Predicting Transcript Production Rates in Yeast With Sparse Linear Models

by

Yezhou Huang

Department of Computer Science
Duke University

Date: ______________

Approved:

Alexander J. Hartemink, Supervisor

Raluca Gordan

L. Ryan Baugh

Sayan Mukherjee

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computer Science in the Graduate School of Duke University 2016
Abstract
Predicting Transcript Production Rates in Yeast With Sparse Linear Models

by

Yezhou Huang

Department of Computer Science
Duke University

Date: ____________________
Approved:

________________________
Alexander J. Hartemink, Supervisor

________________________
Raluca Gordan

________________________
L. Ryan Baugh

________________________
Sayan Mukherjee

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computer Science in the Graduate School of Duke University 2016
Abstract

To provide biological insights into transcriptional regulation, a couple of groups have recently presented models relating the promoter DNA-bound transcription factors (TFs) to downstream gene’s mean transcript level or transcript production rates over time. However, transcript production is dynamic in response to changes of TF concentrations over time. Also, TFs are not the only factors binding to promoters; other DNA binding factors (DBFs) bind as well, especially nucleosomes, resulting in competition between DBFs for binding at same genomic location. Additionally, not only TFs, but also some other elements regulate transcription. Within core promoter, various regulatory elements influence RNAPII recruitment, PIC formation, RNAPII searching for TSS, and RNAPII initiating transcription. Moreover, it is proposed that downstream from TSS, nucleosomes resist RNAPII elongation.

Here, we provide a machine learning framework to predict transcript production rates from DNA sequences. We applied this framework in the *S. cerevisiae* yeast for two scenarios: a) to predict the dynamic transcript production rate during the cell cycle for native promoters; b) to predict the mean transcript production rate over time for synthetic promoters. As far as we know, our framework is the first successful attempt to have a model that can predict dynamic transcript production rates from DNA sequences only: with cell cycle data set, we got Pearson correlation coefficient $C_p = 0.751$ and coefficient of determination $r^2 = 0.564$ on test set for predicting dynamic transcript production rate over time. Also, for DREAM6 Gene Promoter
Expression Prediction challenge, our fitted model outperformed all participant teams, best of all teams, and a model combining best team’s $k$-mer based sequence features and another paper’s biologically mechanistic features, in terms of all scoring metrics.

Moreover, our framework shows its capability of identifying generalizable features by interpreting the highly predictive models, and thereby provide support for associated hypothesized mechanisms about transcriptional regulation. With the learned sparse linear models, we got results supporting the following biological insights: a) TFs govern the probability of RNAPII recruitment and initiation possibly through interactions with PIC components and transcription cofactors; b) the core promoter amplifies the transcript production probably by influencing PIC formation, RNAPII recruitment, DNA melting, RNAPII searching for and selecting TSS, releasing RNAPII from general transcription factors, and thereby initiation; c) there is strong transcriptional synergy between TFs and core promoter elements; d) the regulatory elements within core promoter region are more than TATA box and nucleosome free region, suggesting the existence of still unidentified TAF-dependent and cofactor-dependent core promoter elements in yeast $S.\ cerevisiae$; e) nucleosome occupancy is helpful for representing $-1$ and $+1$ nucleosomes’ regulatory roles on transcription.
To my family, without whom I would never have come as far.
Contents

Abstract iv
List of Tables xii
List of Figures xiv
List of Abbreviations and Symbols xxvi
Acknowledgements xxviii

1 Introduction 1

1.1 Gene expression and the regulation of gene expression in eukaryotes 6

1.2 Regulation of transcription initiation in eukaryotes: introduction 8

1.3 Regulation of transcription initiation in eukaryotes: core promoter and the assembly model of general transcription factors in vitro 10

1.3.1 Only part of the regulatory elements are present in each core promoter 13

1.3.2 Other core promoter regulatory elements are reported 16

1.3.3 Core promoter architectures are different among distinct eukaryotes 16

1.4 Regulation of transcription initiation: complex transcription initiation in vivo 17

1.4.1 TF and its binding preference representation 18

1.4.2 TFs can regulate transcription in various ways 20

1.4.3 The ways in which TFs can act as activators 22

1.4.4 The ways in which TFs can act as repressors 25
1.4.5 The big picture of eukaryotic transcription initiation *in vivo*  
1.5 Regulation of transcription elongation in eukaryotes  
1.6 Regulation of transcription termination in eukaryotes  
1.7 A focus on transcriptional regulation in yeast *S. cerevisiae*  
1.8 Dissertation outline  

2 Hypothesized transcriptional regulation mechanisms in yeast *S. cerevisiae*  
2.1 Hypothesized mechanisms relevant to TF binding to DNA  
2.2 Hypothesized mechanisms relevant to the contribution of TF to transcript production  
2.3 Hypothesized mechanisms relevant to regulatory elements within core promoter  
2.4 Hypothesized mechanisms relevant to the transcriptional synergy between TFs and core promoter regulatory elements  
2.5 Hypothesized mechanisms relevant to the nucleosomes resisting RNAPII elongation  

3 Our feature engineering protocol: extracting and hierarchically grouping features  
3.1 “Fluctuation” group of TF relevant features  
   3.1.1 Distance decay of TF binding instance relative to TSS  
   3.1.2 Stability weight of TF binding instance determined by nucleosome occupancy nearby  
3.2 “Amplifier” group of core promoter relevant features  
   3.2.1 \( k \)-mer count features representing regulatory elements within the core promoter  
   3.2.2 occupancy features representing regulatory elements within the core promoter  
3.3 “Resistance” group of gene body nucleosome features  
3.4 Grouping of main effect features
3.5 Preprocessing feature values ........................................... 75

4 PreTrans: a machine learning framework for predicting transcript 
production rates .......................................................... 76

4.1 The components of PreTrans ........................................... 77

4.2 Linear models representing hypothesized mechanisms .......... 78

4.3 Sparse linear models and regularization techniques ............... 81

4.3.1 The lasso regularization ........................................... 81

4.3.2 The elastic net regularization ................................... 82

4.3.3 The ridge regularization ........................................... 82

4.3.4 The group lasso and its variants ............................... 83

4.4 Our algorithm: a sparse version of hierarchical group ridge .... 85

4.4.1 Step A: determining the best subgroup of features for each 
main effect group. ......................................................... 86

4.4.2 Step B: determining the helpful groups of features. .......... 88

4.4.3 Step C: selecting the best model in terms of predictive perfor-
mina ce. .................................................................. 90

4.4.4 Optional Step D: determining the helpfulness of an additional 
group of features. ....................................................... 91

5 Preparing for predicting yeast dynamic transcript production rates 
in cell cycle ................................................................. 94

5.1 Choosing TFs and genes for cell cycle data set .................... 95

5.2 Preparing inputs for COMPETE to extract features from occupancy 
profiles ................................................................. 96

5.2.1 Sequence specificities ........................................... 97

5.2.2 TF concentrations ............................................... 97

5.2.3 DNA sequences ................................................ 99

5.3 Autocorrelation analysis of transcript production rates throughout cell 
cycle ................................................................. 100
6 Applying PreTrans to predict yeast dynamic transcript production rates in cell cycle

6.1 Excellent results of PreTrans on cell cycle data set

6.2 The gRFE path of feature elimination on DREAM6 data set

6.3 Testing results of hypothesized mechanisms relevant to TFs

   6.3.1 TFs control the fluctuation of transcript production

   6.3.2 Nucleosome’s indirect effects on transcription through TFs are ignorable

   6.3.3 Stability weight, saturation, and distance decay are not helpful for predictive modeling of transcript production rates in cell cycle

6.4 Testing results of hypothesized mechanisms relevant to core promoter elements

   6.4.1 The core promoter amplifies transcript production rate

   6.4.2 The “kCCP” group of features reflect mechanisms more than TATA element, −1 nucleosome, and +1 nucleosome

6.5 Testing results of hypothesized mechanisms about synergistic interactions between TFs and core promoter elements

6.6 Testing results of hypothesized mechanisms about nucleosomes resisting RNAPII elongation

7 Preparing for predicting yeast mean transcript production rate over time for synthetic promoters

7.1 Introduction to DREAM6 data sets

7.2 Scoring metrics in DREAM6 gene expression prediction challenge

7.3 Preparing inputs for COMPETE to extract features from occupancy profiles

8 Applying PreTrans to yeast synthetic promoter sequences

8.1 PreTrans beats all teams participated in the DREAM6 challenge
8.2 The gRFE path of feature elimination on DREAM6 data set 139
8.3 Interpreting our final model ................................. 142
8.4 COMPETE is helpful on feature engineering through providing occupancy profile ............................... 143

9 Discussion and future directions 145

9.1 Our contributions ................................................. 145
9.1.1 First successful attempt to have a dynamic model of transcript production and transcriptional regulation 145
9.1.2 A framework that can test mechanistic hypotheses and identify generalizable features .......................... 146
9.1.3 Presenting the synergistic interactions between TFs and core promoter elements .............................. 148

9.2 Improving our framework ........................................ 149
9.2.1 Improving the application of our framework in yeast S. cerevisiae 149
9.2.2 Improving our framework for its application to metazoans . 150

A k-mer count sequence features extracted from the core promoter 151

B Supplement for cell cycle study ............................... 154

C Supplement for DREAM6 study ............................... 162

Bibliography 171

Biography 182
List of Tables

1.1 Example PSSM of MBF (Mbp1p/Swi6p) in yeast *S. cerevisiae*. Please note that the Mbp1p is the binding component of MBF (Mbp1p/Swi6p). ................................................................. 20

4.1 Hypotheses associated with groups of features for main effects and interactions. ................................................................. 92

4.2 Hypotheses associated with various linear combinations of main effects and interactions. ................................................................. 93

8.1 The scores of PreTrans, being compared with other models. “BOA” is abbreviation for best of all teams. Cp, X2, Sp, R2 are defined in Equations 7.1, 7.2, 7.3, and 7.4 respectively. For Cp and Sp, the higher the better. For X2 and R2, the lower the better. ................................................................. 139

8.2 Fitting models with only “kCCP”, the model performance became worse than the final model (see Table C.4). ................................................................. 142

B.1 The Pearson correlation coefficient between our “ENTCP” and the features in TATA element class of “kCCP” sugbroup. They are sorted in ascending order. ................................................................. 154

B.2 The Pearson correlation coefficient between our “PONOCP” and the features in the initiation class of “kCCP” sugbroup. ................................................................. 155

B.3 The Pearson correlation coefficient between our “MONOCP” and the features in the TATA element, PIC, and scanning classes of “kCCP” sugbroup. ................................................................. 155

B.4 The Pearson correlation coefficient between our “PONOCP” and the features in “kCCP” sugbroup, which could not be assigned into any of TATA element, PIC, scanning, and initiation classes. ................................................................. 156
B.5 The Pearson correlation coefficient between our “MONOCP” and the features in “kCCP” subgroup, which could not be assigned into any of TATA element, PIC, scanning, and initiation classes.

B.6 Iterations of our algorithm on Eser et al. (2014) cell cycle data set (Step A, Part 1). Please see section 4.4 for details of our algorithm.

B.7 Iterations of our algorithm on Eser et al. (2014) cell cycle data set (Continued: Step A-B). At the end of Step A5, our algorithm transited to Step B.

B.8 Iterations of our algorithm on Eser et al. (2014) cell cycle data set (Continued: Step B).

B.9 Iterations of our algorithm on Eser et al. (2014) cell cycle data set (Continued: Step B-C).

B.10 Iterations of our algorithm on Eser et al. (2014) cell cycle data set (Continued: Step C).

C.1 Iterations of our algorithm on DREAM6 data set (Step A, Part 1). Please see Section 4.4 for details of our algorithm.

C.2 Iterations of our algorithm on DREAM6 data set (Continued: Step A-B). At A5, our algorithm transited to Step B.

C.3 Iterations of our algorithm on DREAM6 data set (Continued: Step B).

C.4 Iterations of our algorithm on DREAM6 data set (Continued: Step C).
List of Figures

1.1 The central dogma: a) in transcription, the genetic information flows from DNA to RNA; b) in translation, the gene information flows from RNA to protein. This applies to all living cells. (Adapted from Alberts et al. (2014)). .......................................................... 5

1.2 In a eukaryotic cell, there are various steps in the information flow from DNA to RNA to protein, as shown in (A). And the final level of each protein depends on the regulation at each step, as shown in (B). (Adapted from Alberts et al. (2014)). .......................................................... 7

1.3 A model of the general transcription factors assembling at core promoters to initiate transcription on template DNAs in purified in vitro systems. (A) The core promoter contains a DNA sequence called the TATA box, which is located 25–30 bp upstream of the TSS. (B) Through its subunit TBP, TFIID recognizes and binds the TATA box, and then enables the adjacent binding of TFIIB through causing a large bending in the DNA. (C) For simplicity the DNA bending is not shown. (D) Other general transcription factors, as well as the RNAPII itself, assemble at the core promoter. (E) TFIIH then uses energy from ATP hydrolysis to pry apart the DNA double helix at the TSS, locally exposing the template strand. TFIIH also phosphorylates RNAPII, changing its conformation so that the polymerase is released from the general transcription factors and can start the elongation phase of transcription. As shown, the site of phosphorylation is a long C-terminal polypeptide tail, also called the C-terminal domain (CTD), that extends from the RNAPII molecule. (Adapted from Alberts et al. (2014)). .......................................................... 12
1.4 (A) Consensus sequences found around eukaryotic RNAPII TSS. For the consensus sequences, N indicates any of A, C, G, and T, and two nucleotides separated by a slash indicate an equal probability of either nucleotide at the position. In reality, each consensus sequence is a shorthand representation of a histogram of A, C, G, and T. For most core promoters, only two or three of the four sequences are present. BRE: TFIIB recognition element; INR: initiator, which is a sequence overlapping with TSS and can be recognized by TAF1 and TAF2 subunits of the TFIID (reviewed in Sainsbury et al., 2015); DPE: downstream promoter element, which are probably recognized by TAF6–TAF9 subunits of the TFIID (reviewed in Sainsbury et al., 2015). (B) The TBP is the subunit of TFIID that is responsible for recognizing and binding to the TATA box in the DNA (red). The unique DNA bending caused by TB is thought to serve as a landmark that helps the other general transcription factors assemble onto DNA. TBP is a single polypeptide chain, being folded into two very similar domains (blue and green). (Adapted from Alberts et al. (2014). (B) is originally adapted from Kim et al. (1993).) . . . . . . . . . . . . . 14

1.5 Identified DNA regulatory elements in core promoters from yeast (top), and Drosophila and humans (bottom) are depicted, and elements known to contact TFIID are shown. (Adapted from Sainsbury et al. (2015). . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16

1.6 Distribution of the canonical TATA box consensus sequence TATAWAWR, where W = A or T, and R = A or G. (A) Histogram of TATA box location within the region [−250, +150] relative to the TSS (+1) in of 10,271 human promoters. (B) Histogram of TATA box location within the region [−250, +150] relative to the TSS (+1) in of 6,165 yeast S. cerevisiae promoters. (Adapted from Yang et al. (2007).) . . 18

1.7 The sequence logo of MBF (Mbp1p/Swi6p) in yeast S. cerevisiae . . 20

1.8 Eukaryotic TFs assemble into complexes on DNA. (A) Seven TFs are shown. They are not able to assemble in resolution. (B) The nature and function of the complex, in which a TF participates, depends on the specific binding site that seed the assembly of the complex, and the other participants, especially cofactors. Please note that the emphlight green and dark green TFs are shared by both activating and repressing complexes. (Adapted from Alberts et al. (2014).) . . 21
1.9 Eukaryotic TFs can act as activators at different steps of transcriptional regulation. (A) As an activator, a TF promotes binding of additional activators. (B) As an activator, a TF helps PIC formation and thereby RNAPII recruitment through direct interactions or indirect interactions. (C) As an activator, a TF helps to release already assembled RNAPII from the promoter. (D) As an activator, a TF releases RNAPII molecules that become stalled after transcribing about 50 nucleotides or RNA. (Adapted from Alberts et al. (2014).) 23

1.10 Six ways in which eukaryotic TFs function as repressors. (Adapted from Alberts et al. (2014).) 24

1.11 Eukaryotic transcription initiation by RNAPII in vivo. (Adapted from Alberts et al. (2014).) 27

2.1 Nucleosome competes with TF for DNA binding. (A) A profile using a model of only TFs and DNA. (B) Addition of nucleosomes yields a dramatically altered TF occupancy profile. (Adapted from Wasson and Hartemink (2009).) 37

2.2 Plots of the average nucleosome occupancy for each group centred on the Rap1p motif. X-axis is about absolute distance from Rap1p motif (bp). (A) Rap1p-bound loci are grouped into four categories based on their measured Rap1p residence time: longest, long, short, and shortest. (B) Rap1p-bound loci are grouped into four categories based on their measured Rap1p occupancy: highest, high, low, and lowest. (Adapted from Lickwar et al. (2012).) 38

2.3 The hypothesized mechanisms about TF binding to DNA that we do not model. (A) Protein-protein interaction between a TF and a transcription cofactor can result in modified DNA binding specificity. (B) Protein-protein interaction with a nearby binding TF might facilitate the DNA binding of a TF itself. (C) The base pairs flanking a TF binding site can influence the binding specificity of the TF, possibly by affecting the local DNA shape. (D) Some TFs can induce local DNA bending, which can increase the affinity of other TFs for sites nearby. (A), (C), and (B) are adapted from Levo and Segal (2014); (D) is adapted from Spitz and Furlong (2012).) 40
2.4 We do not model the ~10 bp periodic relationship between gene expression and TF binding instance location. In both panels, the mGal1-10 background and GAL1-10 background are the same sequence derived from the native yeast Gal1-10 promoter in which the known regulatory elements were mutated. (A) Each point corresponds to the average expression of eight sets of promoters in which Sharon et al. (2012) changed the location of the Gcn4p site, where the eight different sets differ in the location of a poly(dA:dT) tract of length 15 bp. To normalize the expression across the eight different sets, the expression is defined as a robust z score, calculated by subtracting the median and dividing by the s.d. of expression differences from the median. There is a ~10 bp periodicity of expression observed over five periods. (B) With settings same as (A), Levo et al. (2015) designed respective regulatory sequences with Gcn4p TF binding site of 9 bp (in blue) or 23 bp, including fixed-flanks site (in red) placed along the same context. In this panel, the y axis is z-score of TF binding, instead of gene expression. Indeed, the ~10 periodicity is about TF binding. However, it might be resulted from random flanks of binding sites: with fixed-flanks, the periodicity disappeared. (Adapted from Sharon et al. (2012) and Levo et al. (2015).)

2.5 Eser et al. (2014) reported time courses of 22 periodically expressed genes that are exclusively annotated as Mbp1p targets. These 22 genes agree well in the timing of gene expression, but show a remarkable diversity in their mean and magnitude of transcript production rate. (Adapted from Eser et al. (2014).)

2.6 Expression is a monotonic function of TF binding sites that saturates at 3–5 binding sites. (A) Within two different promoter backgrounds, Sharon et al. (2012) separately inserted Gcn4p sites in all $2^7(128)$ possible combinations of sites at seven predefined locations within the promoter. Shown are the individual promoter expressions for each background and mean expressions of all promoters that have $k$ Gcn4p sites for $k = 0, 1, 2, \ldots, 7$. Also shown is a fit of a logistic function for each background. (B) As in (A), but for all $2^5(32)$ possible combinations of Gal4 sites at five predefined promoter locations. The outlier promoter in terms of expression in which the two Gal4 sites closest to the core promoter were both added is indicated. These two sites were added at a distance of 1 bp, as opposed to the 5 bp distance used between all other adjacent sites, thus suggesting steric hindrance between Gal4p sites at this distance. (Adapted from Sharon et al. (2012).)
2.7 (A) The effect is about activators. Expression level is related to TF binding site proximity in the promoter. Mean log expression intensity of transcripts is plotted vs. the distance between their TSSs and the nearest putative TF binding sites in liver and 3T3-L1 cells. The analysis in this panel was restricted to genes with binding events in their promoters. (B) The Mata2p-Mcm1p complex is a repressor. Shown are four sets of promoters in which Sharon et al. (2012) modified the location of its repressor site along the promoter, where the four sets differ by the presence of poly(dA:dT) tracts and sites for the activators Gcn4p and Gal4p. Binding position is relative to the 5' end of a ~100 bp TATA-containing core promoter, which is fixed over all four sets of promoters. For each of the four sets, the expression of the promoter without the repressor site is indicated in the inset legend and is higher than all promoters that contain the repressor site. ((A) is adapted from MacIsaac et al. (2010); (B) is adapted from Sharon et al. (2012).) 

2.8 Correlations between the mean transcript production rate of non-periodic and periodic genes grouped into TATA box-containing and TATA-less with core promoter occupancies of general transcription factors involved in PIC formation, which are from Rhee and Pugh (2012). (Adapted from Eser et al. (2014).) 

2.9 Illustration of SAGA-dependent and TFIID-dependent transcriptional activation. Subsequent to TBP recruitment, PIC forms at core promoter, RNAPII is recruited to core promoter, and transcription is initiated. (A) The activator that binds to the UAS of the core promoter help to assemble PIC at the core promoter through interaction with one or more components of the transcriptional machinery, termed as “target”. (B) Such activator-target interactions could be achieved by firstly activator interacting with SAGA, and then SAGA regulate TBP recruitment via Mediator. (C) Activator can directly interact with both SAGA and Mediator, which ultimately both contribute to TBP recruitment. (D) Activator can also interact with TAF subunits of TFIID, and thereby control TBP recruitment. (Adapted from Bhaumik (2011).)
2.10 Shown is the consensus distribution of nucleosomes (grey ovals) around all yeast genes, aligned by the beginning and end of every gene. The arrow under the green circle near the 5' NFR represents the TSS. The green-blue shading in the plot represents the transitions observed in nucleosome composition and phasing (green represents high H2A.Z levels, acetylation, H3K4 methylation and phasing, while blue represents low levels of these modifications). The red circle represents transcriptional termination site within the 3' NFR. (Adapted from Jiang and Pugh (2009), which was reproduced from Mavrich et al. (2008).) .......................... 55

3.1 TSS location information with 346 genes subset of Eser et al. (2014). (A) The histogram of TSS's location relative to TrSS, given data from Rhee and Pugh (2012). TSS are widely distributed, and are condensed with in [-118, -1] of TrSS. The kernel density estimation has a mode as -56. (B) With [-118, -1] of TrSS as core promoter, 194 genes have TSSs located within the core promoter, and 77 genes have TSS out of the core promoter. Other 75 genes miss TSS information in Rhee and Pugh (2012). .......................... 61

3.2 In this work, we chose $d_u = 250$. In other words, we have 250bp as our decay unit, and consequently $\frac{1}{250}$ as our decay constant. With this value, we have the weight decays in an appropriate rate, and in a linear form within the region close to TrSS. But $d_u$ could be set as any reasonable integer by users of our software. The stability weight for TF binding instance is determined by average nucleosome nearby. In this work, we used the $[-30,30]$ window of motif center. .......................... 62

3.3 There are strong correlations between respectively associated features in “OPCP” and the “OPNT” subgroups. The correlation matrix plot for respective features within the “OPCP” and the “OPNT” subgroups. The scatter matrix plot for respective features within the “OPCP” and the “OPNT” subgroups. .......................... 68

3.4 We observed that our “ENTCP” feature has Pearson $r$ correlation coefficient in the range of 0.238 to 0.352 with those TATA element $k$-mer counts in “kCCP” subgroup. Specific values are given in Table B.1. .......................... 69

3.5 We observed that our “PONOCP” has Pearson $r$ correlation coefficient in the range of 0.0002 to 0.1001 with those initiation $k$-mer counts in “kCCP” subgroup. Specific values are given in Table B.2. .......................... 70
3.6 We observed that our “MONOCP” has Pearson $r$ correlation coefficient in the range of -0.272 to 0.075 with the $k$-mer counts in TATA element, PIC, and scanning classes of “kCCP” subgroup. 71

3.7 We observed that our “PONOCP” has Pearson $r$ correlation coefficient in the range of -0.133 to 0.160 with the $k$-mer having no class label in “kCCP” subgroup. 72

3.8 We observed that our “MONOCP” has Pearson $r$ correlation coefficient in the range of -0.162 to 0.112 with the $k$-mer having no class label in “kCCP” subgroup. 73

5.1 Taking the sequence specificities and concentrations of multiple kinds of DBFs, along with a DNA sequence, as inputs, COMPETE generates respective occupancy profiles for studied promoter sequences. Please note here we illustrated with the cell cycle setting, since there are DBF concentrations over time. We used Time=160 as an example, so the promoter occupancy profile is the one at Time=160. 97

5.2 The approach we exploited to estimate TFs real concentrations from its associated gene’s transcript level or transcript production rates. Inferring TF’s normalized concentrations from its associated gene’s deconvolved transcript level. A time lag as 20 time units is used here as an example. Estimating TF’s real concentrations from its normalized concentrations by multiplying a TF specific scaling factor $\alpha_{tf}$. Here $8K_d$ is used for illustration. Here TF MBF is used as an example. Please note MBF is a complex of Mbp1 and Swi6, and its binding component is MBP1. Therefore, we used MBP1 as the associated gene. Moreover, we used the Guo et al. (2013) data set as an example. 100
5.3 For Eser et al. (2014), we showed that almost all genes in the 346 genes subset have reasonably small autocorrelations with time lag 10 mins, as shown in (A), given the 95% confidence interval is usually \([-0.2, 0.2]\), and the 99% confidence interval is usually \([-0.3, 0.3]\).

Also, when we were inferring time lags for estimating TF concentrations, we followed Eser et al. (2014)’s suggestion to use piecewise linear function for estimating transcript production rates between given time points. With ERS1 gene as an example, we estimated the transcript production rates at all 70 mins, as shown in (B). With such estimated profiles, we calculated the autocorrelations with different time lags for all 346 genes. As shown in (C), 10 mins is generally a good time lag, with autocorrelation in the 95% confidence interval. The horizontal lines displayed in the plot correspond to 95% and 99% confidence intervals.

6.1 Scatter plots of our final model’s performance. Please note the performance metrics in (A) are calculated after all hold out sets are aggregated. But when we do model selection, we use the mean cross-validation metrics.

6.2 Without TF relevant features, the core promoter relevant features alone produced a model much worse than our final model (See Figure 6.1).

6.3 Set nucleosome concentration as zero for COMPETE to generate occupancy profile and extract features. With this setting, the predictive performance of the final model is even slightly better in terms of CV mean MSE: 0.067 is within one standard error of that of our final model \([0.072 - 0.005, 0.072 + 0.005]\), and the model is simpler without nucleosome concentrations. (See Figure 6.1 and Table B.10.)

