence of
126   any interaction could not be rejected, and we concluded that there is little evidence for
127   interaction between the -35 and -10 sites in determining promoter strength.

128   Our model assumed that both the predictors carried independent information about the
129   promoter strength, and together they are able to provide sufficient information about
130   the strength. The basis of this assumption was probed to determine if both predictors
131   are necessary to the model. Could one predictor provide sufficient information about the
132   promoter strength in the absence of the other? There are at least three angles to address
133   this question, and all of them were considered to interpret the model better.

134   (1) Comparing the raw, unadjusted R^2 with the adjusted R^2. The corresponding values
135   were:

136   $R^2 = 0.69$

137   Adj. $R^2 \approx 0.65$

138 Since there is not much difference between R² and adj. R², we could say that both
139 predictors contribute substantially to the response variable (promoter strength) and
140 account for about 65% of its variance.

141 (2) Since the p-values of both predictors are significant, it would be interesting to
142 observe their effect on the response variable in more detail. This was performed using
143 the `effects` package in `R`:

144 `library(effects)`

145 `fit = lm(logStrength~ PWM35+ PWM10, data)`

146 `plot(allEffects(fit))`

147 The results are shown in Fig. 3. Confidence in the effect of −35 site increases with the −
148 35 score, as evidenced by decreasing uncertainty in logStrength. Such an effect is
149 however not observed for −10 hexamer: the uncertainty widens at both the ends due to
150 edge effects. The effect of the −35 sequence is also steeper than the effect of the −10
151 sequence.

152 (3) Another robust method to address the question is to compute the correlation
153 coefficients between all the variables of interest, including a variable with the combined
154 effects of −35 and −10 sites. This is shown in Table 3. Three features were used, namely
155 PWM_{-10} score, PWM_{-35} score, and the combined score (i.e., $PWM_{-10} + PWM_{-35}$). These
156 feature variables were correlated with two response variables, namely promoter strength
157 and its corresponding log transformation. It was first observed that the PWM_{-10} and
158 PWM_{-35} scores were uncorrelated (with a correlation coefficient of just ~0.05). It was
159 significant that the highest correlation between the features and response variable was
160 observed between the combined score and log of the promoter strength (~0.83). This
161 validated our modelling process and was in keeping with similar observations for the
162 strength of σ^E promoters (Rhodius and Mutalik, 2010). It was further observed that the
163 combined score showed a relatively moderate correlation with the promoter strength
164 prior to log transformation (about 0.66). This underscored the logarithmic dependence
165 between the promoter strength and sequence, and provided independent validation of
166 Berg and von Hippel's theoretical model (1987).

167 Finally, the assumptions of linear modelling were investigated with reference to our
168 problem. Model diagnostics of four basic assumptions were plotted (shown in Fig. 4).
169 Specifically:

170 Plot 1: The residuals were plotted against the fitted values. No trend was visible in the
171 plot, indicating the residuals did not increase with the fitted values and followed a
172 random pattern about zero. This validated the assumption that the errors were
173 independent.

174 Plot 2: The square root of the relative error (standardized residual) was plotted against
175 the fitted value. No distinct trend was observed, indicating that the standardized
176 residual was not a function of the fitted value. This further validated the assumption
177 that the errors were independent.

178 Plot 3: To test the assumption that the errors were normally distributed, the
179 standardized residuals were plotted against the theoretical quantiles of a normal
180 distribution. The residual distribution did not significantly deviate from the theoretical
181 quantiles.

182 Plot 4: Since the least-squares cost function is sensitive to outliers, the number of
183 outliers should be kept to a minimum. This was investigated by plotting the
184 standardized residual against the corresponding instance's model leverage. This plot
185 showed that there were no significant outliers in the dataset that could exert an undue
186 influence on the regression parameters.

187 An alternative univariate regression model using only the combined score of the PWMs
188 found the coefficient of regression significant (p-value $< 10^{-4}$). However, the weights of
189 the PWMs were slightly different in the model equation (eq. (2)), further the
190 uncertainty in their effects was different. The original multiple linear regression model
191 was retained for the estimation of the promoter strength.