6.4 Replacing “ENTF” in our final model by an alternative subgroup of features in “fluctuation” almost makes no change in terms of predictive performance, using “POTF” as an example.

6.5 The correlation matrix and scatter matrix for the respective features of TF Swi4p in the 8 subgroups of “fluctuation” group. Please note all 7 time points for each gene are used, and similar figures could be generated for any other TF. Please note, for all the plots about feature values, we have taken square roots of feature values to deal with skewness.

6.6 The correlation matrix plot for Swi4p’s respective features in the defined 8 subgroups.
6.7 With “fluctuation” (“ENTF”) as the only (sub)group of features, the fitted linear model got poor results. 

6.8 Replacing “kCCP” in our final model with the joint set of “OPCP” and “OPNT”. 

6.9 By adding “OPCP” and “OPNT” into “amplifier” group in our final model, which is previously represented by “kCCP” only, the model does not improve our final model: the CV mean MSE 0.070 is within one standard error of that of our final model [0.072 – 0.005, 0.072 + 0.005], where we favor a simpler model. (See Figure 6.1 and Table B.10.)

6.10 Our final model emphasized synergy between TFs and core promoter elements through selecting only interactions of the two associated main effects. To validate such a selection, we tried an additive function of the two main effects. Such an additive function is much worse than production function that represents synergistic interaction (See Figure 6.1).

6.11 To validate the selected synergistic interactions between TFs and core promoter elements, we tried adding the two main effects into our final model. This did not improve model performance in terms of MSE: the CV mean MSE 0.069 is within one standard error of that of our final model [0.072 – 0.005, 0.072 + 0.005], where we favor a simpler model. (See Figure 6.1 and Table B.10.)

6.12 Probably due to the variable lengths of gene body, which is demonstrated in (A), and the exponential weight function with 250 as decay constant, the expected number of distance decay weighted gene body nucleosomes is just slightly correlated with the expected number of gene body nucleosomes, as shown in (B).

6.13 Adding “ENNGBD” into our final model by extending interactions. It did not improve our model’s performance (See Figure 6.1).

6.14 Adding “ENNGBD” into our final model as main effects. It did not improve our model’s performance (See Figure 6.1).
7.1 An example of the master strain into which respective research groups integrated all tested promoters (adapted from Zeevi et al. (2011)). Please note that, though it might be different among relevant data sets, there are always a gene encoding for a red fluorescent protein (mCherry) followed by a control promoter, and a gene encoding for a yellow fluorescent protein that is the reported gene for tested promoters.

8.1 The respective plots of FiRST and c4lab in DREAM6 (adapted from Conference Slides (Meyer et al., 2013)). The comparison plot of FiRST team’s results. Promoters are sorted by absolute error. The $R$ in the figure is Pearson correlation coefficient $C_p$. The comparison plot of c4lab team’s results. Promoters are sorted by absolute error. With whole test data set, the c4lab team got $r = 0.5386$.

8.2 Scatter plots of PreTrans’ results on DREAM6 data set.

8.3 The comparison plot of our PreTrans results. Promoters are sorted by absolute error.

8.4 The respective histograms of “ENNGBD” feature value in training set and test set.

8.5 A mutation to nucleosome disfavoring sequence at core promoter region (~83 relative to TrSS) would influence not only nucleosome occupancy at core promoter region, but also that at gene body.

A.1 “kmerConservation” subgroup of k-mer features Zeevi et al. (2014) (adapted from Zeevi et al. (2014)). The authors fitted 10 linear models 10 different partitions of the data into training and held-out test sets. A feature is included in the list shown in Part (A), only if it was selected in 8 out of the 10 fitted linear models. This list of features include k-mer existence and counts, features of hits of PSSMs of known RP regulators, and a feature of the predicted intrinsic nucleosome occupancy. Then based on the features’ locations on the promoter and their effect on expression, the authors grouped 25 out of 38 robust features, and provided a schematic representation of the groups of features, as shown in Part (B). Please note we excluded features 8, 14, 15, 22, 29, 30, 31, 32, 33, 37, and 38 in our work. These eleven features are about TF binding, TBP binding, and nucleosome occupancy. We assume that COMPETE is better on this.
A.2 “kmerCorePromoter” subgroup of \(k\)-mer features reported in Lubliner et al. (2015) (adapted from Lubliner et al. (2015)). The authors fitted 10 linear models to different partitions of the data to training and held-out test sets. A feature is included in the list shown in Part (A) only if it was selected in 6 out of the 10 fitted linear models. The authors then manually classified them according to the associated window’s location in core promoter, and its base content, as shown in Part (B).

C.1 Detailed view of the types and locations of mutant RP promoters in DREAM6 (adapted from Zeevi et al. (2011)). The same figure was used in Meyer et al. (2013) as well. Please note DREAM6 has 4 promoters missing information in this figure: RPL3_Mut5, RPL4A_Mut4, RPL4A_Mut5, and RPL4A_Mut6, all of which are about random mutations, according to the Supplementary Table of Meyer et al. (2013).

C.2 Detailed List of the types, locations, and mutated sequences of mutant RP promoters in DREAM6 (adapted from Zeevi et al. (2011)). Please note DREAM6 still has those 4 promoters missing information in this list. Also, the location is relative to translation start site (TrSS). And the NDS means nucleosome disfavoring sequence. A mutation type means that it is a mutation on that type of sequence. For example, a NDS mutation, is a mutation to a sequence which is nucleosome disfavoring before mutation.

C.3 Types of mutations and their effects on RP gene expression (adapted from Zeevi et al. (2011)). For each promoter, shown is activity of the natural promoter and a promoter in which a mutation was executed. Error bars represent two standard errors computed from 24 replicates. The magnitude of the effect of the mutation on promoter activity is indicated above the activity bars of each promoter pair, where two stars mark promoter activity differences that are statistically significant. Here promoter activity is the mean transcript production rates over population exponential growth phase. A) random mutation (18 bp sequence changes). B) Mutations of Fhl1p sites (2bp changes that preserved the G/C content). C) Mutations of Sfp1p sites (2bp changes that preserved the G/C content). D) Mutations of TATA box (2-3bp changes that preserved the G/C content). E) Mutations of A/T-rich nucleosome disfavoring sequences (replacing 16 A/T base pairs with G/C base pairs in a region of 31 base pairs within the promoter that had the lowest predicted nucleosome occupancy (Kaplan et al., 2009)). F) Summary and comparison of the effect of the mutations according to the mutation type.
C.4 Only names of the two best-performing teams are given. Our team was the 3rd one. \(C_p\) (see Equation 7.1) indicates the Pearson correlation coefficient, \(X^2\) is a score based on the \(\chi^2\) metric (see Equation 7.2), \(S_p\) is the Spearman’s rank correlation coefficient (see Equation 7.3), and \(R^2\) the score based on the rank\(^2\) metric (see Equation 7.4). For \(C_p\) and \(S_p\), the higher the score is, the better the team’s predictions are. For \(X^2\) and \(R^2\), the lower (This Figure is adapted from Meyer et al. (2013)).
List of Abbreviations and Symbols

Abbreviations

- bp: base pair
- DBF: DNA binding factor
- DNA: deoxyribonucleic acid
- GTF: general transcription factor
- mRNA: messenger RNA
- NFR: nucleosome free region
- PSSM: position specific scoring matrix
- RNA: ribonucleic acid
- RNAPII: RNA polymerase II
- RP: ribosomal protein
- PIC: pre-initiation complex
- SAGA: Spt-Ada-Gcn5 acetyltransferase
- TAF: TBP-associated factor
- TBP: TATA-binding protein
- TF: transcription factor
- TSS: transcription start site
- TrSS: translation start site
- UAS: upstream activating sequence
Feature name codes

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENNGB</td>
<td>expected number of nucleosomes within gene body</td>
</tr>
<tr>
<td>ENNGBD</td>
<td>expected number of distance decay weighted nucleosomes within gene body</td>
</tr>
<tr>
<td>ENTCP</td>
<td>expected number of TBP instance within core promoter region</td>
</tr>
<tr>
<td>ENTF</td>
<td>expected number of TF binding instance</td>
</tr>
<tr>
<td>ENTFD</td>
<td>expected number of distance decay weighted TF binding instance</td>
</tr>
<tr>
<td>ENTFS</td>
<td>expected number of stable TF binding instance</td>
</tr>
<tr>
<td>ENTFDS</td>
<td>expected number of distance decay weighted and stable TF binding instance</td>
</tr>
<tr>
<td>ENTNT</td>
<td>expected number of TBP instance nearby TSS</td>
</tr>
<tr>
<td>kCCR</td>
<td>$k$-mer count based sequence features in conservation study of RP genes</td>
</tr>
<tr>
<td>kCCP</td>
<td>$k$-mer count based sequence features in core promoter region</td>
</tr>
<tr>
<td>MONOCP</td>
<td>$-1$ nucleosome occupancy within core promoter region</td>
</tr>
<tr>
<td>MONONT</td>
<td>$-1$ nucleosome occupancy nearby TSS</td>
</tr>
<tr>
<td>OPCP</td>
<td>occupancy profile based features in core promoter region</td>
</tr>
<tr>
<td>OPNT</td>
<td>occupancy profile based features nearby TSS</td>
</tr>
<tr>
<td>PONOCP</td>
<td>$+1$ nucleosome occupancy within core promoter region</td>
</tr>
<tr>
<td>PONONT</td>
<td>$+1$ nucleosome occupancy nearby TSS</td>
</tr>
<tr>
<td>POTF</td>
<td>probability of at least one TF binding instance</td>
</tr>
<tr>
<td>POTFD</td>
<td>distance decay weighted probability of at least one TF binding instance</td>
</tr>
<tr>
<td>POTFS</td>
<td>probability of at least one stable TF binding instance</td>
</tr>
<tr>
<td>POTFDS</td>
<td>distance decay weighted probability of at least one stable TF binding instance</td>
</tr>
</tbody>
</table>
Acknowledgements

First and foremost, I would like to thank my advisor Alex Hartemink, not only for great advices, but also for always being supportive and encouraging. When I was down, Alex told me that everybody has ups and downs when doing research, and being down means that probably ups are in the near future. During these years, Alex would always show his respect to my thoughts. Instead of asking me doing something, Alex would always help me find topics worth working on. I deeply appreciate that he has always been patient, thoughtful, supportive, and encouraging. From certain perspectives, I think Alex is a moral leader, and is my moral model.

I would also like to thank my committee members Raluca Gordan, Ryan Baugh, and Sayan Mukherjee, for always being willing to help me, and always being flexible to reschedule my milestone events. Raluca reminds me exploiting appropriate computational techniques to tackle the biological problems. Ryan keeps me interested in and focused on the important biological questions. Sayan introduces me the relevant statistical methods and machine learning algorithms.

Moreover, I would like to thank people in our group, my teachers, my fellow graduates, staff members in computer science department, and other people I have come to know at Duke. Throughout my graduate school career, this community has helped me grow as a person and a scientist.

Last but not least, I would like to thank my family for always being supportive.
Genetic information is encoded within the sequence of DNA. The process of transcription copies the information encoded in the DNA to a new molecule of RNA called a transcript. If the transcript is a messenger RNA (mRNA), it can be translated by a ribosome into a sequence of amino acids to produce a protein. At the same time and under the same condition, distinct genes can be expressed at different levels. Also, at different times and under different conditions, the same gene can be expressed at different levels. These are all governed by the processes of transcriptional and translational regulation. The mechanisms by which these processes manage to govern the level of gene expression over time in response to different conditions remains a huge open problem in science. Here, we mainly focus on the first step, regulation of the process of transcription, the direct consequence of which will be control over the rate of transcript production (ultimate transcript levels will also be modulated by regulated transcript degradation). We aim at producing quantitative and predictive models of how this first step of gene expression control is enacted.

Recently, promising early models have been proposed for relating a promoter’s DNA sequence to the downstream gene’s transcript level or production rate (Bintu
et al., 2005; Segal et al., 2008; Raveh-Sadka et al., 2009; He et al., 2010; MacIsaac et al., 2010; Zeevi et al., 2011; Sharon et al., 2012; Raveh-Sadka et al., 2012; Meyer et al., 2013; Rajkumar et al., 2013; Zeevi et al., 2014; Lubliner et al., 2015). But these have a number of limitations. All efforts aim at predicting the downstream gene’s mean transcript level or mean transcript production rate over time. However, transcript production is dynamic (Pelechano et al., 2010). Most models only employ features based on \( k \)-mer (a \( k \)-tuple or \( k \)-gram of DNA sequence), but these do not lend themselves to easy biological interpretation. Models such as these, even if they were to add richer sets of sequence-encoded features, like lengths of T- or AT-tracts or levels of DNA deformability, or they were to classify features into categories, like TATA elements or pre-initiation complex (PIC, sometimes called basal transcription machinery), will not be able to predict dynamic transcript production rates in response to changes in TF availabilities over time (because the sequence is constant over time) (Zeevi et al., 2011; Sharon et al., 2012; Raveh-Sadka et al., 2012; Meyer et al., 2013; Zeevi et al., 2014; Lubliner et al., 2015). On the other hand, models that attempt to incorporate more mechanistic features, such as number of (weighted) TF binding instances, have thus far failed to achieve sufficient accuracy to be highly predictive, or sufficient generalizability to be highly predictive in different settings, since in these works possible regulatory elements other than TF binding, like core promoters, are fixed or assumed constant (Segal et al., 2008; Raveh-Sadka et al., 2009; He et al., 2010; MacIsaac et al., 2010; Rajkumar et al., 2013). Complicating the matter, but providing a possible new angle of attack, TFs are not the only factors binding to promoters; other DNA binding factors (DBFs) bind as well, especially nucleosomes (Wasson and Hartemink, 2009; Raveh-Sadka et al., 2009). The DNA sequence preference of nucleosome and competition between nucleosomes and TFs result in higher-order structures within factor binding configurations, like nucleosome free regions (NFRs), which are needed at core promoters to recruit PIC
components (Field et al., 2008; Kaplan et al., 2009; Bai et al., 2010; Bai and Morozov, 2010; Venters et al., 2011; Raveh-Sadka et al., 2012; Rhee and Pugh, 2012). And nucleosomes probably influence the stability of TF binding instance (Lickwar et al., 2012). Also, nucleosomes limit TSS selection by impeding RNA Polymerase II (RNAPII) scanning at TATA-less promoters (Rhee and Pugh, 2012). Lastly, nucleosomes regulate transcription by influencing RNAPII elongation activity (Weber et al., 2014; Jordan-Pla et al., 2015).

With all these in mind, we developed a statistical learning framework that includes not only core promoter \(k\)-mer based features, but also biologically mechanistic features extracted from the posterior probabilities of DBFs starting at per position in promoter and gene body, which as a total is called as occupancy profile. These features and their respective interaction terms are then used to estimate models for predicting transcript production rates. With such models, we not only achieve good performance in terms of predicting genes’ dynamic transcript production rates, but also are able to provide biological insights by interpreting the statistically significant features used in the model. Please note, instead of using the posterior probabilities of DBFs starting at per position, usually we use the sums of probabilities of respective DBFs’ all possible states as occupancy profile to do visualization.

The model framework we developed has been trained and evaluated in the \(S.\) cerevisiae cell cycle dynamic transcript production rate dataset from Eser et al. (2014). In this scenario, the genomic promoter sequences are fixed, and we are using data regarding dynamic transcript production rates over time in response to changing TF concentrations throughout the cell cycle in yeast. Also, we have applied the model framework to specific cases, where we used data measuring changes in transcript production rates as a function of variation in genomic promoter sequences, this time with TF concentrations being held fixed as its time average. Specifically, one such dataset is from DREAM6 Gene Promoter Expression Prediction challenge (Meyer
et al., 2013). It measures expression in response to mutated promoter sequences of yeast ribosomal protein genes.

As far as we know, our framework is the first successful attempt to have a model of dynamic transcript production and transcriptional regulation: with cell cycle data set, we got Pearson correlation coefficient \( C_p = 0.751 \) and coefficient of determination \( r^2 = 0.564 \) on test set for predicting dynamic transcript production rate over time; with DREAM6 Gene Promoter Expression Prediction challenge, we outperformed all participant teams, best of all teams, and a combined model of best team and biological mechanistic features, in terms of almost all scoring metrics. Moreover, our model shows its capability of identifying generalizable features by interpreting the highly predictive models, and thereby provide support for associated hypothesized mechanisms. Specifically, the features selected by our models provide support for the following biological insights: a) TFs govern the probability of RNAPII recruitment and initiation possibly through interactions with PIC components and transcription cofactors; b) the core promoter amplifies the transcript production probably by influencing PIC formation, RNAPII recruitment, DNA melting, RNAPII searching for and selecting TSS, releasing RNAPII from general transcription factors and thereby initiation; c) there is strong transcriptional synergy between TFs and regulatory elements within the core promoter, which very likely represent respective DNA sequence signals for recruiting general transcription factors and transcription cofactors, and for TSS scanning and selection; d) the regulatory elements within core promoter region are more than TATA box and nucleosome free region, suggesting the existence of still unidentified TAF-dependent core promoter elements in yeast \( S. \ cervisiae \); e) nucleosome occupancy profile is helpful for representing \(-1\) and \(+1\) nucleosomes’ regulatory roles on transcription.

Before describe the machine learning framework, we first introduce the biological background that motivates our work toward understanding transcriptional regulation
Figure 1.1: The central dogma: a) in transcription, the genetic information flows from DNA to RNA; b) in translation, the gene information flows from RNA to protein. This applies to all living cells. (Adapted from Alberts et al. (2014)).

by providing predictive models of transcript production rates. At first, we discuss the gene expression and transcriptional regulation in eukaryotes, pointing out the specificities of those in yeast *S. cerevisiae*, which serves as a minimal model eukaryote. We emphasize the regulation of transcription initiation. We discuss the regulatory elements in core promoter and their respective prevalence in different eukaryotes, as well as the different architectures of these elements in distinct organisms. We explore the complexity of transcription initiation *in vivo*: the various ways TFs could regulate transcription initiation, being activators or repressors through different mechanisms, especially recruiting transcription cofactors and thereby ultimately interacting with the general transcription factors or even RNAPII. We also briefly discuss elongation and termination, the mechanisms of which mostly remain unclear for us. Finally we focus on transcriptional regulation in yeast *S. cerevisiae*, discussing its advantages for modeling transcriptional regulation as the minimal model eukaryote. We end this first chapter with a detailed outline of this dissertation.
1.1 Gene expression and the regulation of gene expression in eukaryotes

According to the “central dogma” of molecular biology (see Figure 1.1), a cell express the genetic information encoded within the sequence of DNA in two steps: a) in the process called transcription, the information encoded in a gene is transcribed by a RNA polymerase into a new molecule of RNA called a transcript; b) if the transcript is a mRNA, in the process called translation, it is translated by a ribosome into a protein. Though the “central dogma” is universal, there are important variations between organisms about the information flow from DNA to protein. For example, RNA transcripts in eukaryotic cells are processed in a series of steps in the nucleus, including RNA splicing, before they exit from the nucleus and be translated into proteins.

In this work, we loosely use the term gene to refer to a segment of DNA sequence that is transcribed into a RNA molecule. For majority of genes carried in a cell’s DNA, their final products are proteins. The RNA transcripts of such genes are called messenger RNA (mRNA) molecules. There are also many genes having RNA as the final product. In the yeast *S. cerevisiae*, over 1200 genes (more than 15% of the total) produce RNA as their final product (Alberts et al., 2014). These RNAs are known as noncoding RNAs because they do not code for protein. Some of these noncoding RNAs have structural and catalytic roles in the cells. Other noncoding RNAs act primarily as regulators of gene expression. But we have limited knowledge about the roles of many noncoding RNAs. Though noncoding RNAs are important, in this work, we focus on genes producing mRNAs.

Gene expression can be regulated at various steps in the information flow from DNA to RNA to protein, as shown in Figure 1.2. With transcriptional regulation (Step 1 in Figure 1.2B), when, where, and how often a given gene being transcribed
Figure 1.2: In a eukaryotic cell, there are various steps in the information flow from DNA to RNA to protein, as shown in (A). And the final level of each protein depends on the regulation at each step, as shown in (B). (Adapted from Alberts et al. (2014)).
from DNA to RNA could be controlled. With post-transcriptional regulation (Steps 2, 3, 5 in Figure 1.2B), the splicing and processing of RNA transcripts, RNA transport from the nucleus to the cytosol and localization in cytosol, as well as the mRNA degradation in the cytoplasm, could be controlled. With translational regulation (Step 4 in Figure 1.2B), which mRNAs in the cytoplasm being translated by ribosomes could be controlled. With post-translational regulation (Step 6 in Figure 1.2B), activating, inactivating, degrading, or localizing specific protein molecules could be controlled. Please note, different regulation means could be coupled to each other because of the coupling of corresponding steps of gene expression. For example, transcription elongation in eukaryotes is tightly coupled to RNA processing.

In this work, we focus on transcriptional regulation. For most genes, transcriptional regulations are supreme. As illustrated in Figure 1.2B, only transcriptional regulation guarantees that the cell will not synthesize superfluous intermediates for gene expression. In the following sections, we discuss the typical transcriptional regulation in eukaryotes. Specifically, just like gene expression could be controlled at multiple steps, transcription could be regulated at three phases: initiation, elongation, and termination. We discuss the transcriptional regulation at each of the three phases, emphasizing the regulation of transcription initiation.

1.2 Regulation of transcription initiation in eukaryotes: introduction

The initiation of transcription is an extremely important step in gene expression: it is the main step at which the cell regulates which proteins are to be produced and at what rate. Transcription is performed by the enzymes called RNA polymerases. Eukaryotic nuclei have three types of RNA polymerases: RNA polymerase I, RNA polymerase II, and RNA polymerase III. Thee three polymerases have similar structures and share some common subunits, but they transcribe different categories of genes. RNA polymerases I and III transcribe the genes encoding transfer RNA, ri-
bosomal RNA, and various small RNAs. RNA polymerase II transcribes most genes, including all those that encode proteins. We focus on RNA polymerase II (RNAPII), since it transcribes genes producing mRNAs.

For protein-coding genes, the process of transcription initiation requires that the RNAPII be recruited to DNA sequence and slide into a special sequence of nucleotides indicating the transcription start site (TSS). The recruitment of RNAPII could be regulated by gene regulatory proteins assembling at or binding to various DNA regulatory regions. A regulatory region in DNA is a collective term of cis-regulatory sequences (binding sites or motifs of gene regulatory proteins), which are specific sequences of DNA (typically 5–10 base pairs in length) on the same chromosome (that is, in cis) to the genes they regulate and being bound by gene regulatory proteins. In eukaryotes, the DNA regulatory regions include promoters and enhancers (or silencers).

A **promoter** is a DNA region that allows accurate transcription initiation of a particular gene. The promoter is located upstream of the regulated gene (towards the 5′ region of the sense strand). A **core promoter** is the minimal portion of a promoter required for proper transcription initiation. An **enhancer** is a short DNA regulatory sequence that can be bound by regulatory proteins (activators) to enhance the rate of transcription initiation of a gene. Enhancers are usually far away from the regulated gene, and could be either upstream or downstream of the regulated gene. Silencers are similar to enhancers in a lot of ways, except that they are bound by repressors to decrease the rate of transcription initiation of a gene. We will mainly discuss enhancers. But mechanisms relevant to silencers could be understood in a similar way. Please note, some literatures might have different definitions of the DNA regulatory regions. For example, Alberts et al. (2014) only has definition for promoter, which is identical to core promoter in our work. In Alberts et al. (2014), all regions other than their promoter are enhancers, including the upstream
activating sequences (UASs), which are *cis*-regulatory sequences located upstream of and within a few hundred base pairs (bp) of their promoter. In this work, the UASs and DNA sequences between them collectively form the part of the promoter immediately upstream of the core promoter.

1.3 Regulation of transcription initiation in eukaryotes: core promoter and the assembly model of general transcription factors *in vitro*

Among the regulatory regions, the core promoter probably is the most important one. The core promoter surrounds the TSS that directly interacts with the PIC components. The PIC consists of RNAPII and a set of *general transcription factors*. Its main component is RNAPII, which requires the set of general transcription factors to be placed on TSS. The proteins are “general” because they are required at nearly all RNAPII promoters. They help to position eukaryotic RNAPII correctly on TSS, to pull apart the two strands of DNA to allow transcription to begin, and to release RNAPII from the promoter to start elongation. The general transcription factors are denoted arbitrarily as TFIIA, TFIIB, TFIIC, TFIID, and so on, where TFII stands for “transcription factor for polymerase II”.

It is established that the general transcription factors are sufficient for transcription initiation *in vitro*. As shown in Figure 1.3, the general transcription factors assemble at core promoter to initiate transcription. The assembly process starts with TATA-binding protein (TBP) binding to the TATA box. TBP is a subunit of TFIID. The TATA box is typically located 25 to 30 bp upstream of the TSS (reviewed in Smale and Kadonaga, 2003). TATA box is not the only DNA consensus sequence that signals the start of transcription. As shown in Figure 1.4A, there are some other elements in the core promoter could be recognized by associated general transcription factors. Through its subunit TBP, TFIID firstly recognizes and binds
the TATA box. The binding of TFIID causes a large bending in the DNA of the TATA box, as shown in Figure 1.4B. This bending is thought to serve as a landmark for the location of an active core promoter in the midst of a very large genome, and it brings DNA sequences on both sides of the bending closer together to allow for subsequent assembly steps. Other general transcription factors then assemble, along with RNAPII, to form a complete PIC.

The general transcription factors have varied functional and structural roles (reviewed in Alberts et al., 2014; Allen and Taatjes, 2015; Sainsbury et al., 2015): consisting of TBP and 13 different TBP-associated factors (TAFs), 5 of which dimerize (TAF4, TAF5, TAF6, TAF9 and TAF12), TFIID recognizes TATA box through TBP subunit, and other DNA sequences near the TSS through TAF subunits; having two subunits, the auxiliary factor TFIIA is not required for transcription initiation but can stabilize the TBP–DNA complex; containing only one subunit, TFIIB recognizes the TFIIB recognition elements (BREs) upstream and downstream of TATA box, and thereby accurately positions the RNAPII at the TSS; composed of three subunits, TFIIF stabilizes RNAPII interaction with TBP and TFIIIB, and also helps attract TFIIE and TFIIH; containing two subunits, TFIIE binds to RNAPII, and facilitates the recruitment of TFIIH to the PIC, thus providing a bridge between RNAPII and TFIIH; consisting of nine subunits, TFIIH unwinds DNA at the TSS to expose the template strand, phosphorylates the serine located at the fifth position in the repeat sequence (Ser5) of the RNAPII CTD, and releases RNAPII from the promoter by disengaging it from the cluster of general transcription factors. The most complicated one of the general transcription factors is TFIIH. It is nearly as large as RNAPII itself and performs several enzymatic steps needed for the initiation of transcription.

The general transcription factors are highly conserved; some of those from human cells can be replaced in biochemical experiments by their counterparts from simple
Figure 1.3: A model of the general transcription factors assembling at core promoters to initiate transcription on template DNAs in purified *in vitro* systems. (A) The core promoter contains a DNA sequence called the TATA box, which is located 25–30 bp upstream of the TSS. (B) Through its subunit TBP, TFIID recognizes and binds the TATA box, and then enables the adjacent binding of TFIIB through causing a large bending in the DNA. (C) For simplicity the DNA bending is not shown. (D) Other general transcription factors, as well as the RNAPII itself, assemble at the core promoter. (E) TFIIH then uses energy from ATP hydrolysis to pry apart the DNA double helix at the TSS, locally exposing the template strand. TFIIH also phosphorylates RNAPII, changing its conformation so that the polymerase is released from the general transcription factors and can start the elongation phase of transcription. As shown, the site of phosphorylation is a long C-terminal polypeptide tail, also called the C-terminal domain (CTD), that extends from the RNAPII molecule. (Adapted from Alberts et al. (2014)).
yeasts. However, the assembly model of the general transcription factors, which is demonstrated in Figure 1.3, was deduced from experiments performed in vitro, and the exact order in which the general transcription factors assemble on core promoters probably varies from gene to gene in vivo. Moreover, the consensus sequences found around TSS, which are shown in Figure 1.4, do not always appear in a eukaryotic RNAPII promoter. Therefore, we discuss the prevalence of the consensus sequences and the assembly model of the general transcription factors given in Figures 1.3 and 1.4 as follows.

1.3.1 Only part of the regulatory elements are present in each core promoter

For most RNAPII core promoters in all eukaryotes, only two or three of the four regulatory elements in Figure 1.4A are present.

Though TATA box might be strictly conserved and essential for transcription initiation from yeast to human, it is probably true for all eukaryotes that not all core promoters contain TATA box. As reviewed in Smale and Kadonaga (2003), for Drosophila, TATA boxes were present in 43% of 205 core promoters (Kutach and Kadonaga, 2000), and in 33% of 1,941 potential promoters (Ohler et al., 2002); for human, TATA boxes were present in 32% of 1,031 potential core promoters (Suzuki et al., 2001). Recently, Rhee and Pugh (2012) reported that, for yeast S. cerevisiae, only 17% of 3,945 mRNA genes contain TATA box in their respective core promoters.

Similarly, the initiator (INR), which is found in yeast and metazoans, is not always present in core promoters. As reviewed in Smale and Kadonaga (2003), the Drosophila consensus INR sequence, or a sequence containing one mismatch, was present in 69% of 204 core promoters or, in another study, in 69% of 1,941 core promoters (Kutach and Kadonaga, 2000; Ohler et al., 2002). In a study about human, it is reported that 49% of the core promoters contain the INR element (Gershenzon and Ioshikhes, 2005). In yeast S. cerevisiae, Yang et al. (2007) reported that ~40%
Figure 1.4: (A) Consensus sequences found around eukaryotic RNAPII TSS. For the consensus sequences, N indicates any of A, C, G, and T, and two nucleotides separated by a slash indicate an equal probability of either nucleotide at the position. In reality, each consensus sequence is a shorthand representation of a histogram of A, C, G, and T. For most core promoters, only two or three of the four sequences are present. BRE: TFIIB recognition element; INR: initiator, which is a sequence overlapping with TSS and can be recognized by TAF1 and TAF2 subunits of the TFIID (reviewed in Sainsbury et al., 2015); DPE: downstream promoter element, which are probably recognized by TAF6-TAF9 subunits of the TFIID (reviewed in Sainsbury et al., 2015). (B) The TBP is the subunit of TFIID that is responsible for recognizing and binding to the TATA box in the DNA (red). The unique DNA bending caused by TB is thought to serve as a landmark that helps the other general transcription factors assemble onto DNA. TBP is a single polypeptide chain, being folded into two very similar domains (blue and green). (Adapted from Alberts et al. (2014). (B) is originally adapted from Kim et al. (1993).)
of yeast core promoters have mammalian-type INR sequences within the \([-25, +25]\) region relative to TSS, which is similar to the fraction of human core promoters containing INR elements. However, it is believed that the yeast INR elements probably function differently from properties metazoan INR elements (reviewed in Smale and Kadonaga, 2003; Yang et al., 2007): mutation of yeast INR sequence results in repositioning of the TSS (Chen and Struhl, 1985; Hahn et al., 1985; McNeil and Smith, 1985; Nagawa and Fink, 1985; Mösch et al., 1992; Hampsey, 1998), but promoter activity is often unaffected or reduced only modestly (Mösch et al., 1992), while in metazoans an INR often can greatly enhance promoter activity. Thus, rather than being a major contributor to core promoter recognition through being recognized by TAF1 and TAF2 as the metazoan INR elements, the yeast INR element may represent a preferred TSS for the RNAPII after it scans downstream from its initial interaction site, which is 20 bp downstream of the TATA box (Giardina, 1993).

Also, as reviewed in Yang et al. (2007), in human, it has been estimated that only \(~12\%\) of core promoters have a DPE, while \(~22\%\) have a BRE (Jin et al., 2006); as reviewed in Butler and Kadonaga (2002), it has been estimated that \(~40\%\) of core promoters have a DPE (Kutach and Kadonaga, 2000). Moreover, though BRE and DPE are identified in metazoans (e.g. human and Drosophila), they are not found in yeast core promoters (reviewed in Yang et al., 2007; Sainsbury et al., 2015). Therefore, as depicted in the top panel of Figure 1.5, it is believed yeast core promoters contain at most TATA box and INR.

However, we note that yeast TAFs are important for transcriptional regulation, suggesting the existence of still unidentified TAF-dependent core promoter elements in yeast (reviewed in Green, 2000). This is supported by recent studies about core promoter sequence features in yeast (Lubliner et al., 2013; Zeevi et al., 2014; Lubliner et al., 2015). And, Lubliner et al. (2013) indicates that probably there are unidentified core promoter elements in human as well.
1.3.2 Other core promoter regulatory elements are reported

Some regulatory elements within the core promoter are not given in Figure 1.4A. As reviewed in Juven-Gershon et al. (2008) and Sainsbury et al. (2015), in addition to the original BRE that is identified as TFIIB-binding sequence and is immediately upstream of TATA box, it was found that TFIIB can bind upstream or downstream of the TATA box at the BRE$_u$ (upstream BRE, which is the same as the original BRE) or the BRE$_d$ (downstream BRE) sequences (Deng and Roberts, 2005, 2006, 2007). As reviewed in Sainsbury et al. (2015), the motif ten element (MTE) was identified in *Drosophila melanogaster* and human core promoters (Lim et al., 2004), while the downstream core element (DCE) was identified in human promoters only (Lewis et al., 2000). Just like DPE, MTE is probably recognized by TAF6–TAF9 (Burke and Kadonaga, 1997; Theisen et al., 2010). Moreover, the DCE is possibly contacted by TAF1 (Lee et al., 2005). These regulatory elements are depicted in the bottom panel of Figure 1.5.

1.3.3 Core promoter architectures are different among distinct eukaryotes

The architecture of regulatory elements within core promoter might vary significantly among different eukaryotes. As shown in Figures 1.4A and 1.6A, it is believed that TATA box is almost always located 25–30 bp upstream of the TSS. However,
this is not true for yeasts (reviewed in Struhl, 1989; Smale and Kadonaga, 2003; Yang et al., 2007). As shown in Figure 1.6B, in yeast *S. cerevisiae*, TATA boxes are typically located 40–120 bp upstream of the TSSs, and could be located 250 bp upstream of the TSSs. But it is known that, in both yeast and metazoans, core promoter DNA melting occurs ~20 bp downstream of the TATA box (Giardina, 1993), and thereby the core promoter sequence 30 bp downstream of the TATA box is at the RNAPII active center (Bushnell, 2004; Miller and Hahn, 2006).

Giardina (1993) proposed that following PIC formation, the yeast RNAPII performs a downstream scan of the melted template strand, searching for TSS sequence signals. This scanning model could explain the difference of TSS location between yeast and metazoans, and is consistent with relevant studies about INR element and TSS sequence signal in yeasts (Chen and Struhl, 1985; Hahn et al., 1985; McNeil and Smith, 1985; Nagawa and Fink, 1985; Mösch et al., 1992; Hampsey, 1998; Kuehner and Brow, 2006; Sugihara et al., 2011). Some other studies also showed that TFIIB, TFIIF and TFIIH affect TSS selection in a way depending on the sequences upstream and downstream of TSS (Faitar et al., 2001; Khaperskyy et al., 2008; Fishburn and Hahn, 2012).

1.4 Regulation of transcription initiation: complex transcription initiation *in vivo*

Transcription initiation *in vivo* is more complex and requires more proteins than it does *in vitro*, mainly because DNA in eukaryotic cells is packaged into nucleosomes, and nucleosomes are further arranged in higher-order chromatin structures. In a eukaryotic cell, with a core promoter that is packaged in nucleosomes, the general transcription factors and RNAPII are not capable to assemble onto the core promoter on their own. Therefore, RNAPII additionally requires specific transcription factors (hereinafter referred to as transcription factors, or TFs) and transcription cofactors
to initiate transcription. The cofactors include the Mediator, chromatin-remodeling complexes, and histone-modification enzymes. Transcription cofactors are generally recruited by TFs bound to specific DNA sequences, because most of them do not have DNA binding preference. In following subsections, we discuss the DNA binding specificities of TFs, and the various roles TFs could play in transcription regulation, most of which require recruiting transcription cofactors.

1.4.1 TF and its binding preference representation

A TF recognizes and binds to its own cis-regulatory sequence (hereinafter referred to as binding sites), which is typically 5–10 bp in length and different from those recognized by all the other TFs. Such a binding trigger a series of reactions that ultimately contribute to the regulation of a gene to be transcribed at a certain rate,
including directly interacting with transcription cofactors and thereby indirectly in-
teracting with the RNAPII. Approximately 10% of the protein-coding genes (mRNA
genes) of most organisms are devoted to TFs, making them one of the largest groups
of proteins in the cell (reviewed in Alberts et al., 2014). Transcription of each gene
is regulated by its own collection of TF binding sites. These binding sites typically
are in the intergenic region immediately upstream from the TSS of the gene. The
majority of the genes have complex arrangements of TF binding sites, which are
recognized by different TFs respectively. Therefore, it is the positions, identity, and
arrangement of TF binding sites that ultimately control when, where and to what
level each gene is transcribed.

A TF recognizes a binding site because the surface of the TF is extensively
complementary to the special surface features of the double helix that displays the
DNA sequence of that binding site. Each TF makes a series of contacts with the
DNA, involving hydrogen bonds, ionic bonds, and hydrophobic interactions (reviewed
in Alberts et al., 2014). Although each individual contact is weak, the 20 or so
contacts that are typically formed at the TF–DNA interface add together to ensure
that the interaction is both highly specific and very strong.

For each TF, the common features of binding sites are often summarized in the
form of a motif. A motif is derived by comparing the DNA sequence of many TF
binding sites and tallying up the most common nucleotides found at each position. It
therefore serves as a summary or average of a large number of individual TF binding
sites. Therefore, motif is usually used to represent the sequence preference of a TF
binding to DNA. It is usually given via position-specific scoring matrix (PSSM). A
PSSM is a matrix with four rows, which correspond to nucleotides A, C, G, and
T, and \( n \) columns, where \( n \) is the number of nucleotides commonly bound by the
TF. The four cells of a column in the PSSM contain real numbers between 0 and
1, representing the relative frequency with which each of the four nucleotides is
Table 1.1: Example PSSM of MBF (Mbp1p/Swi6p) in yeast *S. cerevisiae*. Please note that the Mbp1p is the binding component of MBF (Mbp1p/Swi6p).

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2297</td>
<td>0.9379</td>
<td>0.0008</td>
<td>0.0455</td>
<td>0.0007</td>
</tr>
<tr>
<td>0.0661</td>
<td>0.0019</td>
<td>0.9896</td>
<td>0.0010</td>
<td>0.9355</td>
</tr>
<tr>
<td>0.5371</td>
<td>0.0144</td>
<td>0.0041</td>
<td>0.9528</td>
<td>0.0009</td>
</tr>
<tr>
<td>0.1671</td>
<td>0.0458</td>
<td>0.0055</td>
<td>0.0007</td>
<td>0.0629</td>
</tr>
</tbody>
</table>

Figure 1.7: The sequence logo of MBF (Mbp1p/Swi6p) in yeast *S. cerevisiae*.

observed in the corresponding position of the experimentally observed DNA binding sites. These four frequencies sum to 1. Therefore, with PSSM, we could model the TF binding process in a probabilistic approach, as in COMPETE (Wasson and Hartemink, 2009). An example PSSM of MBF, which is a complex of Mbp1p and Swi6p and has Mbp1p as binding component, is shown in Table 1.1. An accurate way of displaying a PSSM is through the use of a sequence logo. As shown in Figure 1.7, the height of each letter is proportional to the frequency at which that nucleotide occurs at that position of the motif. The total height of all the letters at each position is proportional to the information content (expressed in bits) at that position.

1.4.2 TFs can regulate transcription in various ways

TFs can act as either activators, which increase the rate of transcription as much as 1000-fold (reviewed in Alberts et al., 2014), or repressors, which decrease the rate of transcription, including almost eliminating it in some genes. The roles of activators and repressors TFs played are determined by a broad class of multisubunit proteins termed transcription cofactors, which assemble on DNA with TFs. Correspondingly, these transcription cofactors can act as coactivators and corepressors. Typically,
Figure 1.8: Eukaryotic TFs assemble into complexes on DNA. (A) Seven TFs are shown. They are not able to assemble in resolution. (B) The nature and function of the complex, in which a TF participates, depends on the specific binding site that seed the assembly of the complex, and the other participants, especially cofactors. Please note that the emphasized green and dark green TFs are shared by both activating and repressing complexes. (Adapted from Alberts et al. (2014).)

These cofactors do not have DNA binding preference themselves; they are brought to specific DNA sequences by associated TFs. As shown in Figure 1.8, often the protein–protein interactions between TFs and cofactors are too weak for them to assemble in solution; however, the appropriate DNA sequences at TF binding site can “crystallize” the assembly of these complexes on DNA.

Please note, an individual TF can often participate in more than one type of regulatory complex. With coactivator, a TF might function as part of a complex that activates transcription. With corepressor, the same TF might function as part of a complex that represses transcription. Thus, individual eukaryotic TFs function as regulatory parts that are used to build complexes whose function depends on the final assembly of all of the individual components, as demonstrated in Figure 1.8B. Therefore, we firstly discuss the ways in which TFs can act as activators, and then discuss the ways in which TFs can act as repressors.
1.4.3 The ways in which TFs can act as activators

For being activators, TFs mainly function through attracting and positioning RNAPII at the core promoter, and releasing it so that transcription can begin. As shown in Figure 1.9A, a TF could be an activator through changing the local chromatin to allow the binding of additional activators. Repeatedly using this principle, large assemblies of proteins can form on regulatory regions of genes to regulate their transcription. Several statistical thermodynamic theory-based models, including COMPETE developed by our group previously, have incorporated such a mechanisms by allowing TFs competing with nucleosome to bind to DNA sequences (Raveh-Sadka et al., 2009; Wasson and Hartemink, 2009; He et al., 2010; MacIsaac et al., 2010).

Some activators interact directly to one or more of the general transcription factors, accelerating their assembly on a core promoter. Such direct interactions will require DNA looping to bring the general transcription factors in proximity to the activators, if they are binding to enhancers. However, most TFs function as activators by attracting coactivators that then perform the biochemical tasks needed to initiate transcription. One of the most prevalent coactivators is the large Mediator complex, consisting of more than 30 subunits. Almost as large as RNAPII itself, Mediator serves as a bridge between DNA-bound activators, RNAPII, and the general transcription factors, facilitating the assembly of PIC at the core promoter. Moreover, the eukaryotic general transcription factors and RNAPII are not capable to assemble on a core promoter that is packaged in nucleosomes on their own. Thus, in addition to directing the assembly of the PIC at the core promoter, eukaryotic activators promote transcription by triggering changes to the chromatin structure of the core promoters, enhancing the DNA accessibility. These various mechanisms about transcription activators helping PIC formation and thereby RNAPII recruitment are demonstrated in Figure 1.9B.
Figure 1.9: Eukaryotic TFs can act as activators at different steps of transcriptional regulation. (A) As an activator, a TF promotes binding of additional activators. (B) As an activator, a TF helps PIC formation and thereby RNAPII recruitment through direct interactions or indirect interactions. (C) As an activator, a TF helps to release already assembled RNAPII from the promoter. (D) As an activator, a TF releases RNAPII molecules that become stalled after transcribing about 50 nucleotides or RNA. (Adapted from Alberts et al. (2014).)

In some cases, as shown in Figure 1.9C, transcription initiation requires that a DNA-bound activator releases RNAPII from the core promoter so that it begins elongation. In other cases, as demonstrated in Figure 1.9D, the RNAPII pauses after transcribing about 50 nucleotides of RNA, and further elongation requires a activator bound behind it. These paused polymerases are common in metazoans, where a significant fraction of genes that are not being transcribed have a paused RNAPII located just downstream from the promoter (reviewed in Adelman and Lis, 2012;
Alberts et al., 2014). For yeast *S. cerevisiae*, RNAPII usually displays a relatively uniform distribution across the promoter and the corresponding coding sequence (reviewed in Wade and Struhl, 2008; Adelman and Lis, 2012). This observation suggests a rapid transition from initiation to elongation and RNAPII experiences few regulatory barriers (reviewed in Wade and Struhl, 2008; Adelman and Lis, 2012), which is consistent with the very strong correlation between the amount of core promoter-bound TBP and the transcription level (Kuras and Struhl, 1999; Li et al., 1999).

The release of RNAPII can occur in several ways. In some cases, the activator brings in a chromatin remodeling complex that removes a nucleosome, which is barrier of the elongating RNAPII. In other cases, the activator communicates with RNAPII (typically through a coactivator), signaling it to move ahead. Finally, RNAPII requires elongation factors to effectively transcribe through chromatin. In some cases, the key step in gene activation is the loading of these factors...
onto RNAPII, which can be directed by DNA-bound activators. Once loaded, these factors allow the RNAPII to move through barriers imposed by chromatin structure and transcribing the gene successfully. Having RNAPII already poised on a core promoter in the beginning stages of transcription bypasses the step of assembling PIC components at the core promoter, which is often slow. This mechanism can therefore allow cells to begin transcribing a gene as soon as receiving an extracellular signal.

At most genes, more than one of the above mechanisms work together to increase the rate of transcription.

1.4.4 The ways in which TFs can act as repressors

As shown in Figure 1.10, there are six ways in which a TF can act as a repressor. (A) To be a repressor, a TF could compete binding site with an activator. Several statistical thermodynamic theory-based models, including COMPETE developed by our group previously, have incorporated such a mechanisms by allowing TFs competing with each other to bind to DNA sequences (Raveh-Sadka et al., 2009; Wasson and Hartemink, 2009; He et al., 2010; MacIsaac et al., 2010). (B) A repressor can mask the activation surface of an activator, preventing it from carrying out its function. (C) A repressor can block the assembly of the general transcription factors, by binding to the interaction domain of a certain general transcription factor, like TFIID. (D) A repressor can recruit a chromatin remodeling complex, which returns the nucleosomal state of the core promoter region to its pre-transcriptional form: being unaccessible for PIC components and RNAPII. (E) A repressor can attract a histone modifying enzyme, which is histone deacetylase in this case, to the core promoter, and therefore reverse the histone acetylation that can stimulate transcription initiation. (F) A repressor can attract a histone methyltransferase, which modifies certain positions on histones by attaching methyl groups; the methylated histones,
in turn, are bound by proteins that maintain the chromatin in a transcriptionally silent form.

Among these mechanisms, (D), (E), and (F) are about that repressors act by bringing corepressor to DNA, which is typical for eukaryotic repressors. Also, like transcription activation, more than one mechanisms above can function together to ensure extremely efficient repression.

1.4.5 The big picture of eukaryotic transcription initiation in vivo

In this subsection, we summarize the big picture of eukaryotic transcription initiation *in vivo* that we have discussed. Although only one activator is shown in Figure 1.11, a typical eukaryotic gene utilizes many TFs, which in combination determine its transcript production rate and pattern of transcription. TFs can act as either activators or repressors, mainly through directly binding the RNAPII, or indirectly interacting with the RNAPII via binding transcription cofactors. The transcription cofactors are regulatory proteins that typically have no DNA binding specificities: the general transcription factors, Mediator, chromosome remodeling complexes, and histone modifying enzyme. In addition, TFs attract ATP-dependent chromatin remodeling complexes and histone-modifying enzymes to increase or depress transcript production rates. For simplicity, nucleosomes are not shown in the figure. Also, as we discussed in this and previous sections, many proteins (well over 100 individual subunits in total) must assemble at the TSS to initiate transcription in a eukaryotic cell (reviewed in Alberts et al., 2014). This is illustrated in Figure 1.11 too. The order of assembly of these proteins does not follow a prescribed pathway; rather, the order differs from core promoter to core promoter. Indeed, some of these different protein complexes may be brought to core promoter as preformed sub-assemblies.

Sometimes acting from an enhancer that is several thousand bp away (indicated by the dashed DNA molecule in Figure 1.11), the TFs still can help RNAPII, the
general transcription factors, and Mediator all to assemble at the core promoter. In metazoans, it is common for dozens of TFs to regulate the transcription of a single gene, with enhancers (and silencers) spread over tens of thousands of bp (reviewed in Bulger and Groudine, 2011; Alberts et al., 2014; Allen and Taatjes, 2015; Kim and Shiekhattar, 2015). It is the DNA looping that allows the DNA-bound TFs to interact transcription cofactors and ultimately with RNAPII at the core promoter. Precisely how the loops are formed and stabilized remains unclear, but Mediator, in conjunction with other proteins, seems to have important roles (Kagey et al., 2010; Phillips-Cremins et al., 2013; Hnisz et al., 2013; Muto et al., 2014). In the yeast *S. cerevisiae*, genes are not regulated by long distance enhancer–core promoter interactions; TF binding sites are usually limited to UASs that are located within a few hundred bp of the core promoter (reviewed in Bulger and Groudine, 2011; Allen
and Taatjes, 2015; Kim and Shiekhattar, 2015). Relocation of a UAS at greater distances generally results in loss of function. However, it is notable that, probably through DNA looping, mutant forms of Sin4p were identified in *S. cerevisiae* that enabled UASs to activate transcription at longer distances from the core promoter than normally observed in yeast (UASs are normally only 100–200 bp from the core promoter TATA element) (Dobi and Winston, 2007). These observations together led to the realization that enhancer activity is the primary mechanism for the precise control of spatiotemporal patterns of gene expression in multicellular organisms.

Furthermore, in yeast *S. cerevisiae*, DNA looping allows the core promoter–terminator interactions: Mediator has been implicated in the formation of looped DNA structures between the 5′ and 3′ ends of genes (Mukundan and Ansari, 2013). Because some yeast genes are in a looped configuration during activated transcription, Mukundan and Ansari (2013) proposed that the DNA looping plays a role in termination of transcription, and may be facilitating the transfer of RNAPII from the terminator to the core promoter for reinitiation, and thereby promote the efficiency of transcription.

### 1.5 Regulation of transcription elongation in eukaryotes

Once the RNAPII has begun elongation, those general transcription factors released from the DNA are available to initiate another round of transcription with a new RNAPII molecule. Meanwhile, the elongating RNAPII is associated with a series of elongation factors, which decrease the possibility that RNAPII will dissociate before it reaches the 3′ end of a gene. These factors typically associate with RNAPII shortly after initiation and help it walk through gene body DNA sequence. Moreover, in eukaryotes, a RNAPII must cope with chromatin structure as it moves along a template DNA, and they are typically aided by ATP-dependent chromatin remodeling complexes that either move with the RNAPII or simply seek out and rescue the
occasionally stalled RNAPII. In addition, histone chaperones help by partially disassembling nucleosomes in front of a moving RNAPII and assembling them behind. As RNAPII moves along a gene, some of the enzymes bound to it modify the histones, leaving behind a record of where the polymerase has been. This information may aid in transcribing a gene over and over again once it has become active for the first time. It may also be used to coordinate transcription elongation with the processing of RNA as it emerges from RNA polymerase.

In eukaryotes, transcription is only the first step for producing a mature mRNA molecule. Other important steps include the covalent modification of the ends of the mRNA, and the removal of intron sequences, which are discarded from the middle of the mRNA transcript by RNA splicing. Both ends of eukaryotic mRNAs are modified: by capping on the 5 end and by polyadenylation of the 3 end. These special ends allow the cell to assess whether both ends of an mRNA molecule are present (and if the information is therefore intact) before it exports the mRNA from the nucleus and translates it into protein. Alternative RNA splicing joins together the exon sequences, and it provides eukaryotes with the ability to synthesize multiple different proteins from the same gene.

A strategy has evolved to couple all of the above RNA processing steps to transcription elongation. The phosphorylation of the RNAPII tail (C-terminal domain, CTD), which proceeds gradually as the RNAPII initiates transcription and moves along the DNA, not only helps dissociate the RNAPII from other proteins present at the TSS, but also allows a new set of proteins to associate with the RNAPII tail that function in transcription elongation and RNA processing. Some of these processing proteins are thought to “hop” from the RNAPII tail onto the nascent RNA molecule to begin processing it as it emerges from the RNAPII. Therefore, we can view elongating RNAPII as an RNA factory that not only moves along the template DNA synthesizing an RNA molecule, but also processes the nascent RNA molecule. Fully
extended, the RNAPII tail is nearly 10 times longer than the remainder of RNAPII (reviewed in Alberts et al., 2014). As a flexible protein domain, the RNAPII tail serves as a scaffold, holding a variety of proteins close by so that they can rapidly act when needed.

Although alternative RNA splicing is very common in metazoans, it is extremely rare in single-celled yeast *S. cerevisiae*. There are ~6,200 genes in yeast *S. cerevisiae*; only about 300 of them are subject to RNA splicing, and nearly all of these 300 genes have only a single intron (reviewed in Alberts et al., 2014).