192 We implemented our model in Python (www.python.org). Since the modelling results
193 are dependent on the dataset, our implementation provides a facility to augment the
194 learning based on user-provided inputs. A web service for the same has been initiated.
195 The web interface is based on Python web module (web.py) and nginx server. The
196 computational layer is based on numpy, Biopython and matplotlib. The user is provided
197 with an option to add any number of promoter instances with −10 and −35 sequences
198 and the corresponding strengths to augment the training data of the supervised model.
199 The goodness of fit of the updated model is re-computed, along with a 3D plot of the
200 regression surface. Based on the trained model, the user could predict the strength of
201 any uncharacterised promoter given its −10 and −35 hexamers.

202 **CONCLUSION**

203 The following important conclusions were drawn from our study. (1) Sequence-based
204 modelling yielded a logarithmic dependence between promoter strength and sequence.
205 (2) The −10 and −35 sites were equally important in determining promoter strength. (3)
206 The combined sum of the scores ($PWM_{-35} + PWM_{-10}$) emerged as the single most
207 important predictor of the promoter strength. It is straighforward to extend our
208 methodology to the study of new promoter classes of other σ factors. Our
209 implementation and web service could be useful in characterizing unknown promoters
210 of newly sequenced genomes as well in the selection of promoters for synthetic biology

211 experiments. The dynamic feature of our implementation would enable users with own
212 data to obtain more reliable estimates of promoter strength. The service will be
213 periodically updated based on the availability of new training instances, user input data
214 and/or models for promoters of other σ factors.
215

216 **Acknowledgments**

217 We would like to thank computing facilities at SASTRA Deemed University for support.
218

219 **REFERENCES**
220

221 Berg, O.G. and von Hippel, P.H. (1987). Selection of DNA binding sites by regulatory
222 proteins. Statistical-mechanical theory and application to operators and promoters. J
223 Mol Biol **193**:723–750.

224 Basu,R.S. et al. (2014) Structural Basis of Transcription Initiation by Bacterial
225 RNAPolymerase Holoenzyme. J Biol Chem **289**: 24549 –24559

226 Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: A sequence logo
227 generator, *Genome Research*, **14**:1188-1190

228 de Jong, A. et al. (2012) PePPER: a webserver for prediction of prokaryote promoter
229 elements and regulons. BMC Genomics **13**:299

230 Feklistov, A. and Darst, S.A. (2011) Structural Basis for Promoter −10 Element
231 Recognition by the Bacterial RNA Polymerase σ Subunit. Cell **147**: 1257–1269

232 Hook-Barnard, I. *et al.* (2006) *Escherichia coli* RNA Polymerase Recognition of a σ70-
233 Dependent Promoter Requiring a −35 DNA Element and an Extended −10 TGn Motif. J
234 Bacteriol. **188**:8352–8359.

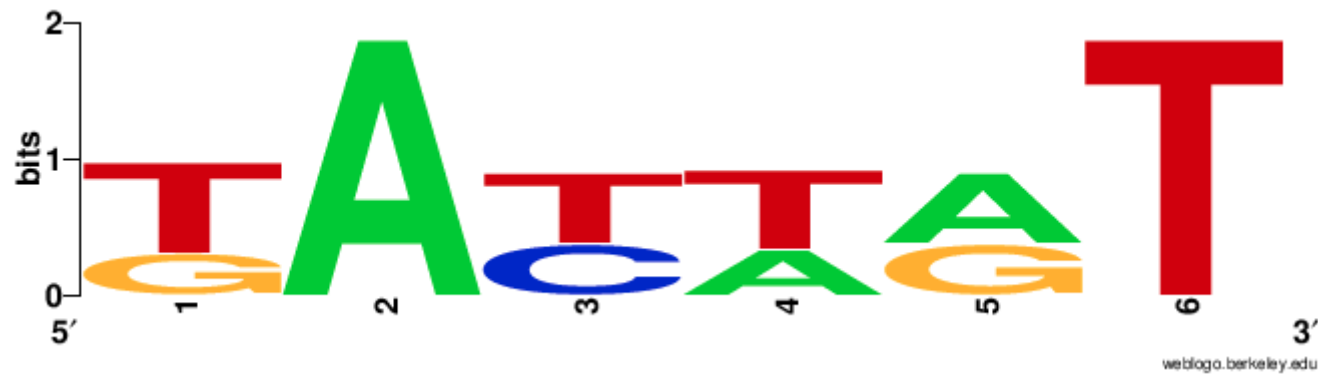
235 Kadonaga, J.T. (2012) Perspectives on the RNA Polymerase II Core Promoter. Wiley
236 interdisciplinary reviews Developmental biology. **1**:40-51.