1.6 Regulation of transcription termination in eukaryotes

The position of the 3′ end of each mRNA molecule is specified by termination signals encoded in the genome, which are a variety of individual cleavage and polyadenylation signals (see Alberts et al., 2014, for details). These signals are transcribed into RNA as the RNAPII walks through them, and they are then recognized (as RNA) by a series of RNA-binding proteins and RNA-processing enzymes. There are two such multisubunit proteins of special importance: CstF (cleavage stimulation factor) and CPSF (cleavage and polyadenylation specificity factor). Both of them travel with the RNAPII tail and are transferred to the 3′-end processing sequence on an RNA molecule as it emerges from the RNAPII. Once CstF and CPSF bind to their recognition sequences on the emerging RNA molecule, additional proteins assemble with them to create the 3′ end of the RNA. Then, the RNA is cleaved from the RNAPII. Next an enzyme called poly-A polymerase (PAP) adds approximately 200 A nucleotides to the 3′ end produced by the cleavage, one at a time. The nucleotide precursor for these additions is ATP, and the same type of 5′-to-3′ bonds are formed as in conventional RNA synthesis. However, unlike RNAPII, PAP does not require a template DNA; therefore the poly-A tail of eukaryotic mRNAs is not directly encoded in the genome. As the poly-A tail is synthesized, poly-A-binding proteins
assemble onto it and help determine the final length of the tail by a poorly understood mechanism.

After the 3′-end of a eukaryotic pre-mRNA molecule has been cleaved, the RNAPII continues to transcribe, in some cases for hundreds of nucleotides. Once 3′-end cleavage has occurred, the newly synthesized RNA that emerges from the polymerases lacks a 5′ cap; this unprotected RNA is rapidly degraded by a 5′ → 3′ exonuclease carried along on the RNAPII tail. It is this continued RNA degradation that eventually causes the RNAPII to be released from the template DNA and terminate transcription.

1.7 A focus on transcriptional regulation in yeast *S. cerevisiae*

As we discussed in previous sections, the complexity of transcriptional regulation in eukaryotes is daunting. But transcriptional regulation is the most important step in the regulation of gene expression. As a starting point to unravel this complexity, it makes sense to select a model organism that is unicellular and as simple as possible. The popular choice for this role of minimal model eukaryote has been the budding yeast *S. cerevisiae*, which is the species used by brewers of beer and bakers of bread, and therefore called as “baker’s yeast” and “brewer’s yeast”.

Comparing to higher order eukaryotes, which are multicellular, the yeast *S. cerevisiae* has the following advantages with regard to modeling simplicity:

- The genome of yeast *S. cerevisiae* is exceptionally small by eukaryotic standards. Its genome has \( \sim 13 \times 10^6 \) nucleotide pairs, and \( \sim 6600 \) genes, while the human genome contains \( \sim 3200 \times 10^6 \) nucleotide pairs, and \( \sim 30,000 \) genes. The smallest size of yeast *S. cerevisiae* genome means that we have the smallest number of genes, TFs, and thereby the smallest regulatory network, among all eukaryotes.
• As discussed in section 1.4.5, the yeast *S. cerevisiae* genes are not regulated by long distance enhancer–core promoter interactions; TF binding sites are usually limited to UASs that are located within a few hundred bp of the core promoter. Therefore we have no need to identify enhancers and predict their respective interactions with the core promoter, which remain an extremely difficult challenge and is an active research topic.

• As discussed in section 1.5, alternative RNA splicing is extremely rare in single-celled yeast *S. cerevisiae*. Therefore we have not need to model the synthesis of multiple mRNAs from the same gene.

Moreover, though it is a minimal model eukaryote, the yeast *S. cerevisiae* suffices for all the basic tasks that every eukaryotic cell must perform. Therefore, we choose yeast *S. cerevisiae* as our model organism and tackle this big question: whether it is possible to predict dynamic transcript production rates over time from DNA sequences and DBF binding specificities. A good answer to this question will promote our understanding of the precise regulation of temporal gene expression in yeast *S. cerevisiae*, and thereby will help to obtain insights for the precise control of spatiotemporal patterns of gene expression in multicellular organisms.

1.8 Dissertation outline

The rest of this dissertation is organized as follows. In chapter 2, we summarize the hypothesized mechanisms about TFs binding to DNA, TFs interacting with transcriptional machinery, the various functioning approaches of regulatory elements within core promoter, and the nucleosomes resisting RNAPII elongation. Because of various reasons, we exclude some hypothesized mechanisms for modelling. We discuss the respective reasons of exclusion in corresponding sections.

In chapter 3, we mainly introduce how we design features to represent the hy-
pothesized mechanisms discussed in chapter 2 that we decide to model. We design three groups of features: the “fluctuation” group for the TFs’ roles, the “amplifier” group for the core promoter’s role, and the “resistance” group for the gene body nucleosomes’ roles. For each group, we design subgroups of features to reflect the multiple mechanisms for the associated regulatory factors and elements. Moreover, we add certain interaction terms to represent hypothesized mechanisms we discuss in chapter 2. We then describe the hierarchical grouping structure of these features we have because of our prior knowledge. Also, we discuss the within-group and group-wise correlations we have for these features. We lastly describe the way in which we preprocess the feature values.

In chapter 4, we describe a machine learning framework called PreTrans, which can take as input DNA sequences, DBF PSSMs, and DBF concentrations, to learn models for predicting the mean or dynamic transcript production rate that any promoter DNA sequence will produce when receiving the DBF concentrations. Then, as we need interpretable and highly predictive models, we discuss the advantages and disadvantages of respective sparse linear models, especially group-wise sparse linear models. And finally we introduce our algorithm for this complicate problem: in a cross-validation approach, we perform gRFE combining the ridge in a sequence of steps to promote subgroup-wise and group-wise sparsity.

In chapter 5, we prepare for predicting yeast dynamic transcript production rates in cell cycle. We firstly describe how we determine the TFs and genes set we study. Secondly we describe how we prepare inputs for running COMPETE to get occupancy profiles for extracting features. Especially, we describe how we estimate TF concentrations. Then we discuss why time series analysis is not appropriate for our work, and the autocorrelation problem that we take care of. Lastly, we introduce the different ways to do training-test splitting for cross-validation on cell cycle data set. And we explain why we choose stratified training-test split, where two randomly
selected time points for each gene are assigned into test set, and, one time point
for one fold, all remaining time point for each gene are randomly splitted into the
validation folds.

In chapter 6, we apply our framework to cell cycle dataset. We record the iter-
ations that our algorithm performs on the chosen data set, and discuss the certain
decisions it makes on eliminating (sub)groups of features. We demonstrate that Pre-
Trans performs very well on test set for predicting dynamic transcript production
rates over cell cycle. By interpreting our final model with excellent predictive perfor-
ance, and comparing our final model and its counterparts, we provide support for
certain hypotheses. Especially, we show that transcriptional synergy between TFs
and core promoter regulatory elements is crucial for predicting transcript produc-
tion rates over cell cycle. We also discuss the possible explanations for why certain
hypothesized mechanisms could not get support.

In chapter 7, we prepare for predicting yeast mean transcript production rate over
time for synthetic promoters. Firstly, we briefly introduce the experimental protocol
about synthesizing promoter sequences, which is recently widely used for studying
transcriptional regulation. Then we introduce a specific data set: the data set of
DREAM6 gene promoter expression prediction challenge. This data set is about
promoters of ribosomal protein (RP) genes. We describe the scoring metrics used
in the challenge, which will be helpful for comparing the performance of our model
with that of other models in chapter 8. We then prepare inputs for COMPETE to
generate occupancy profiles and extract features.

In chapter 8, we apply our framework to DREAM6 data set, to show that our
modeling framework could work well for synthetic promoter sequences too. Espe-
cially, we show that our framework PreTrans beats all participant teams of DREAM6,
the “best of all” pseudo model that takes the best score of all teams at each metric,
and the Meyer et al. (2013) model that combined the FiRST team in DREAM6
and the Zeevi et al. (2011). We compare the gRFE path of eliminating features on DREAM6 data set with that on cell cycle data set, showing that the features selected by our models on different data sets are consistent, and thereby our algorithm works in a consistent way. Also, with DREAM6 results, our framework shows its capability to identify the relevant features for specific setting, even those features are not selected for cell cycle study. Moreover, we show the helpfulness of core promote relevant features, and discuss the difficulty of defining it. And given that such synthetic promoter data sets usually have fixed gene body, we demonstrate that COMPETE is capable of inferring influence of mutations far away and thereby is very helpful for feature engineering. In chapter 9, we discuss lessons our contributions and directions for further improvements
Hypothesized transcriptional regulation mechanisms in yeast *S. cerevisiae*

For predicting transcript production rates, our approach is modeling hypothesized transcriptional regulation mechanisms by extracting and grouping features from DNA sequence and DNA occupancy profile. Please note the DNA occupancy profile is an intermediate output, which is predicted from DNA sequence and DBF binding specificities by COMPETE (Wasson and Hartemink, 2009). Therefore, our approach could not directly model the transcription cofactors, like the general transcription factors, Mediator, chromatin-remodeling complexes, and histone-modification enzymes, because most of them do not have DNA binding specificities. But the transcription cofactors are usually recruited by TFs. And, as we discussed in section 1.3, there are sequence features reflecting binding specificities of certain subunits of some transcription cofactors, like TBP and TAFs, as well as signals relevant to the functioning mechanisms of some cofactors. Therefore, we assume at least some hypothesized mechanisms relevant to functioning of transcription cofactors could be represented by the features regarding associated TFs and sequence signals.
In this chapter, we summarize the hypothesized mechanisms about TFs binding to DNA, TFs contributing to transcript production, the various functioning approaches of regulatory elements within core promoter, and the nucleosomes resisting RNAPII elongation. Moreover, we exclude some hypothesized mechanisms for modelling, for which we discuss the respective reasons of exclusion in corresponding sections.

2.1 Hypothesized mechanisms relevant to TF binding to DNA

As demonstrated in Figure 2.1, there is a hypothesis that the nucleosome competes with TFs for binding sites and, thereby regulate transcription. Wasson and Hartemink (2009) and Raveh-Sadka et al. (2009) suggested that nucleosomes should be modeled, since they compete with TFs for DNA binding sites. Raveh-Sadka et al. (2012) showed that nucleosome-disfavoring sequences, specifically poly(dA:dT) tracts, next to TF binding sites significantly influence the transcript production of the reporter gene, by constructing 70 distinct promoter in a sequence context based on yeast HIS3 promoter, with poly (dA:dT) tracts varying in their length, composition and distance from several distinct TF binding sites (mainly used Gcn4p sites, but used Gal4p sites and Pho4p sites too in 5 out of 70 promoter variants for showing
that the effect of poly(dA:dT) tracts is independent of the TF identity). To show these findings generalize more broadly, Sharon et al. (2012) devised 777 promoter variants by separately inserting consensus sites for 14 TFs into two distinct promoter background sequences. Among these promoter variants, either the TF consensus site location or the location, length and/or orientation of the poly(dA:dT) tract were varied. Moreover, probing the corresponding transcriptional activity in single cells with time-lapse microscopy, Bai et al. (2010) found that TFs’ binding sites contained in nucleosome depleted region, which is the other name of NFR, reduce the variability of gene expression: with nucleosome buried Swi4p/Swi6p cell-cycle box (SCB), SCB-binding factor (SBF) activation is strong in some cell cycles but is undetectable in others; in contrast, SCBs contained in NFR lead to reliable activation, once per cell cycle. In section 3.1, we will see that the mechanism about nucleosome-mediated effect on TF binding is modelled by our framework implicitly, exploiting COMPETE to get occupancy profiles.

In addition to nucleosomes influencing TF binding probability, we propose that nucleosomes might influence the stability of nearby TF binding instances, and consequently their respective contributions to transcript production. This is inspired by the work of Lickwar et al. (2012). As shown in Figure 2.2, the authors observed that

![Figure 2.2: Plots of the average nucleosome occupancy for each group centred on the Rap1p motif. X-axis is about absolute distance from Rap1p motif (bp). (A) Rap1p-bound loci are grouped into four categories based on their measured Rap1p residence time: longest, long, short, and shortest. (B) Rap1p-bound loci are grouped into four categories based on their measured Rap1p occupancy: highest, high, low, and lowest. (Adapted from Lickwar et al. (2012).)]](image-url)
nucleosomes in proximity to the DNA site bound by Rap1p would have a strong influence on Rap1p residence time, which is not observed for Rap1p occupancy. They proposed that TFs probably need long residence time to recruit the RNAPII and initiate transcription, and that TFs competing with nucleosomes for sites bind DNA in short pulses, which are not sufficient for transcription from DNA into RNA. In section 3.1, we will see that our framework models this mechanism explicitly by weighting relevant features with TF binding stability.

Moreover, there are some hypothesized mechanisms we do not model (See Figure 2.3). As shown in Figure 2.3A, it is proposed that TF binding specificity could be modulated by transcription cofactors (reviewed in Spitz and Furlong, 2012; Levo and Segal, 2014). As demonstrated in Figure 2.3C, it is proposed that The base pairs flanking a TF binding site can influence the binding specificity of the TF, possibly by affecting the local DNA shape (reviewed in Levo and Segal, 2014). We do not model these two mechanisms because they require sufficient experimental data to update relevant TF PSSMs.

As illustrated in Figure 2.3B, it is proposed that TF binding to DNA might be facilitated by direct interactions between TF instances that bind to adjacent sites (Spitz and Furlong, 2012; Levo and Segal, 2014). Cooperative TF binding is often thought to result from protein–protein interactions between TFs bound to adjacent sites, thus enhancing each other’s binding to DNA, and thereby possibly promoting PIC formation and RNAPII recruitment (Spitz and Furlong, 2012).

We do not model this mechanism because: a) there are rare evidences about this mechanism in yeast \textit{S. cerevisiae}; b) modelling this mechanism by allowing interactions between features representing TFs’ roles will make our already large feature set extremely large; c) COMPETE has implicitly modelled cooperative TF binding by allowing TFs bound to adjacent sites to compete with the same nucleosome for DNA binding; d) Levo et al. (2015) showed that, in yeast \textit{S. cerevisiae}, TF binding
events to multiple sites are generally independent to each other.

Lastly, we discard another hypothesis for modeling: the ~10bp periodic relationship between gene expression and TF binding instance location (See Figure 2.4A), which is specific to the combination of TF and promoter and depends on the helical phase (Sharon et al., 2012). However, as shown in Figure 2.4B, Levo et al. (2015) has shown that this ~10bp periodic relation might be resulted from the changes of TF binding strength caused by site flanks.
Figure 2.4: We do not model the ~10 bp periodic relationship between gene expression and TF binding instance location. In both panels, the mGal1-10 background and GAL1-10 background are the same sequence derived from the native yeast Gal1-10 promoter in which the known regulatory elements were mutated. (A) Each point corresponds to the average expression of eight sets of promoters in which Sharon et al. (2012) changed the location of the Gcn4p site, where the eight different sets differ in the location of a poly(dA:dT) tract of length 15 bp. To normalize the expression across the eight different sets, the expression is defined as a robust z score, calculated by subtracting the median and dividing by the s.d. of expression differences from the median. There is a ~10 bp periodicity of expression observed over five periods. (B) With settings same as (A), Levo et al. (2015) designed respective regulatory sequences with Gcn4p TF binding site of 9 bp (in blue) or 23 bp, including fixed-flanks site (in red) placed along the same context. In this panel, the y axis is z-score of TF binding, instead of gene expression. Indeed, the ~10 periodicity is about TF binding. However, it might be resulted from random flanks of binding sites: with fixed-flanks, the periodicity disappeared. (Adapted from Sharon et al. (2012) and Levo et al. (2015).)

2.2 Hypothesized mechanisms relevant to the contribution of TF to transcript production

Recently, researchers have applied statistical thermodynamic theory in the study of transcriptional regulation (Bintu et al., 2005; Segal et al., 2008; Raveh-Sadka et al., 2009; He et al., 2010; Zeevi et al., 2011). These thermodynamic-based models assume that each TF binding instance contributes independently to the total attractive force recruiting the RNAPII to the promoter, with activators helping to increase transcript production rate and repressors help to decrease transcript production rate. However, as discussed in section 1.4.2, the roles of activators and repressors TFs played are usually determined by transcription cofactors. And an individual TF could be either
activator or repressor, depending on with which type of cofactor it participating in the regulatory complex. Therefore, we do not explicitly model a TF to be activator or repressor. Instead, we just let our modelling framework assign the appropriate sign to the features associated with a TF: positive sign would mean increasing transcript production rate as activator, and negative sign would mean decreasing transcript production rate as repressor. Also, we take the hypothesis that different TF’s contribution to transcript production might be distinct and non-redundant from each other, and therefore each TF would have its own associated features.

As shown in Figure 2.5, Eser et al. (2014) proposed that TFs govern the timing but not the magnitude of the transcription rate of their target genes. Also, it has been suggested that TFs influence the PIC formation and RNAPII recruitment by interacting with PIC components directly or indirectly through Mediator and other cofactors (See sections 1.4.3 and 1.4.4). We propose a hypothesis that TFs govern the probability of RNAPII recruitment and PIC formation for all genes: it is not only about the timing, but also about magnitude of RNAPII recruitment and PIC formation. Our hypothesis is based on the knowledge that TF has binding sites with different strength across distinct promoters and, thereby the total RNAPII recruitment force determined by TF binding instances might be different. Especially, with fixed core promoter, recent works reported that expression can be fine-tuned through changing TF binding site affinities and locations (Sharon et al., 2012; Rajkumar et al., 2013).

Moreover, there should be a maximum value of transcript production rate for a certain gene. Otherwise, the transcript production rate could be extremely large, if we keep adding activator binding sites to the promoter. Segal et al. (2008) proposed there should be saturation for the contribution of TFs to transcript production rate, and they used logistic function to model this. Further, we propose that, for each TF, a single TF binding instance would be enough for its functioning.
Figure 2.5: Eser et al. (2014) reported time courses of 22 periodically expressed genes that are exclusively annotated as Mbp1p targets. These 22 genes agree well in the timing of gene expression, but show a remarkable diversity in their mean and magnitude of transcript production rate (Adapted from Eser et al. (2014)).

This hypothesis is based on two recent works. Rajkumar et al. (2013) constructed 209 variants of the single *S. cerevisiae* yeast PHO5 promoter, the core promoters of which are identical. These PHO5 mutant promoters consisting of Pho4p binding-site variants, ablations of Pho4p or Pho2p sites. They showed, a relatively simple model based on *in vitro* TF binding specificities was able to explain 95% of the observed variance in gene expression. Specifically, they defined and used the TF occupancy probability as feature in their model, which is equal to a probability of having at least one TF binding instance. For two *S. cerevisiae* yeast background promoter sequences, Sharon et al. (2012) respectively inserted the consensus sites for Gcn4p and Gal4p in all $2^7(128)$ and $2^5(32)$ possible combinations of sites at 7 and 5 certain locations. As shown in Figure 2.6, they observed that expression is a monotonic function of TF binding sites that saturates at 3–5 binding sites. Then they extended
Expression is a monotonic function of TF binding sites that saturates at 3–5 binding sites. (A) Within two different promoter backgrounds, Sharon et al. (2012) separately inserted Gcn4p sites in all $2^7(128)$ possible combinations of sites at seven predefined locations within the promoter. Shown are the individual promoter expressions for each background and mean expressions of all promoters that have $k$ Gcn4p sites for $k = 0, 1, 2, \ldots, 7$. Also shown is a fit of a logistic function for each background. (B) As in (A), but for all $2^5(32)$ possible combinations of Gal4 sites at five predefined promoter locations. The outlier promoter in terms of expression in which the two Gal4 sites closest to the core promoter were both added is indicated. These two sites were added at a distance of 1 bp, as opposed to the 5 bp distance used between all other adjacent sites, thus suggesting steric hindrance between Gal4p sites at this distance. (Adapted from Sharon et al. (2012).)

the set to 13 additional TFs at lower resolution: only one promoter is constructed for each TF in every combination of promoter background sequence and number of binding sites. They found that, on average, expression increases monotonically with more sites of activators and then mostly saturates at 3 or 4 sites, and that the TF identity would influence the expression level at saturation point significantly. We think that, together, Sharon et al. (2012) and Rajkumar et al. (2013) suggest that transcript production of a gene in a single cell would saturate with a single TF binding instance with respect to the contribution of the TF.

In section 3.1, we will see that this mechanism about the saturation of TF binding instances is modeled by extracting the probability of at least one binding instance for each TF. We do not take the approach of exploiting logistic function, since it would require us to rescale our target values to be within the range $[0, 1]$. This would require us to know the maximum possible transcript production rate, which is not available actually.
Another hypothesis is that a TF binding instance’s contribution to RNAPII recruitment, and consequently transcript production (activation or repression), decreases as it is placed further away from TSS. As shown in Figure 2.7A, MacIsaac et al. (2010) reported that, in mouse liver, cerebellum, and 3T3-L1 cells, there is an inversely proportional relationship between the distance of TF binding site to TSS and the mean log transcript abundance on average. As shown in Figure 2.7B, with a S. cerevisiae yeast synthetic promoter library, Sharon et al. (2012) observed a similar relationship between the distance of TF binding site to TSS and its influence on gene expression on average. In section 3.1, we will see that this mechanism about the influence of TF binding instance’s distance to TSS is modeled by applying a distance decay weight to all relevant TF features.

Moreover, for some TFs, the orientation of a TF binding site probably influences the transcript production rate of the downstream gene. To profile the activity of yeast
TFs, Sharon et al. (2012) constructed 150 promoters by inserting the consensus sites of 75 distinct yeast TFs into the same promoter in two orientations. They found significant site-orientation effects for 6 of the 75 TFs (P \textless 0.05): Aft2p, Yrm1p, Adr1p, Cep3p, Rap1p, and Fhl1p. In chapter 3.1, we will model this mechanism about the orientation of TF binding instance through extracting two versions of relevant features for each TF: forward version and reverse version. Just like \textasciitilde 10 bp periodicity actually is caused by site flanks (See Figure 2.4), we think that the real causes of the observed site-orientation effects could be site flanks as well, especially given that only 6 out of 75 TFs were detected for site-orientation effects.

2.3 Hypothesized mechanisms relevant to regulatory elements within core promoter

The RNAPII core promoter is the regulatory DNA sequence around TSS, to which RNAPII and other relevant general transcription factors are recruited, form the PIC, and act to initiate transcription (Smale and Kadonaga, 2003). As we discussed in section 1.3.3, in yeast, TSS is between 40–120 bp downstream from the TATA element (could be even 250 bp downstream of the TATA element), and typical core promoters are 100–200 bp long (Smale and Kadonaga, 2003; Lubliner et al., 2013).

Though core promoter is important for transcription initiation, the regulatory elements in core promoter and the associated mechanisms mostly remain unclear for eukaryotes, especially for yeast \textit{S. cerevisiae}. For most eukaryotes, as shown in Figure 1.4A, the main identified regulatory elements in core promoter are BREs, TATA box, INR, and DPE. However, as discussed in sections 1.3.1, BREs and DPE do not present in yeast \textit{S. cerevisiae} core promoters, and INR might function differently in yeast \textit{S. cerevisiae} than that in metazoans: with INR mutations, promoter activity is often unaffected or reduced only modestly in yeast. Also, as we discussed in section 1.3.1, for all eukaryotes, TATA box does not always present in a core promoter: for
yeast S. cerevisiae, only 17% of 3,945 RNAPII core promoters contain TATA box (Rhee and Pugh, 2012). And Kamenova et al. (2014) claimed ~15% yeast RNAPII core promoters are TATA box-containing promoters, based on Basehoar et al. (2004) and Seizl et al. (2011). Correspondingly, approximately 80-90% of all S. cerevisiae core promoters are designated as “TATA-less”, and have a predominant PIC assembly mechanism and chromatin architecture that differs substantially from those in the “TATA box-containing” category of core promoters (Lee et al., 2000; Basehoar et al., 2004; Huisinga and Pugh, 2004; Tirosr and Barkai, 2008). So, in yeast S. cerevisiae, the only identified regulatory element reflecting the DNA binding specificities of the general transcription factors (and other transcription factors) is TATA box, which is not contained in most RNAPII core promoters.

On the other hand, some general transcription factors and their subunits, which might have DNA binding specificities and thereby are highly promoter selective, are shown to be important for core promoter-dependent transcription initiation in yeast S. cerevisiae. Green (2000) reported that, as in higher eukaryotes, yeast TAFs, are important for core promoter-dependent transcriptional regulation. Specifically, in yeast, TFIIB and some TAFs were shown to be crucial at the different steps of two distinct transcription activation pathways: the TFIID pathway and the SAGA pathway (Bhaumik, 2011; Kamenova et al., 2014, reviewed in). Note TFIIB recognizes BREs in metazoans, and some TAFs have DNA binding specificities to core promoter (See section 1.3). Therefore, all these suggest the existence of still unidentified TAF-dependent and other cofactor-dependent core promoter motifs (regulatory elements) in yeast.

Fortunately, though the associated mechanisms remain uncertain, sequence features extracted from core promoters have been shown to be helpful for predicting promoter activity, i.e., mean transcript production rate over time (Meyer et al., 2013; Lubliner et al., 2013; Zeevi et al., 2014). In the DREAM6 Gene Promoter Expres-
esion Prediction challenge, as shown in Figure 8.1A, it was reported that sequence features extracted from the window of $[-100, -1]$ of TrSS are predictive of time average promoter activity for a group of natural and mutant promoters of ribosomal protein gene (Meyer et al., 2013). More recently, Lubliner et al. (2013) figured out that sequence features extracted from the $[-75, 50]$ region of TSS could be predictive of maximal promoter activity for yeast and human sequence; Zeevi et al. (2014) reported that features extracted from the window of $[-200, -1]$ of TrSS could explain 65% of the variance in orthologous yeast ribosomal promoter activity (See Figure A.1); and Lubliner et al. (2015) showed k-mer count features extracted from the window of $[-118, -1]$ of TrSS could determine time average expression level of 13,000 synthetic yeast ribosomal protein gene promoters to a significant extent (See Figure A.2). Consequently, it is not a surprise that Eser et al. (2014) reported that PIC components measurements from Rhee and Pugh (2012) are correlated with time average transcript production rates, as shown in Figure 2.8. Instead of TSS, these works all used TrSS as pivot to identify the core promoter region for extracting sequence features. Possible reasons are: a) some genes miss TSS information; b) a few genes have mapped TSS far away from TrSS; c) lots of genes have mapped TSS within the $[-118, -1]$ region of TrSS. These are demonstrated in Figure 3.1.

Given that sequence features extracted from core promoters have been shown to be helpful for predicting promoter activity in human and especially yeast, and PIC components amount are correlated with promoter activity, we propose that core promoter amplifies transcript production rates, possibly by influencing RNAPII recruitment, PIC formation, and transcriptional initiation.

We also want to test the hypothesis that the regulatory role of the core promoter could be explained by the mechanisms about TATA elements and nucleosomes within core promoter. Examples of such mechanisms that we are able to test include: a) TATA box motif provides enough information for TBP binding onto TATA box-
Figure 2.8: Correlations between the mean transcript production rate of non-periodic and periodic genes grouped into TATA box-containing and TATA-less with core promoter occupancies of general transcription factors involved in PIC formation, which are from Rhee and Pugh (2012). (Adapted from Eser et al. (2014).)

containing promoters (Kamenova et al., 2014); b) to allow the PIC components being recruited to the DNA, core promoters typically contain a NFR around TSS, and thereby positioning of the −1 and +1 nucleosomes could be major obstacles for DNA access to the general transcription factors at core promoters (Field et al., 2008; Kaplan et al., 2009; Bai et al., 2010; Bai and Morozov, 2010; Venters et al., 2011; Raveh-Sadka et al., 2012); c) +1 nucleosomes might function to limit TSS selection by impeding RNAPII scanning at TATA-less promoters (Rhee and Pugh, 2012); d) +1 nucleosomes could be high barriers to RNAPII elongation (Weber et al., 2014; Bai and Morozov, 2010).

In section 3.2, we will exploit k-mer based sequence features from Zeevi et al. (2014) and Lubliner et al. (2015) to model the core promoter regulatory elements, and represent the hypothesis that core promoter amplifies transcript production rates,
possibly by influencing RNAPII recruitment, PIC formation, and transcriptional initiation. Moreover, we will design TBP binding based features and nucleosome occupancy based features to test the hypothesis that the regulatory role of the core promoter could be explained by the mechanisms about TATA elements and nucleosomes within core promoter.

2.4 Hypothesized mechanisms relevant to the transcriptional synergy between TFs and core promoter regulatory elements

As discussed in sections 1.4, 2.1, and 2.2, TFs regulate transcription by recruiting transcription cofactors through direct protein-protein interactions, and ultimately influence PIC formation and RNAPII recruitment. As discussed in sections 1.3, 1.4, and 2.3, core promoter regulatory elements influence PIC formation, RNAPII recruitment, DNA melting, RNAPII searching for and selecting TSS, releasing RNAPII from general transcription factors and thereby initiation, and thereby transcription initiation. Therefore, TFs and core promoter regulatory elements work together to regulate transcription initiation. In general, where several factors (elements) work together to enhance a reaction rate, the joint effect is not merely the sum of the respective contributions of each factor, but the product (Alberts et al., 2014). Based on these, we propose that there is transcriptional synergy between TFs and core promoter regulatory elements. This hypothesized mechanism about transcriptional synergy is consistent with recent literatures, though none of them propose such a synergy. We discuss the relevant papers as follows.

Recently, Eser et al. (2014) observed that some cell cycle regulators share similar shape of time course with their respective target genes (See Figure 2.5). They observed such kind of match for TFs Swi4p, Yox1p, and Mbp1p. Moreover, the authors observed that PIC components measurements from Rhee and Pugh (2012) are correlated with time average transcript production rates (See Figure 2.8). There-
fore, they proposed that TFs control the timing of transcript production, while the core promoter governs the magnitude. However, Eser et al. (2014) did not provide a computational model. Also, measurements of PIC components are only available from experiments: we are able to infer them from neither DNA sequence nor occupancy profiles. Lastly, we are still not sure that whether there is only additive function between TFs and the core promoter, or there are both additive function and synergy, or there is only synergy. Therefore, testing the existence of transcriptional synergy would be helpful for determining relationship between TFs and core promoter elements.

Moreover, several literatures have studied the mechanisms that TFs, especially activators, regulate transcription through TFIID pathway and SAGA pathway. Note that TFIID and SAGA have a set of common TAFs as subunits (Huisinga and Pugh, 2004; Bhaumik, 2011; Venters et al., 2011), which might have DNA binding specificities. As shown in Figure 2.9, though these two coactivators have additional activities, they both play key roles in activator-dependent recruitment of TBP (reviewed in Kamienova et al., 2014).

In contrast to the *in vitro* assembly pathway of general transcription factors (See section 1.3 and Figure 1.3), an oddity of TBP *in vivo* is that, when it is part of the TFIID complex, it tends to bind promoters that lack the TATA box consensus TATAWAWR, where W = A or T; R = A or G (Basehoar et al., 2004); when TBP is not part of the TFIID complex, the SAGA complex directs TBP to TATA box-containing RNAPII core promoters (Dudley et al., 1999; Bhaumik and Green, 2002; Mohibullah and Hahn, 2008). Genome-wide studies found that 85%-90% of the yeast genome can be generally characterized as being TFIID dominated (Basehoar et al., 2004; Huisinga and Pugh, 2004). Such TFIID-dominated genes are typically TATA-less and utilize TFIID to assemble the PIC. The remaining 10%-15% are SAGA dominated. Such SAGA-dominated genes typically contain a TATA box and
Figure 2.9: Illustration of SAGA-dependent and TFIID-dependent transcriptional activation. Subsequent to TBP recruitment, PIC forms at core promoter, RNAPII is recruited to core promoter, and transcription is initiated. (A) The activator that binds to the UAS of the core promoter help to assemble PIC at the core promoter through interaction with one or more components of the transcriptional machinery, termed as “target”. (B) Such activator-target interactions could be achieved by firstly activator interacting with SAGA, and then SAGA regulate TBP recruitment via Mediator. (C) Activator can directly interact with both SAGA and Mediator, which ultimately both contribute to TBP recruitment. (D) Activator can also interact with TAF subunits of TFIID, and thereby control TBP recruitment. (Adapted from Bhaumik (2011).)

preferentially exploit SAGA to assemble the PIC. SAGA-dominated genes tend to be environmental stress-induced genes that are more highly regulated. The environmental stress include heat, DNA damage and metabolic starvation (Huisinga and Pugh, 2004). TFIID-dominated genes tend to be housekeeping genes, many of which are downregulated in response to environmental stress. Because a gene may be involved to varying degrees in both stress and housekeeping responses, it may use differing blends of the two pathways, rather than one or the other exclusively (Venters et al., 2004).
To identify TBP-binding motif at TATA-less core promoters, Rhee and Pugh (2012) searched for sequence elements with up to two mismatches to the TATAWAWR consensus, limiting search to measured PIC locations. Remarkably, 99% of the TATA-less core promoters contained a sequence having two or less mismatches to the TATA box consensus. The authors referred to these mismatched elements as “TATA-like”, as they did not form a consensus, whereas those conforming to the consensus retain the “TATA box” designation. Rhee and Pugh (2012) referred to the two elements together as “TATA elements”.

However, Kamenova et al. (2014) reported that mutations of the TBP DNA binding surface reduce transcription of TATA box-containing promoters, not TATA-less promoters; and probably it is TBP, not an intact DNA binding surface of TBP, that is required for transcription. Their results are consistent with a hypothesis that the sequence-specific DNA binding activity of TBP is not important for transcription from TFIID-dominated TATA-less promoters (Lee and Struhl, 1995; Arndt et al., 1995). As reviewed in (Green, 2000; Bhaumik, 2011), it is probably the core promoter elements are responsible for the TFIID-dependency. Even in yeast, TFIID-dependency of some genes has been mapped to the core promoter elements (Shen and Green, 1997). For SAGA-dependency, Venters et al. (2011) provided a theory: probably the TATA box provides added TBP-promoter stability against encroaching nucleosomes, while at the same time making TBP accessible to its many negative regulators such as Mot1p, which favors the SAGA pathway. In the SAGA pathway, TBP is stabilized against such negative regulators by TFIIA (reviewed in Lee and Young, 1998; Pugh, 2000).

To recruit SAGA and TFIID, activators directly interact with respective subunits of SAGA and TFIID. An essential component of SAGA, Tra1, has been implicated as the target of several activators in yeast (Brown et al., 2001; Bhaumik et al., 2004;
Green, 2005). Also, Mencía et al. (2002) have shown that the Rap1p, which is associated with the UASs of the yeast TFIID-dependent ribosomal protein genes, is sufficient to recruit TAFs (TFIID) to both TFIID dependent and independent core promoters.

In summary, as shown in Figure 2.9, TFs can directly interact with mediator, SAGA, and TAFs, and ultimately interact with RNAPII, to regulate transcription. Also, generally the functioning mechanism of TFs would require the existence of corresponding elements in core promoter. Therefore, we propose that TFs and core promoter elements always work together to regulate transcript production rate, and thereby there is transcriptional synergy between them.

In chapter 3, we will model the transcriptional synergy between TFs and core promoter regulatory elements, by introducing interaction terms between TF relevant features and core promoter relevant features.

2.5 Hypothesized mechanisms relevant to the nucleosomes resisting RNAPII elongation

During elongation, RNAPII always confronts barriers to its progression, including nucleosomes and other polymerases. Recently, researchers have observed that nucleosomes are barriers to RNAPII elongation (Jiang and Pugh, 2009; Weber et al., 2014; Jordan-Pla et al., 2015). Specifically, +1 nucleosomes (relative to TSS) could be high barriers, while gene body nucleosomes are low barriers and their function as barriers probably decrease from 5’ to 3’ (Weber et al., 2014). Also, it is likely that +1 nucleosomes regulate RNAPII elongation in a H2A.Z-dependent way (Jordan-Pla et al., 2015). Confronting nucleosomes as barriers, RNAPII may be backtracking, stalling, and even being arrested; and arrested RNAPII usually requires TFIIIS-stimulated transcript cleavage for reactivation (Saeki and Svejstrup, 2009). But RNAPII backtracking does not always result in stalling: when the leading RNAPII backtracks,
the trailing RNAPII would be most likely in the active mode, pushing the leading RNAPII forward (Saeki and Svejstrup, 2009; Nudler, 2012). With such a “group action” mechanism, in the highly expressed gene, the pushing by the polymerase behind might make leading RNAPII conquer nucleosomes as obstacles. However, in the lowly expressed gene, TFIIS-stimulated transcript cleavage for reactivation is still required.

We are not able to model mechanisms relevant to H2A.Z and TFIIS, because we have limited knowledge about them. Fortunately, COMPETE is able to infer nucleosome occupancy in gene body. In section 3.3, we will be able to model that +1 nucleosomes are high barriers, that gene body nucleosomes are low barriers and their function as barriers probably decrease from 5’ to 3’ , and that there is “group action” between leading RNAPII and trailing RNAPII.
Our feature engineering protocol: extracting and hierarchically grouping features

The main idea of our work is: a) model various mechanistic hypotheses by designing and extracting certain groups of features; b) provide support for certain hypotheses by interpreting the selected groups of features. In this chapter, we mainly introduce how we design our feature engineering protocol according to the hypothesized mechanisms we discussed in chapter 2. With these features, we will introduce our machine learning framework, which selects (sub)groups of features so that provide support for associated hypotheses, in chapter 4.

We design three groups of features: the “fluctuation” group for the TFs’ roles, the “amplifier” group for the core promoter’s role, and the “resistance” group for the gene body nucleosomes’ roles. For each group, we design subgroups of features to reflect the multiple mechanisms for the associated regulatory factors and elements. Moreover, we add “fluctuation × amplifier” interaction terms between each feature in the “fluctuation” group and each feature in the “amplifier” group. These interaction terms represent the transcriptional synergy between TFs and core pro-
moter regulatory elements. Similarly, we add “fluctuation × resistance” interaction terms, “amplifier × resistance” interaction terms, as well as “fluctuation × amplifier × resistance” interaction terms. Such interaction terms model the assumption that nucleosomes within gene body (ORF region) modulate transcript production via regulation of RNAPII elongation. Especially, they represent the assumption that collisions between leading and trailing RNAPII themselves are helpful for conquering nucleosomes as barriers, and thereby make highly expressed genes be less influenced by nucleosomes. Without such a “group action” assumption, interactions involving “resistance” features would not be needed, because “resistance” features as main effects would be enough.

In the following sections, we will describe how we design and extract features for each group of main effects. Especially, each group contains subgroups of features. These subgroups represent alternative hypotheses for corresponding group of main effects. Then we summarize the hierarchical grouping structure of our features. And we introduce how we preprocess feature values.

3.1 “Fluctuation” group of TF relevant features

The “fluctuation” group is for the TFs’ roles. As discussed in 2.2, we do not explicitly model a TF to be activator or repressor: we expect the sign of a relevant feature reflect the TF’s role with the associated hypothesized mechanism. However, we take the hypothesis that different TF’s contribution to transcript production might be distinct and non-redundant from each other, by allowing each TF have its own associated features.

Firstly, we design two basal features for each TF: expected number of TF binding instances and probability of at least one TF binding instance. Secondly, we propose two weighting functions: being weighted by distance with respect to TSS; and being weighted by nucleosome occupancy nearby. In combination, there are 8 features for
each TF. Since TFs are the major group of driving forces of fluctuations in promoter outputs, we assign these 8 subgroups of TF relevant features into the “fluctuation” group:

1. “ENTF”: expected number of TF binding instance;
2. “ENTFD”: expected number of distance decay weighted TF binding instance;
3. “ENTFS”: expected number of stable TF binding instance;
4. “ENTFDS”: expected number of distance decay weighted and stable TF binding instance;
5. “POTF”: probability of at least one TF binding instance;
6. “POTFD”: distance decay weighted probability of at least one TF binding instance;
7. “POTFS”: probability of at least one stable TF binding instance;
8. “POTFDS”: distance decay weighted probability of at least one stable TF binding instance.

If we use the TrSS as origin, the DNA sequence as a 1 dimensional coordinate axis, and from TrSS to 5’ of the promoter as positive direction, we will have $i$ as the index of a genomic location in a promoter region, relative to its downstream gene’s TrSS.

Also, denote $P(i, tf)$ as the probability of a transcription factor $tf$ starting binding at location $i$, and $P(i, nuc)$ as the probability of a nucleosome starting binding at location $i$.

We have the nucleosome occupancy at location $i$, $OP(i, nuc)$, as,

$$OP(i, nuc) = \sum_{j=i-147}^{i} P(j, nuc), \quad (3.1)$$
where the 147 is the nucleosome motif length. Please note Equation 3.1 also applies to other DBFs, though they might have motif length different from 147. From this equation, we know that the occupancy of a DBF is the sum of probabilities that it starts binding at all valid locations. Please note we sum from \( i - 147 \) to \( i \), simply because COMPETE calculates probabilities from 5' to 3'. Also, please note the occupancy is a posterior probability: it is within \([0, 1] \). 

Denote \( l_{pro} \) as the length of promoter. In this work, we chose \( l_{pro} = 600 \), and have tried 800, 500, and some other smaller values. Basically 600 bp as native promoter length provides the best results, while 800 bp is slightly worse and other values is worse. We define the expected number of TF binding instance (ENTF) as,

\[
E[N_{tf}] = \sum_{i=0}^{l_{pro}} P(i, tf),
\]

and the probability of at least one TF binding instance (POTF) as,

\[
P(one + tf) = 1 - \prod_{i=0}^{l_{pro}} (1 - P(i, tf)).
\]

Please note \( P(one + tf) \) is a saturated version of \( E[N_{tf}] \),

\[
P(one + tf) = 1 - \prod_{i=0}^{l_{pro}} (1 - P(i, tf))
\]

\[
= \sum_{i=0}^{l_{pro}} P(i, tf) - \sum_{i=0}^{l_{pro}} \sum_{j=i+1}^{l_{pro}} P(i, tf) \times P(j, tf) + \sum_{i=0}^{l_{pro}} \sum_{j=i+1}^{l_{pro}} \sum_{k=j+1}^{l_{pro}} P(i, tf) \times P(j, tf) \times P(k, tf) - \ldots
\]

\[
= E[N_{tf}] - \sum_{i=0}^{l_{pro}} \sum_{j=i+1}^{l_{pro}} P(i, tf) \times P(j, tf)
\]

\[
+ \sum_{i=0}^{l_{pro}} \sum_{j=i+1}^{l_{pro}} \sum_{k=j+1}^{l_{pro}} P(i, tf) \times P(j, tf) \times P(k, tf) - \ldots
\]

59
3.1.1 Distance decay of TF binding instance relative to TSS

Denote \( l_{\text{tss}} \) as TSS’s relative location to TrSS. When TSS info is not available in Rhee and Pugh (2012), we use \(-56\) relative to TrSS as TSS. Inspired by observations in MacIsaac et al. (2010) and Sharon et al. (2012), we define the distance decay weight for TF binding instance as,

\[
w^d(i) = e^{-\frac{l_{\text{tss}}}{w_s}}, \forall 0 \leq i \leq l_{\text{pro}}.
\]  

(3.5)

We chose the exponential decay function because it is a common distance decay function, and it nicely decreases from 1 to 0 as the distance goes from 0 to infinity. We illustrate our distance decay weight in Figure 3.2. Also, in the figure, we used the TSS location \(-42\) (relative to TrSS) as an example. But as shown in Figure 3.1, we miss TSS info for a significant fraction of genes, and sometimes have TSS downstream of TrSS.

3.1.2 Stability weight of TF binding instance determined by nucleosome occupancy nearby

Inspired by Bai et al. (2010) and Lickwar et al. (2012), we define the stability weight, which is determined by the average nucleosome occupancy nearby, as,

\[
w^s(i) = 1 - \frac{1}{2} \times \frac{w_s}{w_s} \times \sum_{j=-w_s+\frac{3}{2}l_{\text{motif}}}^{w_s+\frac{3}{2}l_{\text{motif}}} OP(i + j, nuc), \forall 0 \leq i \leq l_{\text{pro}},
\]  

(3.6)

where \( l_{\text{motif}} \) is the motif length of \( tf \). And the window, \([-w_s, w_s]\), of motif center is the window we used for calculating average nucleosome occupancy nearby. We arbitrarily chosen \( w_s = 30\) based on Figures in Lickwar et al. (2012). But \( w_s \) could be set as any reasonable integer by users of our software. We demonstrated our stability weight in Figure 3.2.

Using \( d \) for distance and \( s \) for stability, we define the expected number of distance
Figure 3.1: TSS location information with 346 genes subset of Eser et al. (2014). (A) The histogram of TSS’s location relative to TrSS, given data from Rhee and Pugh (2012). TSS are widely distributed, and are condensed within [-118, -1] of TrSS. The kernel density estimation has a mode as -56. (B) With [-118, -1] of TrSS as core promoter, 194 genes have TSSs located within the core promoter, and 77 genes have TSS out of the core promoter. Other 75 genes miss TSS information in Rhee and Pugh (2012).
Figure 3.2: In this work, we chose $d_u = 250$. In other words, we have 250bp as our decay unit, and consequently $\frac{1}{250}$ as our decay constant. With this value, we have the weight decays in an appropriate rate, and in a linear form within the region close to TSS. But $d_u$ could be set as any reasonable integer by users of our software. The stability weight for TF binding instance is determined by average nucleosome nearby. In this work, we used the $[-30, 30]$ window of motif center.

decay weighted and stable binding instance of transcription factor $t_f$ (ENTFDS) as

$$E[N_{t_f}^{ds}] = \sum_{i=0}^{l_{pro}} P(i, t_f) \times w^d(i) \times w^s(i), \quad (3.7)$$

and the probability of at least one distance decay weighted and stable binding instance of transcription factor $t_f$ (POTFDS) as

$$P(\text{one}+d_{t_f}^{ds}) = 1 - \prod_{i=0}^{l_{pro}} \left(1 - P(i, t_f) \times w^d(i) \times w^s(i)\right), \quad (3.8)$$

Similar to Equations 3.7 and 3.8, there are other weighted versions of expected number of binding instances features and at least one binding instance probability features.

An important issue about these subgroups of features are pairwise strong correlations. Basically all subgroups other than “ENTF” are its variants by applying saturation or weighting. So this is an expected issue. We demonstrated this issue in
Figures 6.5, 6.6. Moreover, though they are highly correlated to each other, we believe that, if the mechanism represented by saturation or weighting is relevant, there should be additional information contained in the corresponding subgroup of features and, thereby the subgroup would be selected for better predictive performance.

Lastly, there are corresponding orientation version subgroups for each of these 8 subgroups, by considering the orientation of TF binding instance. Basically, for each of the 8 features above, there will be a feature value for forward TF binding instance, and a feature value for reverse TF binding instance. Their sum would be equal to the original feature value for TF binding instance. But such an orientation version would be helpful for predictive performance, if the orientation is important for TF’s functioning on transcriptional regulation.

3.2 “Amplifier” group of core promoter relevant features

Because there is limited knowledge about DNA specificities of general transcription factors binding to core promoter in *S. cerevisiae* Yeast, we decided to represent the function of core promoter with sequence features, TATA binding protein occupancy at core promoter, and nucleosome occupancy at core promoter. Specifically, for sequence features we resorted to the *k*-mer count features in Zeevi et al. (2014) and Lubliner et al. (2015); for occupancy features, we designed two groups of features based on the definition of core promoter.

3.2.1 *k*-mer count features representing regulatory elements within the core promoter

As for the 38 *k*-mer based features given in Zeevi et al. (2014), we exclude features 8, 14, 15, 22, 29, 30, 31, 32, 33, 37, and 38 in our work. These eleven *k*-mer counts and sequence content features are about TF binding, TBP binding, and nucleosome occupancy. The reasons that we excluded them are: a) they are fixed over time; b) we
assume that TF relevant features extracted from COMPETE dynamic occupancy
profiles are better generally. Then the left 27 features are named as “k-mer count
based sequence features in conservation study of RP genes” (“kCCR”). Details of
these features are included in Figure A.1.

The Lubliner et al. (2015) defined $[-118, -1]$ of TrSS as core promoter, and
reported 64 k-mer count features extracted from the core promoter region (They
indexed to 65, but missed the 60th, probably due to reindexing.). We use all of
them as “k-mer count based sequence features in core promoter region” (“kCCP”)
subgroup. We believe “kCCP” is a refined version of “kCCR”, since they are both
from the Segal group. Details of these “kCCP” features are included in Figure A.2.

Zeevi et al. (2014) and Lubliner et al. (2015) had tried to categorize their re-
spective sets of core promoter sequence features. Specifically, Lubliner et al. (2015)
manually partitioned their defined core promoter into the TATA element region, the
PIC region, the scanning region, and the initiation region, and then classified most
of the features into classes determined by the feature’s location and G/C content.
However, such a partitioning is based on their manually annotation of subsequences
within native $[-118, -1]$ promoter regions of seven genes (RPL3, RPL28, RPL25,
RPL4B, GAL7, HSP12, and RPB10). Given that, as shown in Figure 3.1, we have
TSSs located out of the core promoter for a lot of genes, such a partitioning and,
thereby the classification of sequence features, might not be applicable for interpret-
ing these features for all the genes.

3.2.2 occupancy features representing regulatory elements within the core promoter

To see whether there could be better features representing clearer mechanisms, we
design another two subgroup features and extracted their values from the COMPETE
outputs. We call them the “occupancy profile based features in core promoter region”
(“OPCP”) subgroup and “occupancy profile based features nearby TSS” ("OPNT")
subgroup. The hypothesized mechanisms are: a) TATA box motif provides enough information for TATA binding; b) To allow the PIC being recruited to the DNA and scanning for TSS, core promoters typically contain a NFR (Field et al., 2008; Kaplan et al., 2009; Bai and Morozov, 2010), and if there is a nucleosome within NFR, transcription would be harmed; c) +1 nucleosomes (relative to TSS) could be high barriers to RNAPII elongation (Weber et al., 2014; Bai and Morozov, 2010).

Please note the NFR region is between −1 nucleosome and TSS. So the pivot should be TSS. However, as shown in Figure 3.1, we miss TSS info for a significant fraction of genes, and have TSSs located out of the core promoter, which is as [−118, −1] of TrSS, for some genes. With these info in mind, we designed 3 features for the “OPCP” subgroup:

1. “expected number of TBP instance within core promoter region” ("ENTCP"),
2. “+1 nucleosome occupancy within core promoter region” ("PONOCP"),
3. “−1 nucleosome occupancy within core promoter region” ("MONOCP"),

where we follow Lubliner et al. (2015) to set [−118, −1] of TrSS as core promoter, and follow Zhong et al. (2014) to model TATA binding protein (TBP). Therefore, “ENTCP” is the expected number of TBP binding instance within the [−118, −1] of TrSS. Denote \( l_{tss} \) as TSS’s relative location to TrSS. “PONOCP” is the average nucleosome occupancy within the \([l_{tss}, -1]\) of TrSS, and “MONOCP” is the average nucleosome occupancy within \([-118, l_{tss}]\). If TSS is out of the core promoter, we force it to be at the closes core promoter end.

And there are 3 features for the “OPNT” correspondingly,

1. “expected number of TBP instance nearby TSS” ("ENTNT"),
2. “+1 nucleosome occupancy nearby TSS” ("PONONT"),
3. “−1 nucleosome occupancy nearby TSS” (“MONONT”),

where we follow Lubliner et al. (2013) to set \([l_{tss} - 75, l_{tss} + 50]\) as the region nearby TSS for extracting these feature values. Though we use the upstream length 75 bp and the downstream length 50 bp, they could be set by users of our software. Therefore, “ENTNT” is the expected number of TBP binding instance within the \([l_{tss} - 75, l_{tss} + 50]\) region. “PONONT” is the average nucleosome occupancy within the \([l_{tss}, l_{tss} + 50]\) region, and “MONONT” is the average nucleosome occupancy within the \([l_{tss} - 75, l_{tss}]\) region.

The “OPCP” features are designed for a direct comparison with sequence features extracted from core promoter region. If these three features could provide same amount of information as the “kCCP” subgroup, then we could claim most of the sequence features could be explained by functioning of TATA elements and nucleosomes within the \([-118, -1]\) region of TrSS. However, because they are about occupancy profiles of the exact same promoter region as “kCCP”, they still have the same limitation: though the \([-118, -1]\) promoter region is reasonable as core promoter, we did have some TSSs that are upstream of the core promoter, making the interpretation of k-mer features less meaningful. Therefore, the “OPNT” features are designed for tackling this limitation. However, we have to admit that this subgroup has some limitations as well: a) we need to choose upstream and downstream boundaries for defining the region nearby TSS, which is just our own version of core promoter; b) some TSSs are even downstream of TrSS in our data, even though we have taken care of reference genome version in different data sources; c) some transcripts have alternative TSSs. An example for limitation b) is gene PRP2 (YNR011C). It is one of the genes that have TSS downstream of TrSS, as shown in Figure 3.1. For PRP2 (YNR011C), we have TSS location as 646938 on chromosome XIV, while we have the gene location as 644322 to 646952. Please note PRP2 is “−”
Because more than half of the genes we studied have TSS within the \([-118, -1]\) region of TrSS, and we use the upstream length 75 bp and the downstream length 50 bp of TSS for defining region nearby TSS, we should expect that these two subgroups of features have strong pairwise correlation, as shown in Figure 3.3.