237 Paget, M.S. and Helmann, J.D. (2003). The σ70 family of sigma factors. Genome Biology
238 **4**:203.

239 Rhodius, V.A. and Mutalik, V.K. (2010) Predicting strength and function for promoters
240 of the *Escherichia coli* alternate sigma factor, σE. Proc. Natl. Acad. Sci. USA **107**: 2854-
241 2859

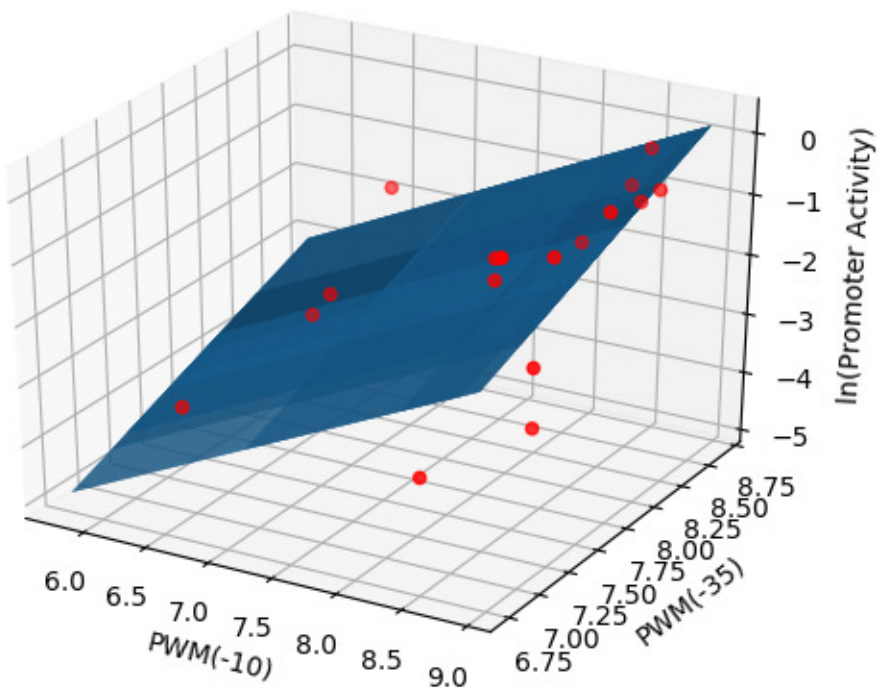