Moreover, we try to check whether our features in “OPCP” subgroup are correlated with certain classes of sequence features in “kCCP” subgroup according to the categorization in Lubliner et al. (2015).

Because our “ENTCP” and those \(k\)-mer counts are all based on TATA box, we did see a relative strong correlation among them, as shown in Figure 3.4 and Table B.1. The feature name “1_ATAAA_counts_m118_2_m61-.Consensus_TATA” means that this feature is indexed as 1 in Lubliner et al. (2015), about ATAAA counts, extracted from the \([-118, -61]\) region, and is classified as consensus TATA in Lubliner et al. (2015). And for simplicity, we will call it feature 1 in “kCCP”. Other feature names could be explained in the same way.

Since the initiation region is annotated as where the TSSs are located, we check the correlation coefficient between “PONOCP” and those initiation \(k\)-mer counts in “kCCP” subgroup. Hypothetically, those \(k\)-mer features in initiation class could represent the functioning of the +1 nucleosome. However, as shown in Figure 3.5, they are weakly correlated. Moreover, as given in in Table B.2, though the features 10, 25, and 44 have maximum pearson correlation coefficient absolute value around 0.1, the respective signs are opposite to what expected. According to their effects in Lubliner et al. (2015), features 10 and 25 are expected to positive effects transcription, while feature 44 is expected to have negative effects on transcription. Because “PONOCP” is also expected to have negative effects on transcription, we expect it to be positively correlated with feature 44, and negatively correlated with features 10 and 25. Therefore, it could be either that these three features are weakly correlated
Figure 3.3: There are strong correlations between respectively associated features in “OPCP” and the “OPNT” subgroups. The correlation matrix plot for respective features within the “OPCP” and the “OPNT” subgroups. The scatter matrix plot for respective features within the “OPCP” and the “OPNT” subgroups.
Figure 3.4: We observed that our “ENTCP” feature has Pearson $r$ correlation coefficient in the range of 0.238 to 0.352 with those TATA element $k$-mer counts in “kCCP” subgroup. Specific values are given in Table B.1.

with “PONOCP” only by chance, or that these initiation $k$-mer counts reflect some mechanisms other than +1 nucleosome as barrier to elongation.

Moreover because the TATA element region is annotated as where containing strong or weak TATA elements, the PIC region is annotated as where PIC formed and initially unwinds the DNA, and the scanning region is annotated as where RNAPII is expected to pass when it scans for possible TSSs, we checked the correlation between “MONOCP” and the $k$-mer counts in those classes within “kCCP” subgroup. Hypothetically, those $k$-mer features in TATA element, PIC, and scanning class could represent the functioning of the $-1$ nucleosome. As demonstrated in Figure 3.6, most of them are not correlated. However, as listed in Table B.3, features 1, 2, 3, and 9 are negatively correlated with “MONOCP”, with $r$ from $-0.154$ to $-0.272$. Given that features 1, 2, and 3 are TATA element features, we did expect TATA
Figure 3.5: We observed that our “PONOCP” has Pearson $r$ correlation coefficient in the range of 0.0002 to 0.1001 with those initiation $k$-mer counts in “$kCCP$” subgroup. Specific values are given in Table B.2.

box would produce lower $-1$ nucleosome occupancy within the core promoter, since COMPETE is capable to model the competition between TBP and nucleosome for binding at same DNA sequence location. Moreover, feature 9 is in PIC class, and is A\T-rich. Lubliner et al. (2015) reported that it leads to higher predicted expression. Therefore, feature 9 is opposite to our “MONOCP” feature with regarding to transcriptional regulation, since we assume more $-1$ nucleosome in the core promoter region would make less RNAPII recruited and, thereby less PIC formed. This makes the negative correlation reasonable. Given these relatively strong correlations, it is possible that $k$-mer features in TATA element, PIC, and scanning classes, more or less reflect the regulatory role of $-1$ nucleosome.

Lastly, Lubliner et al. (2015) reported some $k$-mer features could not be classified. So we checked whether they could be explained by $+1$ or $-1$ nucleosome occupancy
Figure 3.6: We observed that our “MONOCP” has *Pearson r* correlation coefficient in the range of -0.272 to 0.075 with the k-mer counts in TATA element, PIC, and scanning classes of “kCCP” subgroup. Within core promoter. As shown in Figures 3.7 and 3.8, as well as Tables B.4 and B.5, most features are not correlated with our nucleosome features. The features 47, 45, 41 have *r* ≈ −0.1 with our “PONOCP”. However, they are reported to have negative effects on transcription. The correlation between these k-mer features and our nucleosome feature should be positive, if they represent +1 and/or −1 nucleosome’s regulatory role. The feature 65 has 0.160 correlation coefficient with our “PONOCP”. This is a correlation with expected sign, given that the feature 65 is reported to lead to lower predicted expression. Similar observations are obtained for our “MONOCP” feature.
3.3 “Resistance” group of gene body nucleosome features

Denote $i$ as the distance of a location within ORF to 5' boundary of ORF (TrSS) and $l_{ORF}$ as the length of ORF region. We define the “expected number of nucleosomes within gene body” (“ENNGB”) as,

$$E(N)_{nuc} = \sum_{i=0}^{l_{ORF}} p_{nuc}(i),$$

(3.9)

where $l_{ORF}$ is the length of ORF.

With reason similar to that for Equation 3.5, we use exponential decay function and $d_u = 250$ as distance decay unit for defining the “expected number of distance
Figure 3.8: We observed that our “MONOCP” has Pearson $r$ correlation coefficient in the range of -0.162 to 0.112 with the $k$-mer having no class label in “kCCP” subgroup.

decay weighted nucleosomes within gene body” (“ENNGBD”),

$$E(N)_{\text{nuc}}^d = \sum_{i=0}^{l_{\text{ORF}}} p_{\text{nuc}}(i) \times e^{-\frac{i-l_{\text{tss}}}{d_u}}$$,  \hspace{1cm} (3.10)

where $d$ indicates it is a distance decay weighted version, the $l_{\text{tss}}$ is the TSS’s location to 5’ boundary of ORF (TrSS). We assume the resistance effects of a gene body nucleosome decrease as it moves away from TSS. For those native promoters without TSS information available in Rhee and Pugh (2012), we use the TrSS as the TSS.

3.4 Grouping of main effect features

We grouped the extracted features bottom-up and listed the grouping as below. Please note there are associated interaction groups and subgroups, which are de-
scribed in Table 4.1.

1. To represent TFs’ roles in transcriptional regulation, we extracted “fluctuation” group of TF relevant features consisting of 8 subgroups (We do not list the orientation version of these 8 subgroups here.):
   
   (a) “ENTF”,
   (b) “ENTFD”,
   (c) “ENTFS”,
   (d) “ENTFDS”,
   (e) “POTF”,
   (f) “POTFD”,
   (g) “POTFS”,
   (h) “POTFDS”.

2. To reflect the core promoter’s function in transcriptional regulation, we used “amplifier” group of core promoter relevant features, consisting of 4 subgroups respectively:
   
   (a) “kCCR”,
   (b) “kCCP”,
   (c) “OPCP”,
   (d) “OPNT”.

3. To study the nucleosome’s influence on transcription within gene body, we extracted “resistance” group of gene body nucleosomes features, consisting of 2 single feature subgroups:
3.5 Preprocessing feature values

Our feature values are all based on respective counts of \( k \)-mer and DBF binding instance. Therefore, for all data sets, we transform the feature values by taking square roots to deal with skewness. Moreover, sparse linear models, like Lasso, usually suggest standardizing each feature to have mean zero and unit norm, since regularization depends on the scaling of the input features. Therefore, after taking square roots for input features, we standardize each of them.
In chapter 3, we introduced how we design and extract features according to the hypothesized mechanisms that we discussed in chapter 2. In this chapter, we present a machine learning framework that learns sparse linear models for the transcriptional regulation.

One of our main ideas is providing support for certain hypotheses by interpreting the selected (sub)groups of features in our learned models. This idea requires a modeling framework with not only good predictive performance, but also strong interpretability. Moreover, we have large number of features, especially because we have interactions among main effects. Therefore, we decide to exploit sparse linear models.

Our framework takes as input position-specific scoring matrices (PSSMs) for a set of DBFs, concentrations for the set of DBFs, and DNA sequences of promoter region, gene body and flanking sequences on either side, and learns models for predicting the mean or dynamic transcript production rate that any promoter DNA sequence will
produce when receiving the DBF concentrations. Then, given our needs, we discuss the advantages and disadvantages of respective sparse linear models, especially group-wise sparse linear models. And finally we introduce our algorithm for this complicate problem: in a cross-validation approach, we perform gRFE combining the ridge in a sequence of steps to promote subgroup-wise and group-wise sparsity. We have implemented this framework in our PreTrans software package in Python.

4.1 The components of PreTrans

Our framework PreTrans has three main components. The first component, given inputs, generates features representing certain hypotheses, where these features are hierarchically grouped because of our domain knowledge. We have introduced our feature engineering protocol in Chapter 3. From the perspective of implementation, this feature engineering component has three parts. The first part extracts features from the occupancy profile. The occupancy profile is produced by COMPETE Wassen and Hartemink (2009) , a software package implementing a model of continuous DNA binding that considers competition by diverse DBFs at given concentrations, and capable of doing so at genome-wide scale and single-nucleotide resolution. The second one extracts $k$-mer count features from promoter DNA sequences, according to reported features in Zeevi et al. (2014) and Lubliner et al. (2015). The third one adds interaction terms among different groups of features.

The second component of PreTrans, in a cross-validated approach, selects helpful features in a procedure, proposed by us, that combines ridge and a group version of recursive feature elimination (Hoerl and Kennard, 1970; Friedman et al., 2010; Guyon et al., 2002). PreTrans’ third component exploits the selected features to predict dynamic or mean transcript production rates with the best model selected by comparing fitted lasso, ridge, and elastic net models in terms of predictive performance (Hoerl and Kennard, 1970; Tibshirani, 1996; Zou and Hastie, 2005; Friedman
et al., 2010). The second and third components of our framework are programmed with Python, utilizing the scikit-learn package (Pedregosa et al., 2011).

4.2 Linear models representing hypothesized mechanisms

We hypothesized that TFs control the fluctuation of RNAPII recruitment and transcriptional initiation, core promoter amplifies such fluctuation, and gene body nucleosomes resist RNAPII elongation. Correspondingly, we designed three groups of features for main effects: “fluctuation” for the TFs’ roles, “amplifier” for the core promoter’s role, and “resistance” for the gene body nucleosome’s role. Moreover, we added four groups of features representing interactions among main effects: “fluctuation \times amplifier”, “fluctuation \times resistance”, “amplifier \times resistance”, and “fluctuation \times amplifier \times resistance”. Given that we face a regression problem, we exploited linear models to predict transcript production rates, because of the excellent scalability and interpretability of linear models.

As shown in Tables 4.1 and 4.2, we explained the associated hypotheses for respective groups of main effects and interactions, as well as for certain linear combinations of these groups. Although we did not enumerate all possible linear combinations, the underlying hypotheses could be inferred in a way similar to those in the table.

Our full linear regression model has the form

\[
y = \beta_0 + X^{(f)}\beta^{(f)} + X^{(a)}\beta^{(a)} + X^{(r)}\beta^{(r)} + X^{(f:a)}\beta^{(f:a)} + X^{(f:r)}\beta^{(f:r)} + X^{(a:r)}\beta^{(a:r)} + X^{(f:a:r)}\beta^{(f:a:r)} + \epsilon, \tag{4.1}
\]

where \( y = (y_1, y_2, \ldots, y_n)^T \) is the target: \( n \) transcript production rates, \( X \) is the \( n \times p \) feature matrix:

\[
X = (X^{(f)}, X^{(a)}, X^{(r)}, X^{(f:a)}, X^{(f:r)}, X^{(a:r)}, X^{(f:a:r)}), \tag{4.2}
\]
\( f \) indicates “fluctuation” group, \( a \) indicates “amplifier” group, \( r \) indicates “resistance” group, and : indicates pairwise interactions: \( X^{(f:a)} = X^{(f)} \times X^{(a)} \). \( \beta_0 \) is intercept, the other \( \beta \)'s are associated coefficient vectors for feature submatrices of different groups, and \( \epsilon \) is a vector of independent and identically distributed errors with mean 0 and variance \( \sigma^2 \).

With our problem-specific hypotheses, which we believe can lead to greater prediction accuracy, we have predefined groups of main effects. And for each group of main effects, subgroups of features, which reflect alternative hypotheses, compete for representing the certain group (See corresponding sections in chapter 3.). Because of our sets of alternative hypotheses, we know that there are highly correlated features across subgroups and groups. For example, across different subgroup within “fluctuation” group, we know that associated features for the same TF would be highly correlated, since they are all extracted from the same occupancy profiles, as shown in Figures 6.5, 6.6. Moreover, complicating the matter is the hierarchy introduced by subgroups contained in groups, and interactions between main effects.

Therefore, in our situation, we know that some subgroups of features will be discarded in our fitted linear model. It makes sparse linear models, especially group-wise sparse linear models, be a natural choice. Moreover, we have another reason for exploiting sparse linear models: the sparsity would advance the predictive power of fitted models by providing good generalization. Like many other recent applications of machine learning, our problems involve large numbers of features: with all studied datasets, we have the \( p \gg n \) problem. Removing irrelevant or redundant features often improves prediction accuracy, by reducing overfitting to training data. Recently, sparse linear models based on regularization techniques, which simultaneously embed feature selection and promoter sparsity, have received a lot of attention on different applications. However, different regularizers have distinct limitations. To choose an appropriate model, We should be clear about our needs.
Firstly, we would like to identify the most relevant subgroup for each main effect group, and thereby identify the most relevant subgroup for each interaction group. In this step, we want the interactions to obey strong hierarchy: an interaction subgroup can be selected only if all of its main effect subgroups are selected. In contrast, weak hierarchy would only require either subgroup of its main effects are selected. These strong and weak hierarchy concepts are introduced in Lim and Hastie (2015). In our situation, there are two cases in which a subgroup would have smaller total importance than its counterparts and, thereby would be discarded: a) it has no features that could match important features in maintained subgroups; b) it has features containing less information than corresponding features in maintained subgroups, though there are pairwise strong correlations. While case a) is not difficult to deal with for almost all sparse linear models, case b) should be handled carefully. For the pairs of features from different subgroups with strong correlations, we would like the linear models assign them coefficients tend to be equal in absolute value. Extremely, if a pair of features are exactly identical, we would like the linear models assign them identical coefficients. By doing so, we would be able to discard a subgroup simply because it has smaller total importance (sum of absolute feature coefficients) than its counterparts, instead of being resulted from randomness more or less.

Secondly, with only one subgroup selected for each main effect group, and consequently only one subgroup selected for each interaction group, we would like to identify the most relevant subgroups at final. For certain groups of main effects and interactions, possibly the only subgroups will be excluded in the final model. For instance, after we select “POTF” for “fluctuation”, “kCCP” for “amplifier”, “ENNGBD” for “resistance”, and the associated subgroups for corresponding interactions, we may discard the “fluctuation” group by excluding all “POTF” features, and keep all the interaction subgroups associated with “POTF”, because the respective functions of TFs all depend on core promoter sequence features. So, in short,
for this step, we do not require even weak hierarchy: a group of interactions could be selected without any group of its main effects are selected. In other words, we would like to select the most relevant subgroups: it is not necessary that our fitted model is a full model given in equation 4.1.

Thirdly, we would like to estimate models that are sparse on the level of (sub)groups of features. In other words, we want “(sub)groupwise sparsity” (number of (sub)groups with at least one nonzero coefficient). In contrast, we do not want that much of “within-group sparsity” (number of nonzero coefficients within each nonzero (sub)group), especially for those subgroups for TFs, since we would like each TF has features representing its role selected by the model.

Given these needs, we discuss the advantages and disadvantages of relevant regularizers as follows. Then we will introduce the algorithm we proposed for meeting our needs. Please note the sparse linear models we discussed here require features being standardized. Therefore, the importance of a feature could be reflected by its absolute coefficients in the fitted sparse linear model.

4.3 Sparse linear models and regularization techniques

4.3.1 The lasso regularization

A standard approach is $l1$ regularization, often referred to as lasso (Tibshirani, 1996). One of the reasons for the lasso’s popularity is that it performs feature selection by setting some coefficients to zero. Although the lasso has been successful in many cases, it has a couple of limitations (Zou and Hastie, 2005): a) if there are a set of correlated or nearly linearly dependent features, the lasso tends to randomly select only one feature from the set, even if many or all of these features are relevant in the unknown true model; b) in the $p > n$ case, the lasso would select at most $n$ features. These limitations both are hazardous for our situation. We know that there are pairwise strong correlations among alternative subgroups of TF features.
We definitely do not want a subgroup is selected in a random manner: a subgroup of features should be selected simple because it is better for explaining the target value. Moreover, on specific sets of synthetic promoters, we often have \( n \) be very small, like \( \sim 90 \). With only 3 TFs and 64 core promoter sequence features, we would have \( 3 \times 64 = 192 \) interactions which are very likely to be relevant. In such cases, if lasso only could select 90 features, a lot of relevant features might be discarded and, thereby the predictive performance of the fitted model would be bad.

4.3.2 The elastic net regularization

To tackle the limitations of the lasso, Zou and Hastie (2005) proposed the elastic net. The elastic net regularization is a linear combination of \( l_1 \) and \( l_2 \) regularizations. With \( l_1 \) regularization term, the elastic net still does automatic feature selection. By introducing a \( l_2 \) regularization term, the elastic net assign highly correlated features coefficients tend to be equal in absolute value. This phenomenon is known as “grouping effect” in statistics: grouping highly correlated features together and make them be in or out of the model at the same time (please note such grouping effect is different from predefined grouping structure in our work). Theoretically, a strictly convex regularization provides a sufficient condition for the grouping effect. Here for the elastic net, it is the \( l_2 \) regularization that guarantees strict convexity. Although the “grouping effect” is exactly what we want, the elastic net has a limitation in our situation: it tends to miss the features representing true effects with small coefficients, while such features could be important in biological settings (Xie and Zeng, 2010).

4.3.3 The ridge regularization

The \( l_2 \) regularization itself, often referred to as the ridge, is a standard approach for improving predictive performance by continuously shrinking the coefficients in linear models (Hoerl and Kennard, 1970). Moreover, continuous shrinkage makes the ridge
tend to fit stable models, i.e., regression coefficients are usually little affected by small changes in the data on which the ridge regression model is fitted. However, the ridge cannot produce a sparse linear model, because it cannot set coefficients of features to exactly zero, though it shrinks them towards zero. To combat this disadvantage, a couple of sparse versions of the ridge regularization have been proposed (Wang et al., 2006; Aseervatham et al., 2011). Specifically, Wang et al. (2006) proposed a ridge based feature selection procedure. With this procedure, after fitting the ridge model at each iteration, the increase of loss function by deleting a feature is calculated for each feature, using the remaining features and learned coefficients at that iteration. Using the calculated results in such a way, the respective features with smallest increase of loss function at each iteration are recursively removed until a stop criterion is satisfied: the increase of loss function is bigger than a certain fraction of the optimal loss function, where the fraction is a parameter determined by user. Please note, if we standardize the features, this feature selection protocol is equivalently to recursively removing features according to the absolute values of their respective coefficients, i.e., importance of features. The biggest disadvantage of this procedure is that it does not consider the possible grouping structure of features. But later we will show that it coincide with our procedure in terms of general ideas.

4.3.4 The group lasso and its variants

Moreover, given that we have prior knowledge on the hierarchical grouping structure of features, methods enforcing specific structured sparsity seem to be ideal in our situation (Yuan and Lin, 2006; Wang and Leng, 2008; Jacob et al., 2009; Liu et al., 2013; Simon et al., 2013; Lim and Hastie, 2015). But they all have severe limitations for our situation. To select predefined groups, Yuan and Lin (2006) proposed the group lasso. Its regularization term, called $l_1/l_2$ regularization, is the sum (i.e., $l_1$ norm) of the groupwise $l_2$ norms of the coefficients in different group. With
such a regularizer, the group lasso promote sparsity at group level, i.e., groupwise
sparsity. However, the group lasso has several limitations: a) it requires that the
features across different groups are not too correlated to work well; b) it suffers model
selection inconsistency; c) it could not enforce hierarchical selection of (sub)groups
of features: with hierarchical grouping, (sub)groups of features could be selected
only if their parents, or even ancestors, in the hierarchy are selected too; d) it does
not enforce within-group sparsity. Limitations a), b), and c) are all crucial in our
situation. The only exception, d), is not crucial for now, because all the subgroups of
features are learned from specific settings and we do not expect that much sparsity
within respective subgroups. However, it could be crucial in the future, if users of
our software would like to directly test some subgroups of features designed with
possible irrelevance or redundancy.

To tackle at least part of these problems, variants of the group lasso have been
proposed (Wang and Leng, 2008; Jacob et al., 2009; Liu et al., 2013; Simon et al.,
2013; Lim and Hastie, 2015). Wang and Leng (2008) proposed the adaptive group
lasso to bring in estimation efficiency and model selection consistency. Jacob et al.
(2009) proposed the overlapped group lasso that allows groups of features have over-
laps: some features can show up in more than one group and get a new coefficient
in each group; their ultimate coefficients are the respective sums of associated coef-
ficients in all groups they show up. Liu et al. (2013) proposed the smoothed group
lasso to tackle the problem of strong correlation among adjacent groups with the
natural grouping structure in SNPs. They would like to select all groups of SNPs
associated with response, even if some groups are highly correlated. Therefore, the
smoothed group lasso regularization is proposed as the sum of the group lasso reg-
ularization and a quadratic regularization on difference of coefficients of adjacent
groups. Simon et al. (2013) proposed the sparse group lasso to perform both group-
wise and within-group feature selection by blending the lasso and the group lasso
regularizations. Lim and Hastie (2015) proposed the hierarchical group lasso regularization to obtain solutions that satisfy the strong hierarchy property: an interaction can be present in the fitted model only if all of its main effects are present. They did so by representing the strong hierarchy property in terms of overlapped group lasso regularization: main effects appear not only in a group for main effects, but also in a group for interactions; when interactions are selected, main effects are guaranteed to be selected.

Though none of these variants of the group lasso alone can conquer all limitations mentioned above, we know there is a possibility that combining the relevant techniques proposed in these variants could solve the problem. However, to get a simple and intuitive modeling process, we decided to propose a sparse version of hierarchical group ridge as below.

4.4 Our algorithm: a sparse version of hierarchical group ridge

Given ridge assigning weights to features (the coefficients in the fitted linear models), we recursively eliminating (sub)group of features in a way similar to the RFE of individual features in Guyon et al. (2002). We call it gRFE: group version of RFE. Moreover, because we are interested in selecting all features helpful for prediction, instead of keeping certain number of features, we also utilized “one standard-error rule” to stop gRFE.

We exploited ridge because it assign highly correlated features coefficients tend to be equal in absolute value, instead of picking one in a random manner, like lasso. Though elastic net also has this advantage, its origin of such advantage is ridge. Also, elastic net has another hyperparameter for controlling the ratio of $l1$ in the linear combination of $l1$ and $l2$, which makes it be less efficient in time with cross-validation. Moreover, we do need groupwise sparsity, but none of the group lasso variants alone can meet all our needs. Additionally, RFE is a backward elimination
method. We choose it, instead of some forward selection, because RFE will make the importance of features be assessed in the context of other features included most cases, in contrast to forward selection methods. Lastly, we did not exploit best subset of features, simply because exhaustive enumeration of all features, even exhaustive enumeration of all subgroups of features, is not possible, given the enormous number of features we have, and the number of subgroups we have.

In a cross-validation approach, we perform gRFE combining with the ridge in a sequence of steps to promote sub-groupwise sparsity and groupwise sparsity, and then choose the best linear model to relate the remaining features to transcript production rates, as follows.

4.4.1 Step A: determining the best subgroup of features for each main effect group.

In step A, we calculate the total importance for each subgroup of main effects, while we do not compute it for respective subgroups of interactions, and the importance of interactions are attributed to each of the associated subgroups of main effects. The reasons are: a) in step A, our goal is promote the subgroup-wise sparsity within each main effects group – we want a group of main effects are only represented by the relevant subgroups; b) and thereby we require strong hierarchy for step A: an interaction would not present if any of its two main effects is excluded.

As a result, in this step, the importance of each main effect feature is defined as the sum of the absolute coefficients of the main effect itself and all relevant interactions. Denote $i$ as one of the $S_f$ main effect features in the “fluctuation” group, $j$ as one of the $S_a$ main effect features in the “amplifier” group, and $k$ as one of the $S_r$ main effect features in the “resistance” group. We have $W^{(f)}_i$ as the importance of the $i$th main effect feature in the “fluctuation” group,

$$W^{(f)}_i = |\hat{\beta}^{(f)}_i| + \sum_{j=1}^{S_a} |\hat{\beta}^{(f:a)}_{i:j}| + \sum_{k=1}^{S_r} |\hat{\beta}^{(f:r)}_{i:k}| + \sum_{j=1}^{S_a} \sum_{k=1}^{S_r} |\hat{\beta}^{(f:a:r)}_{i:j:k}|, \quad \forall i \in [1 \ldots S_f].$$

(4.3)
Please note the coefficients used for calculating importance are the CV mean coefficients.

Then, we have the total importance for the 1st subgroup in the “fluctuation” group as,

\[ W^{(f_1)} = \sum_{i=1}^{S_{f_1}} W_i^{(f)}. \]  \hfill (4.4)

And the total importance for the 2nd subgroup in the “fluctuation” group as,

\[ W^{(f_2)} = \sum_{i=S_{f_1}+1}^{S_{f_2}} W_i^{(f)}. \]  \hfill (4.5)

It keeps going if there are more than two subgroups in the “fluctuation” group. Note \( S_f = \sum_{l=1}^{L_f} S_{f_l} \), where \( S_{f_l} \) is the size of the \( l \)th subgroup, and \( L_f \) is the number of subgroups.

Similar equations apply to the “amplifier” and “resistance” groups if needed.

With the total importance for each subgroup of main effects, we are able to perform gRFE for determining the best subgroup of features for each main effect group, as follows.

1. Select the best ridge model with cross-validation on training set, using all given features; set it as current model;

   item With current model, identify the least important subgroup of features for each group of main effects:

   (a) If no group has more than one subgroup, or all groups have been stopped for eliminating subgroups, stop step A;

   (b) If only one group has more than one subgroups, the subgroup with the smallest total importance is the least important subgroup, go to step A2;
(c) For each group with more than one subgroups, calculate the percentage of decrease for each subgroup, comparing to the subgroup with the biggest total importance. The subgroup with the largest percentage of decrease is the least important subgroup. If multiple subgroups have percentages of decrease close to the largest one, they will be the least important subgroups.

2. Remove the least important subgroup(s) of main effects, and consequently associated subgroups of interactions, and then reselect a ridge model with cross-validation:

   (a) If the ridge model selected at this iteration has a mean squared error (MSE) smaller than that of the best model, or its MSE is within one standard error of that of the best model, update the best model to be current model, since we favor simpler model; and then go back to step A1;

   (b) Otherwise, add back the removed subgroups of features, and then stop subgroup elimination for the associated group. Go back to step A1.

4.4.2 Step B: determining the helpful groups of features.

After step A, if we still have more than one subgroups for a certain group, we merge those subgroups into a single group in our grouping structure, since it means that these subgroups are complementary to each other more or less. Therefore, in step B, we have no subgroups; we do not require strong hierarchy: an interaction group could present even if none of its main effect groups is selected. But we may still have correlated groups: a main effect group might be correlated with its associated interaction groups. Also, we want to promoter groupwise sparsity at this step: we want to remove a main effect group or an interaction group if it is less important,
and is not crucial for accurate prediction. As a result, we still exploit gRFE with ridge for this step.

However, if the least important group is a main effects group, we will rank it with all of its interaction groups together based on average importance, and then assign the group with the smallest average importance as the least important group. The reason is that there might be strong correlations between a main effects group and its associated interactions group. Given that ridge tends to assign similar coefficients to correlated features, the interactions group probably becomes important simply because of its bigger group size. An interactions group has its feature number more than or equal to any of its associated main effects group. Moreover, introducing interactions usually makes the model more complicate. To favor simpler model, we should try eliminate an interactions group at first, if the group has smallest average importance among the main effects group and all of its associated interactions groups.

Therefore, we need to update $W_{i}^{(f)}$ and $W^{(f)}$ as

$$W^{(f)} = \sum_{i=1}^{S_{f}} W_{i}^{(f)} = \sum_{i=1}^{S_{f}} |\beta_{i}^{(f)}|,$$  \hspace{1cm} (4.6)

And now we have total importance for every interaction group. For example, we have $W^{(f,a)}$ as

$$W^{(f,a)} = \sum_{i=1}^{S_{f}} \sum_{j=1}^{S_{a}} W_{i,j}^{(f,a)} = \sum_{i=1}^{S_{f}} \sum_{j=1}^{S_{a}} |\beta_{i,j}^{(f,a)}|.$$  \hspace{1cm} (4.7)

With these total importances of respective groups, we perform gRFE to determine the most relevant groups of features, as follows.

1. Select the best ridge model with cross-validation on training set, using all remaining features after Step A; set the best ridge model as current model
2. With current model, identify the least important group of features: the group with smallest total importance. If it is a main effects group, we will rank it with
all of its interaction groups together based on average importance, and then assign the group with the smallest average importance as the least important group.

3. Remove the least important group of features, and then reselect a ridge model with cross-validation:

   (a) If the ridge model selected at this iteration has a mean squared error (MSE) smaller than that of the best model so far, or its MSE is within one standard error of that of the best model so far, go back to step 2 with the selected ridge model as current model, and with the best model being updated to current model if its MSE is smaller;

   (b) Otherwise, add back the removed group of features, and then stop.

4.4.3 Step C: selecting the best model in terms of predictive performance.

Unlike steps A and B are for enforcing sub-groupwise and groupwise sparsity, step C is for getting the best model in terms of predictive performance. In this step, usually we do not have correlated subgroups or groups anymore. Or, even if we have correlated subgroups or groups, we would like to keep them because they have complementary information to each other.

Therefore, we could compare the predictive performance of the lasso, ridge, and elastic net, to select the best one with smallest MSE. With cross-validation on training set, we got the best lasso, ridge, elastic net models with hyper-parameters respectively tuned. The we select the model with the smallest average MSE for predicting target values on test set.
4.4.4 Optional Step D: determining the helpfulness of an additional group of features.

If an additional group of features could be viewed as an alternative subgroup for a certain group of main effects, we add it to the remaining grouping structure obtained at the end of Step B in various reasonable ways, and go through Step C again. If the additional group of features improve model performance, it means that the additional group provides new information, and the associated mechanistic hypothesis gets support. Meanwhile, our results get updated.

With an additional group, we will go through Steps A, B, C again, only if it changes our grouping structure. For instance, adding a group of main effects would increase our number of groups of main effects from 3 to 4, and increase the number of groups of interactions from 4 to 11.
Table 4.1: Hypotheses associated with groups of features for main effects and interactions.

<table>
<thead>
<tr>
<th>Math Expression</th>
<th>Hypotheses</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X^{(f)} )</td>
<td>TFs control the respective fluctuations of transcript production rates.</td>
</tr>
<tr>
<td>( X^{(a)} )</td>
<td>The core promoter amplifies transcript production.</td>
</tr>
<tr>
<td>( X^{(r)} )</td>
<td>The gene body nucleosomes regulate transcription through resisting elongating RNAPII’s.</td>
</tr>
<tr>
<td>( X^{(fa)} )</td>
<td>There is synergy between TFs and the core promoter: the transcript production rates could not be explained by additive function of TFs and the core promoter. Probably TFs interact with mediator and / or PIC components to recruit RNAPII, and then the core promoter sequence would influence the RNAPII recruitment, as well as the following RNAPII scanning for TSS and RNAPII initiation.</td>
</tr>
<tr>
<td>( X^{(fr)} )</td>
<td>There is synergy between the TFs and the gene body nucleosomes: the transcript production rates could not be explained by additive function of the TFs and the gene body nucleosomes. A possible explanation is that gene body nucleosomes could regulate transcription through resisting RNAPII elongation, while the amount of elongating RNAPII is at least partly determined by TFs.</td>
</tr>
<tr>
<td>( X^{(ar)} )</td>
<td>There is synergy between the core promoter and the gene body nucleosomes: the transcript production rates could not be explained by additive function of the core promoter and the gene body nucleosomes. A possible explanation is that gene body nucleosomes could regulate transcription through resisting RNAPII elongation, while the amount of elongating RNAPII is at least partly determined by the core promoter.</td>
</tr>
<tr>
<td>( X^{(fa,r)} )</td>
<td>There is synergy among TFs, the core promoter, and the gene body nucleosomes: the transcript production rates could not be explained by additive function of them. A possible explanation is that gene body nucleosomes could regulate transcription through resisting RNAPII elongation, while the amount of elongating RNAPII is at least partly determined by the interaction of TFs and the core promoter.</td>
</tr>
</tbody>
</table>
Table 4.2: Hypotheses associated with various linear combinations of main effects and interactions.

<table>
<thead>
<tr>
<th>Math Expression</th>
<th>Hypotheses</th>
</tr>
</thead>
</table>
| \( X^{(a)}\beta^{(a)} + X^{(f,a)}\beta^{(f,a)} \) | 1. The core promoter could regulate transcription without interacting with TFs, probably due to the basal amount of transcription: RNAPII could be recruited without helps from TFs;  
2. TFs functioning must be performed through interactions with other regulatory elements: there is synergy between TFs and the core promoter elements, and TFs could not function alone. Probably TFs interact with mediator and/or PIC components to recruit RNAPII, and then the core promoter sequence would influence the RNAPII recruitment, as well as the subsequent RNAPII scanning for TSS and RNAPII initiation.  
3. While TFs and the core promoter elements probably regulate transcription at PIC formation and RNAPII initiation stages, the gene body nucleosomes possibly regulate transcription at RNAPII elongation. And probably because of the “group action” of elongating RNAPII’s, the regulatory effects of gene body nucleosomes depend on the amount of elongating RNAPII. |
| \( X^{(a,r)}\beta^{(a,r)} + X^{(f,a,r)}\beta^{(f,a,r)} \) | 1. The core promoter could regulate transcription without interacting with TFs, probably due to the basal amount of transcription: RNAPII could be recruited without helps from TFs;  
2. TFs functioning must be performed through interacting with other regulatory elements: there is synergy between TFs and the core promoter elements, and TFs could not function alone. Probably TFs interact with mediator and/or PIC components to recruit RNAPII, and then the core promoter sequence would influence the RNAPII recruitment, as well as the subsequent RNAPII scanning for TSS and RNAPII initiation.  
3. While TFs and the core promoter elements probably regulate transcription at PIC formation and RNAPII initiation stages, the gene body nucleosomes possibly regulate transcription at RNAPII elongation. And probably because of the “group action” of elongating RNAPII’s, the regulatory effects of gene body nucleosomes depend on the amount of elongating RNAPII. |
Preparing for predicting yeast dynamic transcript production rates in cell cycle

Cell division cycle is a fundamental process in biology. If we could provide a predictive model that links the changes in TF concentrations to the changes in transcript production rates, we would be able to provide deeper insight into cell division cycle, understanding how cell-cycle events are regulated. Recently, Eser et al. (2014) measured dynamic transcript production rates in yeast \textit{S. cerevisiae}, using metabolic RNA labeling and comparative dynamic transcriptome analysis (cDTA). While some other data sets could only measure transcript levels and assume constant transcript degradation rates, Eser et al. (2014) derived dynamic transcript production rates and degradation rates every 5 min during three cell cycles. The transcript level, i.e. the amount of transcripts, is the result of two opposite reactions: transcript production and transcript degradation. Consequently, the direct object of the transcriptional regulation should be transcript production rate Pelechano et al. (2010). Therefore, Eser et al. (2014) provides a good data set for our study.

Please note traditional time series analysis, which predict future target value
from past target values, is not applicable for such scenarios. The reason is that it will not answer how the transcript production rates are regulated by DBFs and other regulatory elements, and what the underlying mechanisms of transcriptional regulation are.

In this chapter, we prepare for predicting yeast dynamic transcript production rates in cell cycle with Eser et al. (2014) data set. We firstly describe how we determine the TFs and genes set we study. Secondly we describe how we prepare inputs for running COMPETE to get occupancy profiles for extracting features. Especially, we describe how we estimate TF concentrations. Then we discuss why time series analysis is not appropriate for our work, and the autocorrelation problem that we take care of. Lastly, we introduce the different ways to do training-test splitting for cross-validation on cell cycle data set. And we explain why we choose stratified training-test split, where two randomly selected time points for each gene are assigned into test set, and, one time point for one fold, all remaining time point for each gene are randomly splitted into the validation folds.

5.1 Choosing TFs and genes for cell cycle data set

Ideally, we want to study a set of TFs and genes, where a) each TF has a cell cycle averaged TF concentration as scaling factor for estimating its concentrations at every time point, and b) its associated gene has cell cycle transcript production rates to get its relative concentrations at every time point; c) each gene is regulated by only the considered TFs, and d) has cell cycle transcript production rates.

For constraint c), we need associations of TFs and their target genes. We used ChIP-chip derived associations ($P$-value < 0.001) of TFs and their target genes (MacIsaac et al., 2006). Constraints b) and d) are determined by the cell cycle data sets we exploited (Eser et al., 2014). For constraint a), we could either use TF concentrations inferred from MNase data by our group (Zhong et al., 2014), or use
multiples of $K_d$.

With TF concentrations inferred from MNase data in Zhong et al. (2014) and cell cycle transcript production rates from Eser et al. (2014), we identified 12 TFs and 346 genes for study.

5.2 Preparing inputs for COMPETE to extract features from occupancy profiles

The occupancy profiles used by our framework are produced by COMPETE, a thermodynamic model of competitive DNA binding previously developed by our group. COMPETE extends hidden Markov models (HMMs) to model competition between nucleosomes and many TFs at given concentrations, resulting in a profile of continuous levels of DNA binding by all factors. The COMPETE software efficiently implements this thermodynamic model to produce chromatin occupancy profiles at single-nucleotide resolution on a genome-wide scale. With such occupancy profiles, we are able to extract biologically mechanistic features.

COMPETE takes the sequence specificities and concentrations of multiple kinds of DBFs, along with a DNA sequence, as inputs, to generate the promoter occupancy profile, as shown in Figure 5.1. There are important assumptions involved in the preparation of inputs for COMPETE:

1. the respective time lags of different TFs from being transcribed as transcripts (mRNA) to being activated TF in nucleus are independent of each other: they are not required to be identical;

2. the concentration of activated TF in nucleus is proportional to its mRNA level after transcription. In the case of missing mRNA level, transcript production rates are used, though it means that we ignored the distinct transcript decay rates;
Figure 5.1: Taking the sequence specificities and concentrations of multiple kinds of DBFs, along with a DNA sequence, as inputs, COMPETE generates respective occupancy profiles for studied promoter sequences. Please note here we illustrated with the cell cycle setting, since there are DBF concentrations over time. We used Time=160 as an example, so the promoter occupancy profile is the one at Time=160.

In following subsections, we will describe how we get the inputs for running COMPETE.

5.2.1 Sequence specificities

In this paper, the sequence specificities of TFs are position-specific scoring matrices (PSSMs) curated by Gordan et al. (2011); the binding specificities of nucleosomes are generated from the study conducted by Kaplan et al. (2009).

5.2.2 TF concentrations

Unfortunately, the real concentrations of TFs are difficult to obtain (Granek and Clarke, 2005; Stormo and Fields, 1998). We overcome this challenge by estimating the TF concentration from the associated gene’s transcript level, as shown in Figure 5.2A.
Between transcript production and TF concentrations intervene numerous post-transcriptional steps. Yet we have limited knowledge of the effects of these steps. Therefore, we simply apply a time lag to get the normalized TF concentration for corresponding gene’s transcript level. Recently, our assumption is fortified by an observation that most of the variance in protein change is explained by changes in mRNA abundance (Lee et al., 2011).

Denote the cell cycle length as $T$ and the time lag for $tf$ as $l_{tf}$. The normalized concentration of $tf$ at time $t$, $NC_{tf}(t)$, is set as the transcript level of the associated gene $g$ at time $(t + T - l)\%T$, $C_g((t + T - l)\%T)$, divided by its average transcript level in cell cycle AVG($C_g$),

$$NC_{tf}(t) = \frac{C_g((t + T - l_{tf})\%T)}{\text{AVG}(C_g)}, \quad \forall 1 \leq t \leq T$$

To estimate time lags, for a cell cycle data set, we firstly determined an inclusive interval of the possible time lags: $[5, 30]$. Then, for each $tf$, the time lag $l_{tf}$ is estimated as the one producing the highest absolute pearson correlation coefficient between rotated expression profile of the gene corresponding to $tf$ and a target gene of $tf$ according to ChIP-chip derived associations ($P$-value $< 0.001$) of TFs and their targets (MacIsaac et al., 2006). Here is the reasoning for such an approach: a target gene usually is regulated by multiple TFs; however, for each TF, there should be at least one target gene that is mainly regulated by the TF, and thereby its expression profile have the highest pearson correlation coefficient with the TF’s normalized concentration profile, which is the TF’s corresponding gene expression profile right rotated with the underlying time lag.

A possible concern about our approach of inferring TF specific time lag is that we use whole data set. Superficially, we are “seeing data points in test set twice”, and thereby the estimated generalization performance may be overestimated. However,
we argue that time lags are approximately constant for different TFs. Even for those unseen data points of next cell cycle, a time lag would be identical for a certain TF. Therefore, it should be fine if we only use all data points, including those in test set, to estimate TF specific time lags.

Then, as shown in Figure 5.2B, the estimated TF concentration of \( tf \) at time \( t \), \( \tau_{tf}(t) \), is the product of a TF specific scaling factor \( \alpha_{tf} \) and the normalized TF concentration \( NC_{tf}(t) \),

\[
\tau_{tf}(t) = NC_{tf}(t) \times \alpha_{tf}, \quad \forall 1 \leq t \leq T \tag{5.2}
\]

To estimate the TF specific scaling factors \( \alpha_{tf} \)'s, we can either use multiples of equilibrium dissociation constant \( K_d \)'s, or use average TF concentrations across cell cycle inferred from other data sources (Zhong et al., 2014).

The equilibrium dissociation constant of TF for the optimal binding sites, \( K_d \), is calculated as described in (Granek and Clarke, 2005):

\[
K_d = e^{-	riangle G/RT} = e^{-\left(\sum_{i=1}^{k} RT \ln(f_{b,j}/p_{b})\right)/RT} = e^{-(\sum_{i=1}^{k} \ln(f_{b,j}/p_{b}))} = \prod_{i=1}^{k} \frac{p_{b}}{f_{b,j}}. \tag{5.3}
\]

where \( b \) is the base at position \( j \) of optimal binding site, \( p_{b} \) is the prior probability of base \( b \), and \( f_{b,j} \) is the observed frequency of base \( b \) at position \( j \) of all potential binding sites of the corresponding TF.

5.2.3 DNA sequences

For cell cycle data set, we use the natural promoter sequences, which are 600 bp sequences upstream of 5' boundaries of ORFs, the natural ORF sequences,
5.3 Autocorrelation analysis of transcript production rates throughout cell cycle

Although the cell cycle gene expression profiles are time series data sets, we do not refer to time series models, like autoregressive integrated moving average (ARIMA) models. Such time series models will not answer how the transcript levels and transcript production rates are regulated by DBFs, and what the underlying mechanisms
of transcriptional regulation are. In addition, these models are just be capable to predict the transcript production rates at next time point by exploiting transcript production rates at previous time points. They are not able to predict the transcript production rates of the gene that is downstream of a new promoter sequence, like promoter sequence outside of training set and mutated promoter sequence, even if the DBF concentrations at starting time points and DBF binding specificities are given, simply because of the lack of transcript production rates at starting time points. So, we proposed our model as a pipeline consists of extended-HMMs part for generating occupancy profile, feature engineering part for extracting features from occupancy profile and promoter sequence, and linear regression part for relating features to transcript production rate.

However, we cannot ignore the possibility of serious serial correlation (autocorrelation) for cell cycle gene expression profiles. It is reasonable that the gene transcript levels are highly autocorrelated. Firstly, the sampling (time) interval of the gene expression profile is small such that the transcript production rate or transcript level is always smoothly moving to the rate or level at next time point. Secondly, the promoter occupancy would not be significantly changed during a single, or even several, sampling interval(s). Even though the binding TF instances might dissociate very fast, the promoter might also have newly associated binding TF instances very fast to maintain the promoter occupancy profile (Voss and Hager, 2008).

Autocorrelation is known to complicate statistical modeling by reducing the number of independent observations: if the time points we studied are too closely consecutive, we will have serious serial correlation and thereby low effective sample size of the cell cycle data set. So, we need to make sure that there is a big enough sampling (time) interval. With big enough time intervals, we will not waste our computational resources on redundant observations, and will not miss information by discarding time points. Here, as demonstrated in Figure 5.3, we confirmed that
the default time interval 10 in Eser et al. (2014) is a good choice and, thereby we
directly used all \( \mu_5, \mu_{15}, \mu_{25}, \mu_{35}, \mu_{45}, \mu_{55}, \) and \( \mu_{65} \) time points for each gene.

### 5.4 Different ways to do training-test splitting for cross-validation

For cell cycle data sets, there are four ways to do training-test splitting for cross-validation. Using Eser et al. (2014) sub dataset with 12 TFs and 346 genes as an example, we describe these four settings as below.

1. Fixed training-test split, where some certain time points for each gene are assigned into test set, and a certain time point for each gene is assigned into a certain validation fold. For example, it could be time points \( \mu_{55} \) and \( \mu_{65} \) of each gene are assigned into test set, while \( \mu_5, \mu_{15}, \mu_{25}, \mu_{35}, \) and \( \mu_{45} \) of each gene are assigned into validation folds 0, 1, 2, 3, and 4 respectively.

2. Stratified training-test split, where two randomly selected time points for each gene are assigned into test set, and, one time point for one fold, all remaining time point for each gene are randomly splitted into the validation folds: folds 0 to 4 with Eser et al. (2014) data set. With such a setting, the time points in test set, and the time point in a certain validation fold would be different across distinct genes.

3. Random training-test split on both gene and time coordinates, where all time points of all genes are viewed as independent data points and form a single pool, then \( 346 \times 2 \) data points are randomly selected for test set, and 346 data points are randomly selected for each validation set.

4. Random training-test split on gene coordinate, where \( 2/7 \) of the 346 genes are randomly sampled and all their 7 time points are assigned into test set, and
Figure 5.3: For Eser et al. (2014), we showed that almost all genes in the 346 genes subset have reasonably small autocorrelations with time lag 10 mins, as shown in (A), given the 95% confidence interval is usually \([-0.2, 0.2]\), and the 99% confidence interval is usually \([-0.3, 0.3]\). Also, when we were inferring time lags for estimating TF concentrations, we followed Eser et al. (2014)'s suggestion to use piecewise linear function for estimating transcript production rates between given time points. With ERS1 gene as an example, we estimated the transcript production rates at all 70 mins, as shown in (B). With such estimated profiles, we calculated the autocorrelations with different time lags for all 346 genes. As shown in (C), 10 mins is generally a good time lag, with autocorrelation in the 95% confidence interval. The horizontal lines displayed in the plot correspond to 95% and 99% confidence intervals.
then the remaining genes are randomly split into 5 sets and all the 7 time points of genes in each set are forming a validation fold.

We think the 2nd and 3rd settings are more interesting. And we present results of 2nd setting in this dissertation. We also get good results with 1st and 3rd setting, while we could not get good results with 4th setting. A possible explanation is that, with 4th setting, we would have some unseen or rarely seen features for genes in test set. In other words, there would be TF binding features or core promoter sequence features having mostly zeros on training set, yet non-zeros on test set, especially given that we are doing cross-validation on training set. Therefore, we could not get information about how such TFs or sequence features function from training set, i.e., we don’t have coefficients or approximately correct coefficients for such features in our model fitted on training set.

5.5 Discussion

Firstly, we identified a set of 12 TFs and 346 genes for study, as discussed in Section 5.1. To get dynamic occupancy profiles for extracting features, we need dynamic TF concentrations. Unfortunately, the real concentrations of TFs are difficult to obtain (Granek and Clarke, 2005; Stormo and Fields, 1998). We overcame this challenge by estimating the dynamic TF concentration from the associated gene’s dynamic transcript level, as described in Section 5.2.2. Because cell cycle data set from Eser et al. (2014) does not provide transcript level, we used transcript production rate to do estimation. Also, because we need a scaling factor, which could be understood as mean TF concentrations over cell cycle, we exploited the estimated TF concentrations from another project in our group, which are inferred from MNase data (Zhong et al., 2014). Moreover, we arbitrarily selected promoter length to be 600 bp and, thereby the promoter region is the 600 bp DNA sequence upstream of the downstream gene’s
5’ open reading frame (ORF) boundary, i.e., the gene’s translation start site (TrSS).

In Eser et al. (2014), there are transcript production rates at 7 time points through cell cycle for each gene: $\mu_5, \mu_{15}, \mu_{25}, \mu_{35}, \mu_{45}, \mu_{55}$, and $\mu_{65}$. And the authors suggested that piecewise linear function could be used to estimate transcript production rates between given time points. We did autocorrelation analysis to show that there is no significant serial correlation with the default interval of 10 units (See Figure 5.3 for details). Therefore, all 7 time points can be used as independent to each other. Moreover, there are 4 settings about splitting data points for cross-validation scheme, as discussed in Section 5.4. We chose to focus on stratified 5-fold Cross Validation: when doing training-test split and correspondingly cross-validation split, we always randomly assign 2/7 of all data points into test set, and 1/7 of all data points into each fold of the 5 fold cross-validation.
Applying PreTrans to predict yeast dynamic transcript production rates in cell cycle

In this chapter, we apply our framework to Eser et al. (2014) cell cycle dataset to answer a crucial question in biology: how transcript production rates for different genes change during the cell cycle. We demonstrate that PreTrans performs very well on test set for predicting dynamic transcript production rates over cell cycle. We record the iterations that our algorithm performs on the chosen data set, and discuss the certain decisions it makes on eliminating (sub)groups of features. By interpreting our final model with excellent predictive performance, and comparing our final model and its counterparts, we provide support for certain hypotheses. Especially, we show that transcriptional synergy between TFs and core promoter regulatory elements is crucial for predicting transcript production rates over cell cycle. And we show that expected number of binding instances for each TF is good for representing TF’s role, and the subgroup of \( k \)-mer based sequence features from Lubliner et al. (2015) are selected for reflecting core promoter regulatory elements. We then demonstrate the \( k \)-mer based sequence features could be partly replaced by the occupancy profile
based features, which represent TATA elements and -1 and +1 nucleosomes. We also discuss the possible explanations for why certain hypothesized mechanisms could not get support.

6.1 Excellent results of PreTrans on cell cycle data set

PreTrans produced excellent results on the cell cycle data set, as shown in Figure 6.1. We got Pearson correlation coefficient $C_p = 0.751$ and coefficient of determination $r^2 = 0.564$ on test set. Also, we observed that the cross-validation performance provided a good estimate of predictive performance, indicating we did not have that much of overfitting.

Our framework selected the interactions “ENTF × kCCP” only,

$$y = \beta_0 + \mathbf{X}^{(f:a)}\beta^{(f:a)} + \epsilon,$$

(6.1)

where “ENTF” is selected for $X^{(f)}$, and “kCCP” is selected for $X^{(a)}$.

Such a selection means:

1. “ENTF” is the best subgroup of features for representing TF’s regulatory role;
2. “kCCP” is the best subgroup of features for representing core promoter’s regulatory role;
3. there is transcriptional synergy between TFs and core promoter elements: they must work together through synergistic interactions.

6.2 The gRFE path of feature elimination on DREAM6 data set

We recorded the gRFE path of eliminating (sub)groups of features in Tables B.6 – B.10. At each iteration, we recorded the performance metrics of best model so far, that of current model, and the decision of our algorithm: removing or adding
Figure 6.1: Scatter plots of our final model’s performance. Please note the performance metrics in (A) are calculated after all hold out sets are aggregated. But when we do model selection, we use the mean cross-validation metrics.
back (sub)groups of features, and consequently the change of feature set size. Please note we started running the algorithm with only “ENTF” and “ProbOfAtLeastOne-Instance” subgroups in “fluctuation” group. We did this for two reasons: a) we wanted to reduce the feature set size for efficient computation; b) as shown in Figure 6.5, the 8 subgroups of features in “fluctuation” group are highly correlated to each other, probably because all of them are “ENTF” itself or its variants by applying certain functions. But we did replace “ENTF” with the other subgroups to test the associated hypothesized mechanisms (See section 6.3.).