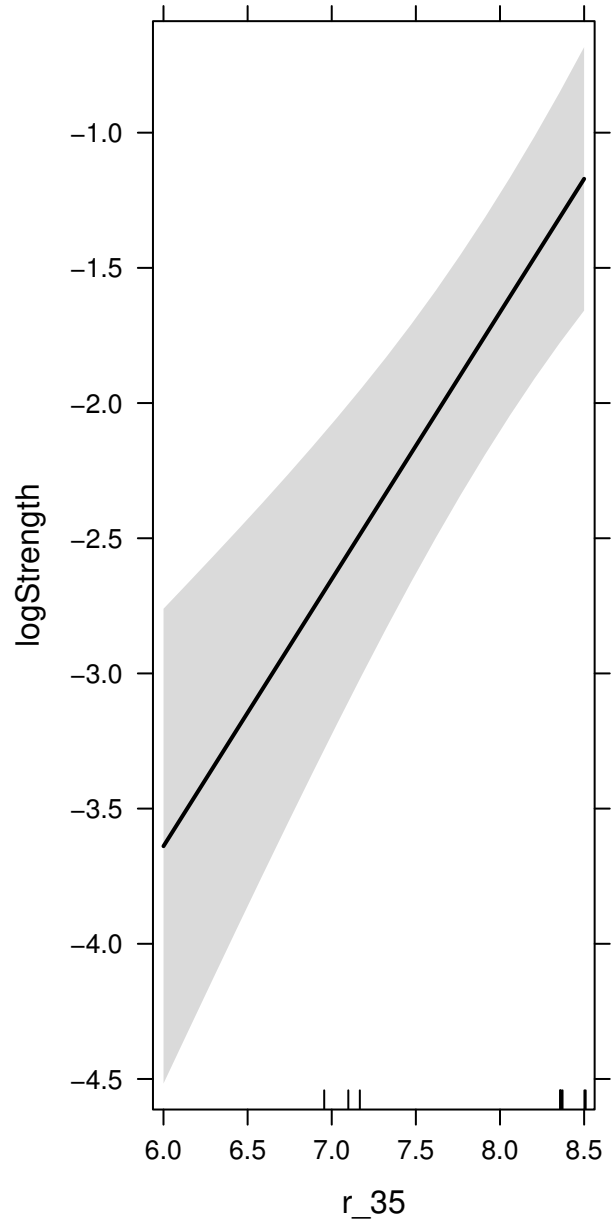
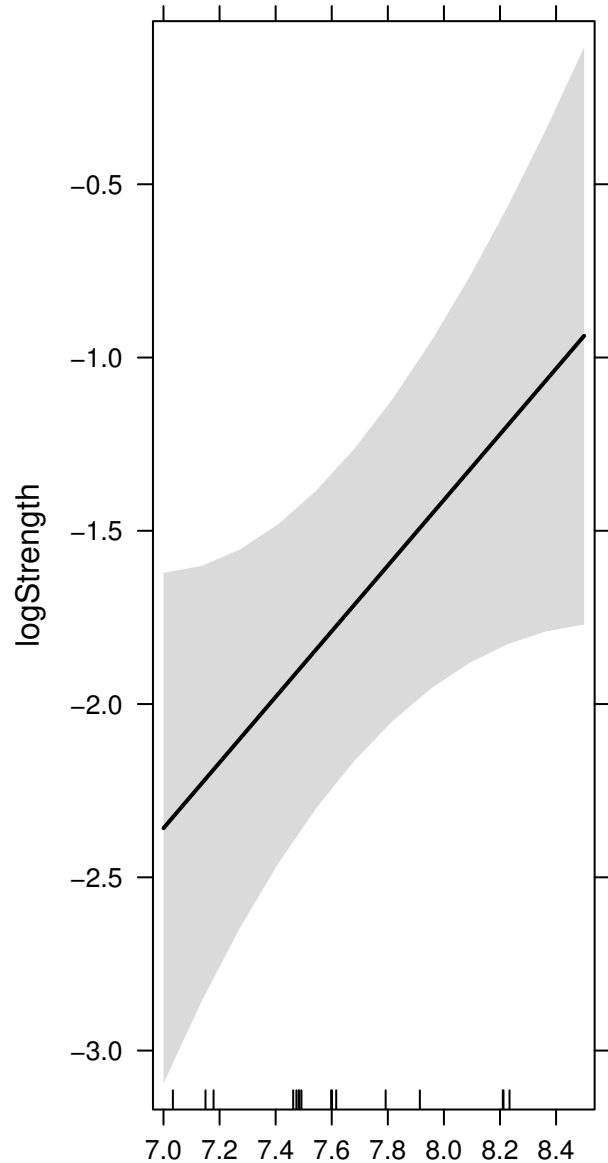
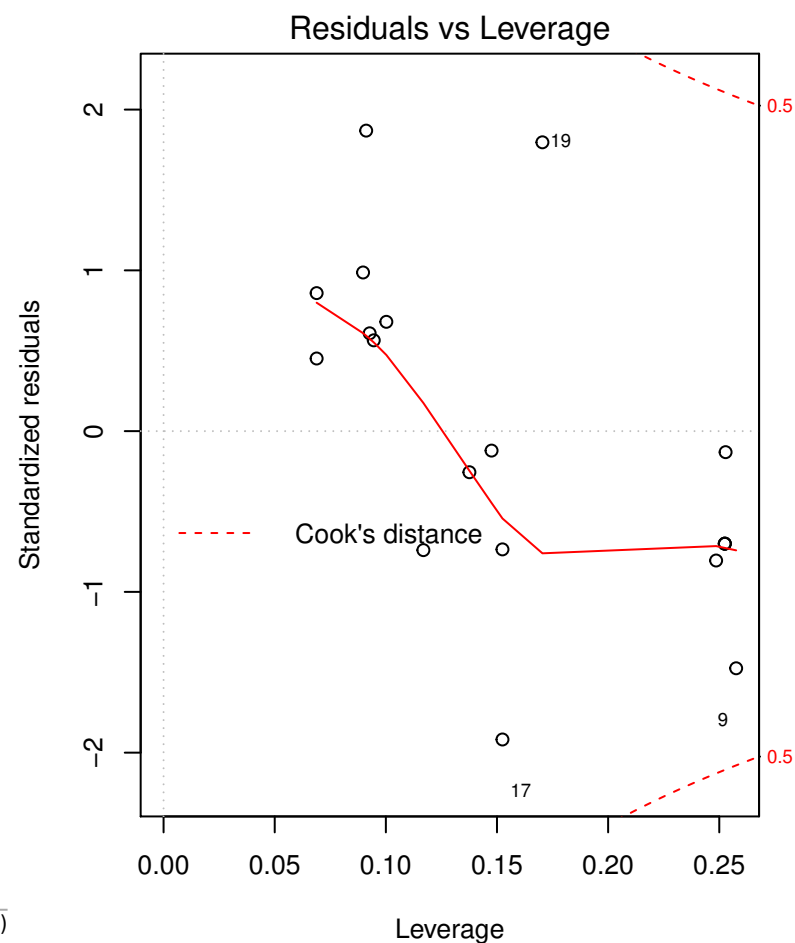
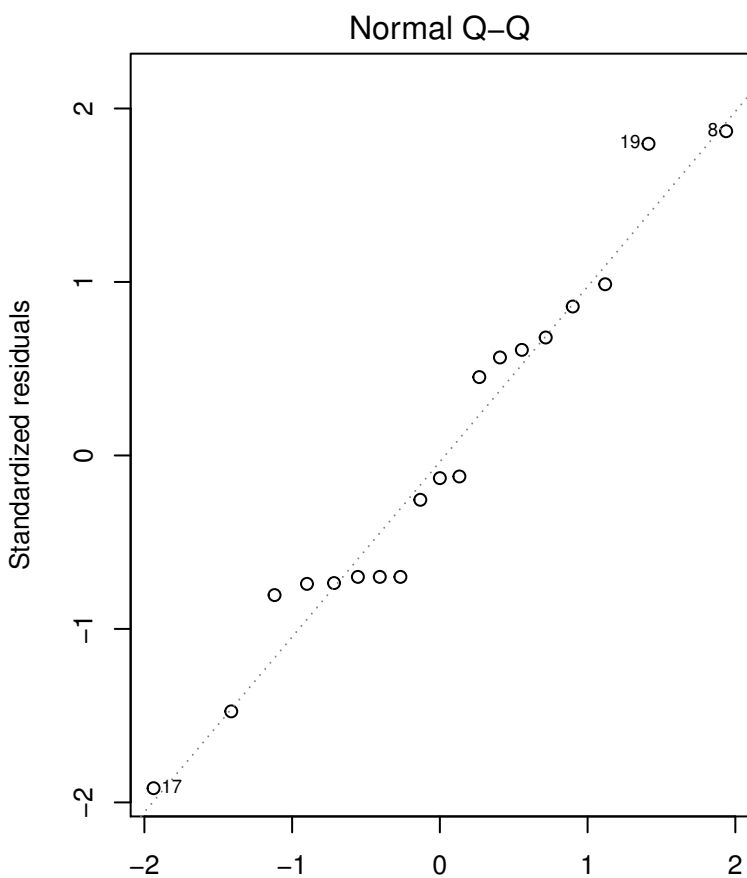
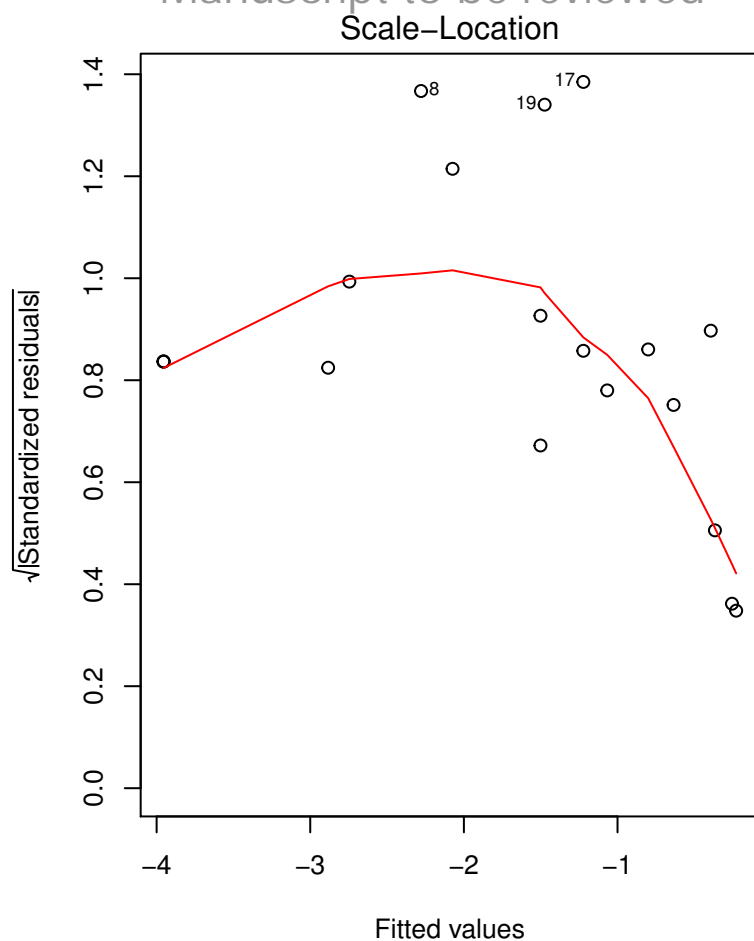
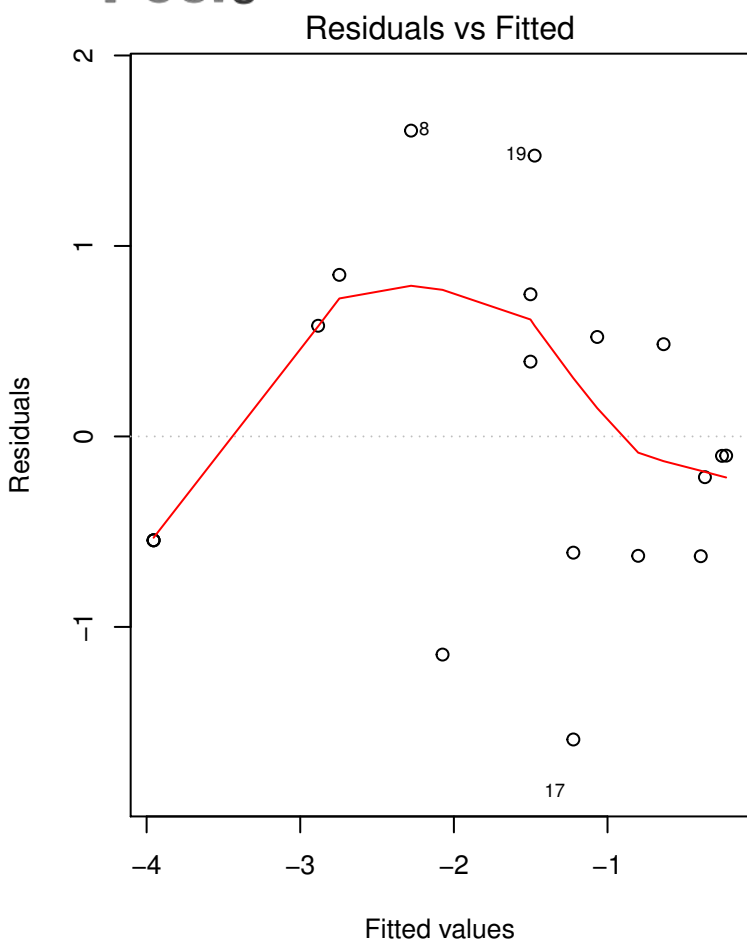
242 Shultzaberger, R.K. *et al.* (2007) Anatomy of *Escherichia coli* sigma70 promoters.
243 Nucleic Acids Res **35**:771–788.

Sequence logos of the –35 and –10 hexamer sequences of the promoters in the Anderson library.

Figure was made using WebLogo (Crooks *et al.*, 2004).

(a) -10 motif

(b) -35 motif

The regression surface (blue) of the estimated model with the training data points(red).

X- and y-axes represent PWM scores and the z-axis (vertical) represents the predicted *ln*(promoter strength).

Effects plots of –35 and –10 promoter sites on promoter strength.

Model diagnostics plots for investigating the assumptions underlying linear modelling.

Summary of promoter information.

The promoter activities (strengths) are seen to span the range [0.0, 1.0]. * indicates promoter strength capped at 0.01. The promoters follow the naming in the Anderson dataset.

Promoter	PWM$_{-35}$ score	PWM$_{-10}$ score	Promoter Activity	ln(Promoter Activity)	Predicted ln(promoter activity)
BBa_J23100	8.80192196	7.05252226	1	0	-1.4669153
BBa_J23101	8.50436801	8.65655364	0.7	-0.35667494	-0.25855671
BBa_J23102	8.94341939	7.79025811	0.86	-0.15082289	-0.62881141
BBa_J23103	5.76111212	7.60539274	0.01	-4.60517019	-4.0308767
BBa_J23104	8.94341939	8.22340587	0.72	-0.32850407	-0.22100527
BBa_J23105	8.36287058	8.22340587	0.24	-1.42711636	-0.80989265
BBa_J23106	8.36287058	7.48567002	0.47	-0.75502258	-1.50446674
BBa_J23107	8.36287058	8.65655364	0.36	-1.02165125	-0.40208651
BBa_J23108	7.16328158	7.91881778	0.51	-0.67334455	-2.31347961
BBa_J23109	8.50436801	6.73909721	0.04	-3.21887582	-2.06383098
BBa_J23110	8.36287058	7.48567002	0.33	-1.10866262	-1.50446674
BBa_J23111	8.80192196	7.48567002	0.58	-0.54472718	-1.05910916
BBa_J23112	5.76111212	7.60539274	0.00*	-4.60517019	-4.0308767
BBa_J23113	5.76111212	7.60539274	0.01	-4.60517019	-4.0308767
BBa_J23114	6.96070112	7.48567002	0.1	-2.30258509	-2.92677594
BBa_J23115	7.10219855	7.48567002	0.15	-1.89711998	-2.78324614
BBa_J23116	8.94341939	7.17224497	0.16	-1.83258146	-1.21066725
BBa_J23117	8.94341939	7.17224497	0.06	-2.81341072	-1.21066725
BBa_J23118	8.80192196	8.22340587	0.56	-0.5798185	-0.36453507

Correlation matrix of features and response variables.

Corr. Coef.	PWM_{-10}	PWM_{-35}	Combined	Strength	Log-strength
PWM_{-10}	1	0.05223042	0.44876850	0.31830991	0.36182231
PWM_{-35}	0.05223042	1	0.91586764	0.59734821	0.76551138
Combined	0.4487685	0.9158676	1	0.6625185	0.8304959
Strength	0.3183099	0.5973482	0.6625185	1	0.8744839
Log-strength	0.3618223	0.7655114	0.8304959	0.8744839	1

Cross-validation results.

In each trial, a random observation was chosen as a test instance for prediction based on a model built with the rest of the dataset. This process was repeated 19 times, once for each test instance and the cross-validation (CV) residuals were obtained.

Fold	Observation	Log(strength)	Predicted	CV Pred	CV Residual
1	3	-4.60	-3.99	-3.77	-0.73
2	15	-1.89	-2.74	-2.82	0.93
3	1	-0.35	-0.25	-0.22	-0.13
4	10	-1.10	-1.50	-1.53	0.43
5	12	-4.60	-3.95	-3.77	-0.73
6	2	-0.15	-0.63	-0.68	0.53
7	4	-0.32	-0.22	-0.21	-0.11
8	00	0.00	-1.47	-1.78	1.78
9	14	-2.30	-2.88	-2.94	0.64
10	5	-1.42	-0.80	-0.71	-0.71
11	13	-4.60	-3.95	-3.77	-0.73
12	7	-1.02	-0.39	-0.18	-0.84
13	17	-2.81	-1.22	-0.93	-1.88
14	8	-0.67	-2.27	-2.44	1.77
15	9	-3.22	-2.07	-1.68	-1.54
16	11	-0.54	-1.06	-1.12	0.58
17	16	-1.83	-1.22	-1.11	-0.72
18	6	-0.75	-1.50	-1.55	0.80
19	18	-0.58	-0.36	-0.33	-0.25