6.3 Testing results of hypothesized mechanisms relevant to TFs

In our final model, the signs of interaction terms varies, and an individual TF can have different signs within distinct interaction terms. These support the hypothesis that the roles of activators and repressors TFs played are usually determined by transcription cofactors, which is proposed in section 2.2. Especially, the other part of the interaction terms in our final model are always sequence features extracted from the core promoter region, most likely representing the mechanisms relevant to transcription cofactors. Moreover, with same core promoter sequence feature, the coefficients of the interaction terms associated with distinct TFs are different, supporting the hypothesis that different TF’s contribution to transcript production might be distinct and non-redundant from each other, as proposed in section 2.2. In following subsections, we discuss the testing results of other hypothesized mechanisms relevant to TF binding to DNA and TF’s contribution to transcript production.

6.3.1 TFs control the fluctuation of transcript production

To test the hypothesis that TFs control the fluctuation of transcript production (see section 2.2), we designed a baseline model by excluding TF relevant features. As shown in Figure 6.2, without TF relevant features, the baseline model performed
much worse than our final model. Given that TFs are the only group of factors have feature values changing over time in our model, it should be TFs control the fluctuation of transcript production.

6.3.2 Nucleosome’s indirect effects on transcription through TFs are ignorable

As discussed in section 2.1, it is suggested that nucleosomes compete with TFs for DNA binding sites, and thereby might influence transcription through TFs. To test this hypothesis, we tried setting nucleosome concentration as zero for COMPETE to generate occupancy profile and thereby extract features. Surprisingly, as shown in Figure 6.3, with this setting, the predictive performance of our final model became even a little bit better.

We think this is probably because nucleosome’s indirect effects on transcription may be performed through regulatory elements within the core promoter: in Section 6.4.1, we will see that the effects of “kCCP” sequences could be partly explained by nucleosomes within the core promoter and nearby TSS. Moreover, it is possible that the observations about the regulatory role of poly(dA:dT) tract on expression resulted from some other mechanisms, instead of nucleosomes competing with TFs for DNA access.

For Raveh-Sadka et al. (2012), one thing need to note is that promoter variants were mainly constructed with the region 150 bp upstream of the TSS in the native HIS3 promoter in yeast. In Sharon et al. (2012), for each of the synthetic promoters, its 103 bp variable promoter region is directly upstream of an identical 100 bp TATA-containing yeast native HIS3 core promoter. Given that these varying regions are contained in the core promoter or the region close to the core promoter at least, in these two studies, the observations about poly(dA:dT) tract, and thereby nucleosome, may be related to the regulatory elements within the core promoter, instead of the TFs. Especially, they are on a set of synthetic promoter variants constructed
Figure 6.2: Without TF relevant features, the core promoter relevant features alone produced a model much worse than our final model (See Figure 6.1).
with one or two promoter background sequences and/or a small set of TFs.

And it is notable that Raveh-Sadka et al. (2012) and Sharon et al. (2012) are both about expression measurements, not about TF binding measurements. By measuring TF binding and nucleosome formation on the same set of sequences in Sharon et al. (2012), Levo et al. (2015) suggest that, the in vivo effect of a poly(dA:dT) tract on expression might result from a combination of multiple mechanisms, including a nucleosome-mediated effect on TF binding, and a direct effect of the tract on TF binding strength through flanks to core binding sites.

6.3.3 Stability weight, saturation, and distance decay are not helpful for predictive modeling of transcript production rates in cell cycle

In sections 2.1 and 2.2, we proposed that nucleosomes might influence the stability of nearby TF binding instances, and consequently their respective contributions to transcript production; for each TF, a single TF binding instance would be enough for its functioning (saturation); a TF binding instance’s contribution to RNAPII recruitment, and consequently transcript production (activation or repression), decreases as it is placed further away from TSS. Correspondingly, in section 3.1, we defined “POTF” subgroup, as well as two weighting functions: being weighted by distance relative to TSS; and being weighted by nucleosome occupancy nearby. Applying the two weighting functions to “ENTF” and “POTF”, we got 8 subgroups of features.

However, our final model did not support these hypotheses, because it selected “ENTF” for the “fluctuation” group of TF relevant features. Replacing the “ENTF” with any other subgroup of TF features, like “POTF”, slightly improved the model performance, or did not improve it at all, as demonstrated in Figure 6.4. For example, with “POTF”, we will have CV mean MSE as 0.071, with standard error 0.006, while our final model has CV mean MSE as 0.072, with standard error 0.005. It is hard to tell the difference.
Figure 6.3: Set nucleosome concentration as zero for COMPETE to generate occupancy profile and extract features. With this setting, the predictive performance of the final model is even slightly better in terms of CV mean MSE: 0.067 is within one standard error of that of our final model \([0.072 - 0.005, 0.072 + 0.005]\), and the model is simpler without nucleosome concentrations. (See Figure 6.1 and Table B.10.)
Figure 6.4: Replacing “ENTF” in our final model by an alternative subgroup of features in “fluctuation” almost makes no change in terms of predictive performance, using “POTF” as an example.
For stability weight, it is notable that the Lickwar et al. (2012) observation is based on Rap1p binding only. Therefore, even if it is true, it could be specific to Rap1p binding.

For saturation of TF binding instances, it is notable that, recently, Levo et al. (2015) studied TF binding to sequences with multiple binding sites, exploiting synthesized sequences containing one to seven Gcn4p consensus sites. They showed that, only with pretty high concentration and with more than 3 or 4 sites, there would be a significant fraction of DNA sequences bound by two TF instances; in other cases, there would be only naked DNA sequences and DNA sequences bound by one TF instance. This observation indicates that, in most cases, we probably have no more than two binding instances at the same promoter for a certain TF. Therefore, it is likely that we may not be able to tell whether a single TF binding instance would be enough for its functioning, because we rarely have promoters bound by more than two instances of a certain TF.

For distance decay, it is important to note that, in section 1.4.5, we introduced that, a) metazoan genes are generally regulated by enhancers through DNA looping; b) yeast *S. cerevisiae* genes are at most regulated by enhancer-like UASs, and are not regulated by long distance enhancer–core promoter interactions; c) probably through DNA looping, mutant forms of Sin4p were identified in *S. cerevisiae* that enabled UASs to activate transcription at longer distances from the core promoter than normally observed in yeast: UASs are normally only 100–200 bp from the core promoter TATA element (Dobi and Winston, 2007). Therefore, it is possible that the observations of distance decay actually is just a correlation of distance and the DNA looping mechanisms: the real cause of decay of TF’s effects might be relevant to DNA looping. Moreover, just like ~10 bp periodicity actually is caused by site flanks (See Figure 2.4), we think that another possible real cause of distance decay could be the flanking sequences of TF binding sites.
Lastly, because these observations probably are caused by effects in a random way, and the hypothesized mechanisms probably are just about correlation, instead of causation, it could be that all the 8 features for every TF are close to each other on every promoter. Therefore, it would not be a surprise that these features are highly correlated, as shown in Figures 6.5 and 6.6. Still, it is possible that our respective feature designs are not good enough for representing these hypotheses.

6.4 Testing results of hypothesized mechanisms relevant to core promoter elements

6.4.1 The core promoter amplifies transcript production rate

Our final model supports this “amplifier” hypothesis (see section 2.3), because the final model contains and only contains interactions between TFs features and core promoter features. It even support the hypothesis about transcriptional synergy, which we will discuss in section 6.5.

But our final model does not support the model of promoter specific basal amount of transcription, because there are no features of the “amplifier” appear in the final model as main effects: they all appear as part of the respective interactions. To further validate our results, we designed a baseline model by taking “fluctuation” group of features only (“ENTF”), the accuracy of the model significantly decreased, as shown in Figure 6.7. Also, we will demonstrate that the interaction between core promoter and TFs are crucial, in section 6.5.

6.4.2 The “kCCP” group of features reflect mechanisms more than TATA element, −1 nucleosome, and +1 nucleosome

Moreover, our final model selected “kCCP” group of features to represent the “amplifier” function of core promoter. This selection indicates that “kCCR” group of features from Zeevi et al. (2014) is less generalizable than “kCCP” group of features from Lubliner et al. (2015). This is reasonable. Lubliner et al. (2015) focused on
Figure 6.5: The correlation matrix and scatter matrix for the respective features of TF Swi4p in the 8 subgroups of “fluctuation” group. Please note all 7 time points for each gene are used, and similar figures could be generated for any other TF. Please note, for all the plots about feature values, we have taken square roots of feature values to deal with skewness.
designing thousands of core promoters to study its function, while Zeevi et al. (2014) studied the conservation of hundreds of native RP gene promoters across different yeast species. Therefore, the “kCCP” group could be viewed as a refined version of “kCCR” group.

We also tested the hypothesis that the regulatory role of the core promoter could be explained by the mechanisms about TATA element’s role and nucleosome’s role within core promoter, which is proposed in section 2.3. Note in section 3.2.2, we designed the “OPCP” and “OPNT” subgroups of features for the “amplifier” group.
Figure 6.7: With “fluctuation” (“ENTF”) as the only (sub)group of features, the fitted linear model got poor results.
Figure 6.8: Replacing “kCCP” in our final model with the joint set of “OPCP” and “OPNT”.

\[ C_p = 0.410, r^2 = 0.165, \text{MSE} = 0.128 \]

\[ C_p = 0.448, r^2 = 0.199, \text{MSE} = 0.137 \]
As demonstrated in Figure 6.8, by replacing “kCCP” in our final model with the joint set of “OPCP” and “OPNT”, we got results within between the performance of a model with only “ENTF” main effects group and the final model (See Figures 6.7 and 6.1).

It is better than a model with only “ENTF” main effects group, meaning that these two group of features provide helpful information. It is worse than our final model, meaning that these two group of features contains less information than “kCCP”. Also, adding them in to “amplifier” group (represented by “kCCP”) in our final model to extend the interactions did not improve the performance, as shown in Figure 6.9. Therefore, we believe “kCCP” is partly explained by “OPCP” and “OPNT”, which are related to TATA box and nucleosome occupancy.

In summary, we believe the $k$-mer counts in “kmerCorePromtoer” subgroup represent mechanisms more than the respective roles of TBP, $-1$ nucleosome, and $+1$ nucleosome in transcriptional regulation. But how to understand them still require further exploration, even though Lubliner et al. (2015) tried to annotate them, as shown in Figure A.2.

6.5 Testing results of hypothesized mechanisms about synergistic interactions between TFs and core promoter elements

In section 2.4, we proposed that TFs and core promoter elements always work together to regulate transcript production rate, and thereby there is transcriptional synergy between them. Our final model supports this hypothesized mechanism, because it selected the interaction terms interaction terms between TF relevant features and core promoter relevant features. And given that only these interactions are selected, we believe that synergistic interactions between TFs and regulatory elements within core promoter region are crucial for cell cycle.
Figure 6.9: By adding “OPCP” and “OPNT” into “amplifier” group in our final model, which is previously represented by “kCCP” only, the model does not improve our final model: the CV mean MSE 0.070 is within one standard error of that of our final model \([0.072 - 0.005, 0.072 + 0.005]\), where we favor a simpler model. (See Figure 6.1 and Table B.10.)
We validated such a selection by trying with additive effects only,

$$y = \beta_0 + X^{(f)} \beta^{(f)} + X^{(a)} \beta^{(a)} + \epsilon. \quad (6.2)$$

As shown in Figure 6.10, such a linear function could not explain the transcript production rates well.

Moreover, we tried adding the two main effects into our final model,

$$y = \beta_0 + X^{(f)} \beta^{(f)} + X^{(a)} \beta^{(a)} + X^{(f:a)} \beta^{(f:a)} + \epsilon. \quad (6.3)$$

As shown in Figure 6.11, such a linear function does not have advantage over our final model. Same results observed for only adding each of the two main effects:

$$y = \beta_0 + X^{(f)} \beta^{(f)} + X^{(f:a)} \beta^{(f:a)} + \epsilon, \quad (6.4)$$

and

$$y = \beta_0 + X^{(a)} \beta^{(a)} + X^{(f:a)} \beta^{(f:a)} + \epsilon. \quad (6.5)$$

These results mean that, only interactions contain the important information for explaining transcript production rates over cell cycle. Adding the core promoter features as main effects does not help. Without TFs, it is known that a gene can be transcribed \textit{in vitro}. This means the existence of basal amount of transcription \textit{in vitro} with general transcription factors only. However, our model did not select the main effects group representing the core promoter’s role. And adding it did not help. A possible explanation is that such a basal amount of transcription is not promoter specific. Instead, it is a general one for all promoters. If so, our model may have included it in the intercept.

6.6 Testing results of hypothesized mechanisms about nucleosomes resisting RNAPII elongation

In our final model, none of the subgroups for “resistance”, which represents the regulatory roles of gene body nucleosomes, is selected. We discussed these selections
Figure 6.10: Our final model emphasized synergy between TFs and core promoter elements through selecting only interactions of the two associated main effects. To validate such a selection, we tried an additive function of the two main effects. Such an additive function is much worse than production function that represents synergistic interaction (See Figure 6.1).
Figure 6.11: To validate the selected synergistic interactions between TFs and core promoter elements, we tried adding the two main effects into our final model. This did not improve model performance in terms of MSE: the CV mean MSE 0.069 is within one standard error of that of our final model \( 0.072 - 0.005, 0.072 + 0.005 \), where we favor a simpler model. (See Figure 6.1 and Table B.10.)
In the following sections.

In section 2.5, we proposed that +1 nucleosomes are high barriers, that gene body nucleosomes are low barriers and their function as barriers probably decrease from 5’ to 3’, and that there is “group action” between leading RNAPII and trailing RNAPII.

In section 3.3, at first, we designed “ENNGB” feature (See Equation 3.9). Then, assuming that nucleosome’s resistance to RNAPII elongation decreased as it moved away from the TSS (and thereby TrSS), “ENNGBD” feature are proposed (See Equation 3.10). Unlike the distance decayed TF features and their associated original features, we found that the two resistance features are not highly correlated (See Figures 6.5, 6.6, and 6.12).

There could be two reasons for this: a) gene body length varies from hundreds to thousands base pairs, while native promoter length has been fixed to 600 (See Figure 6.12A); b) the distance decay weight makes nucleosomes more than 1000bp away from TSS be trivial for “ENNGBD” (See Figure 3.2). Specifically, from Figure 6.12B, we know that the max value of “ENNGB” feature is around 7, because we have taken square roots for feature values to deal with skewness. Its original value is around 51, for VPS13, which has a gene body length 9435. Also, the max value of “ENNGBD” is only about 1.4.

Our final model did not select any feature in “resistance” group. To validate our results, in two ways, we tried to add “resistance” group of features into our final model (See Equation 8.1). Note we only used “ENNGBD”, since it was deleted later than the other feature.

One way of adding it is extending the interactions in our final model,

\[
y = \beta_0 + X^{(f:avr)} \beta^{(f:avr)} + \epsilon. \tag{6.6}
\]
Figure 6.12: Probably due to the variable lengths of gene body, which is demonstrated in (A), and the exponential weight function with 250 as decay constant, the expected number of distance decay weighted gene body nucleosomes is just slightly correlated with the expected number of gene body nucleosomes, as shown in (B).
The other way is adding it into our final model as main effect,

$$y = \beta_0 + X^{(r)}\beta^{(r)} + X^{(f:a)}\beta^{(f:a)} + \epsilon.$$  \hspace{1cm} (6.7)

As shown in Figures 6.13 and 6.14, it turns out, this gene body nucleosome feature did not improve the model performance significantly.

A possible explanation for our results would be that mainly the +1 nucleosomes regulate RNAPII elongation, and the gene body nucleosome’s regulatory effect is too small to be modelled. And even the +1 nucleosome’s regulatory effects probably has been modelled by core promoter sequence features, as discussed in previous section.

Actually, after being recruited, the RNAP II initiate, pause, resume, and terminate. However, it is reported that, in the *S. cerevisiae* yeast, the density of RNAP II at promoters is roughly equivalent to that within the corresponding gene coding regions in almost all cases. Also it is known that the average Traveling Ratio value $= 0.9$ Wade and Struhl (2008). So, another possible explanation for our results could be that, at least for *S. cerevisiae* yeast, the recruited RNAP II always initiate transcription successfully, and mechanisms, like pausing, resume, and termination, do not affect the transcription significantly.
Figure 6.13: Adding “ENNGBD” into our final model by extending interactions. It did not improve our model’s performance (See Figure 6.1).
Figure 6.14: Adding “ENNGBD” into our final model as main effects. It did not improve our model’s performance (See Figure 6.1).
Preparing for predicting yeast mean transcript production rate over time for synthetic promoters

In this chapter, we prepare for predicting yeast mean transcript production rate over time for synthetic promoters. Firstly, we briefly introduce the experimental protocol about synthesizing promoter sequences, which is recently widely used for studying transcriptional regulation. Then we introduce a specific data set: the data set of DREAM6 gene promoter expression prediction challenge. This data set is about promoters of ribosomal protein (RP) genes. We describe the scoring metrics used in the challenge, which will be helpful for comparing the performance of our model with that of other models in chapter 8. We then prepare inputs for COMPETE to generate occupancy profiles and extract features.

7.1 Introduction to DREAM6 data sets

Recently, as shown in Figure 7.1, an experimental protocol has been proposed for gaining understandings on transcriptional regulation: a) synthesizing a lot of promoter sequences; b) inserting them into a fixed loci with identical downstream a
Figure 7.1: An example of the master strain into which respective research groups integrated all tested promoters (adapted from Zeevi et al. (2011)). Please note that, though it might be different among relevant data sets, there are always a gene encoding for a red fluorescent protein (mCherry) followed by a control promoter, and a gene encoding for a yellow fluorescent protein that is the reported gene for tested promoters.

yellow fluorescent protein (YFP) as reporter gene; c) measuring the in vivo YFP level as the promoter activity. With such controlled experiments, researchers have provided multiple data sets about the changes on promoter activity in response to mutating native promoter sequences (Zeevi et al., 2011; Sharon et al., 2012; Raveh-Sadka et al., 2012; Meyer et al., 2013; Rajkumar et al., 2013; Zeevi et al., 2014; Lubliner et al., 2015).

With such experimental protocol, the measurement is about the amount of YFP protein. Because the 3’ UTR of YFP coding region is always identical, we could assume the post-transcriptional regulation is identical for promoter variants and, thereby, the YFP protein production is a proxy of YFP gene’s transcript production. Therefore, the promoter activity represents the average transcript production rate of YFP gene from each promoter, per cell per second, during a certain phase, like the exponential phase, of growth. Please note, all these sets, excepting the PHO5 gene mutant promoter set (Rajkumar et al., 2013), are about ribosomal protein (RP) genes.

Aiming at a better understanding of how the regulatory elements of gene promoters enact transcriptional regulation, the DREAM6 gene promoter expression prediction challenge presented to the participants a training set consisting of 90 natural
RP promoters (Meyer et al., 2013). A test set consisting of 20 natural RP promoters plus 33 mutated RP promoters is hidden from participants for evaluating their models after their respective submissions. This data set was provided by Zeevi et al. (2011). As shown in Figures C.1 and C.2, the respective mutation types, locations, and mutated sequences of the mutant promoters are given in Zeevi et al. (2011). But according to the Supplementary Excel file of Meyer et al. (2013), the DREAM6 data set is a super set of the data set in Zeevi et al. (2011): RPL3_Mut5, RPL4A_Mut4, RPL4A_Mut5, RPL4A_Mut6 are provided in DREAM 6 only. In Meyer et al. (2013), scores of participants’ models were reported. The organizers also provided a post-hoc model combining the best team’s k-mer features with transcription factor binding affinities.

7.2 Scoring metrics in DREAM6 gene expression prediction challenge

According to Meyer et al. (2013) and Cokelaer et al. (2015), the organizers of the DREAM6 challenge exploited the following metrics for scoring participants.

The $C_p$ actually is the Pearson correlation coefficient

$$C_p = \frac{<X_{ip} \cdot \xi_i> - <X_{ip}><\xi_i>}{\sigma_{X_{ip}} \sigma_{\xi_i}}.$$  \hspace{1cm} (7.1)

where $X_{ip}$ is the participant $p$’s predicted activity for promoter $i$, and $\xi_i$ is the measured activity of promoter $i$. Please note $i = 1, 2, \ldots, N$, where $N = 53$, because there are 53 promoters in test set; and $p = 1, 2, \ldots, P$, where $P = 21$, because there are 21 teams that participated in the challenge.

The $X^2$ score is based on the $\chi^2$ metric

$$X^2 = \sum_{i=1}^{N} \frac{(X_{ip} - \xi_i)^2}{\frac{1}{P} \sum_{j=1}^{P} (X_{ij} - \xi_i)^2}.$$  \hspace{1cm} (7.2)
The $Sp$ score is the Spearman’s rank correlation coefficient

\[ Sp = \frac{\frac{1}{N} \sum_{i}^{N} R_{ip} \cdot \rho_i - \frac{1}{N} \sum_{i}^{N} R_{ip} \cdot \frac{1}{N} \sum_{i}^{N} \rho_i}{\frac{1}{N} \sum_{i}^{N} (R_{ip} - \frac{1}{N} \sum_{i}^{N} R_{ip})^2 \cdot \frac{1}{N} \sum_{i}^{N} (\rho_i - \frac{1}{N} \sum_{i}^{N} \rho_i)^2}, \]

(7.3)

where $R_{ip}$ is the rank of participant $p$’s predicted value for promoter $i$ among all of his/her $N = 53$ predicted values, therefore $1 \leq R_{ip} \leq 53$; $\rho_i$ is the rank of the measured value for promoter $i$ among all measured values for $N = 53$ promoters in test set.

The $R^2$ is a score based on the rank\(^2\) metric

\[ R^2 = \sum_{i=1}^{N} \frac{(R_{ip} - \rho_i)^2}{\frac{1}{P} \sum_{j=1}^{P} (R_{ij} - \rho_i)^2}. \]

(7.4)

7.3 Preparing inputs for COMPETE to extract features from occupancy profiles

We follow Zeevi et al. (2011) and Meyer et al. (2013) to choose Fhl1p, Sfp1p, and Rap1p as the TFs regulating RP genes. For synthetic promoter setting, the TF concentrations are set as their respective mean concentration over cell cycle or certain time duration. Here we used TF concentrations inferred from MNase data in Zhong et al. (2014).

For population data averaged over population and time, we could assume that mean TF concentration over cell cycle is representative for this average case. Consequently each promoter would have only the promoter occupancy profile for average case. But it is expected, since we have only one data point for each promoter: mean transcript production rates over population and time.

The DREAM6 challenge organizers provided DNA sequences for each promoter in training set and test set, as well as the fixed flanking sequences on either side.
Therefore, we use the given promoter sequences and flanking sequences to generate occupancy profiles (Meyer et al., 2013), exploiting COMPETE.
Applying PreTrans to yeast synthetic promoter sequences

In this chapter, we apply our framework to DREAM6 data set, to show that our modeling framework could work well for synthetic promoter sequences too. Especially, we show that our framework PreTrans beats all participant teams of DREAM6, the “best of all” pseudo model that takes the best score of all teams at each metric, and the Meyer et al. (2013) model that combined the FiRST team in DREAM6 and the Zeevi et al. (2011). We compare the gRFE path of eliminating features on DREAM6 data set with that on cell cycle data set, showing that the features selected by our models on different data sets are consistent, and thereby our algorithm works in a consistent way. Also, with DREAM6 results, our framework shows its capability to identify the relevant features for specific setting, even those features are not selected for cell cycle study. Moreover, we show the helpfulness of core promote relevant features, and discuss the difficulty of defining it. And given that such synthetic promoter data sets usually have fixed gene body, we demonstrate that COMPETE is capable of inferring influence of mutations far away and thereby is very helpful for
feature engineering.

8.1 PreTrans beats all teams participated in the DREAM6 challenge

Here we compared our model’s scores with scores of all participant teams reported in Meyer et al. (2013). We calculated the scores using DREAMTools (Cokelaer et al., 2015), which is a Python package for scoring collaborative DREAM challenges, including this DREAM6 Gene Expression Prediction Challenge.

As shown in Figure C.4, three teams ranked closely as top ones in the challenge (Meyer et al., 2013). Our team was ranked third. The FiRST team got Pearson correlation coefficient $C_p = 0.65$, while the c4lab team got $C_p = 0.5386$, as shown in Figure 8.1. All teams used only sequence features, like k-mer counts. Though it ignores the respective fluctuations of TFs’ concentrations over time and would have feature values fixed over time, an approach with sequence features only is applicable for this situation: the target variable is the promoter activity, which is the mean transcript production rates over population and time (Population exponential growth phase, specifically), and has a measurement unit as being proportional to the number of transcripts produced per cell per minute. A special approach is from the FiRST team: they only used sequence features extracted from $[−100, −1]$ of TrSS. This was strange, because there were mutations outside of the $[−100, −1]$ window resulting significant changes of expression level, as demonstrated in Figures C.1 and C.3.

To test PreTrans’ performance with our extracted features, we revisited DREAM6 data set. As shown in Figures 8.2 and 8.3, we got pretty good results. Please note the performance metrics in Figure 8.2A are obtained by aggregating the predicted values on 10 hold-out data sets and calculating them as whole, which are different from CV means of these performance metrics in those Tables. The CV means are obtained by calculating the metrics on each set respectively and taking the mean and standard deviation.
Figure 8.1: The respective plots of FiRST and c4lab in DREAM6 (adapted from Conference Slides (Meyer et al., 2013)). The comparison plot of FiRST team’s results. Promoters are sorted by absolute error. The $R$ in the figure is Pearson correlation coefficient $C_p$. The comparison plot of c4lab team’s results. Promoters are sorted by absolute error. With whole test data set, the c4lab team got $r = 0.5386$. 

\[ R = 0.65 \]

\[ 0.62 \text{ (13 cases)} \]

\[ 0.95 \text{ (31 cases)} \]
Table 8.1: The scores of PreTrans, being compared with other models. “BOA” is abbreviation for best of all teams. Cp, X2, Sp, R2 are defined in Equations 7.1, 7.2, 7.3, and 7.4 respectively. For Cp and Sp, the higher the better. For X2 and R2, the lower the better.

<table>
<thead>
<tr>
<th></th>
<th>Cp</th>
<th>X2</th>
<th>Sp</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>PreTrans</td>
<td>0.6727</td>
<td>23.2959</td>
<td>0.6848</td>
<td>24.5147</td>
</tr>
<tr>
<td>Meyer et al. (2013)</td>
<td>0.6729</td>
<td>39.7960</td>
<td>0.6682</td>
<td>30.7543</td>
</tr>
<tr>
<td>FiRST</td>
<td>0.6475</td>
<td>52.6197</td>
<td>0.6469</td>
<td>35.852</td>
</tr>
<tr>
<td>BOA</td>
<td>0.6475</td>
<td>29.7988</td>
<td>0.6469</td>
<td>35.852</td>
</tr>
</tbody>
</table>

To further show that we beat all teams participated in DREAM6, we calculated all the scores used in the DREAM6 challenge for our results, and compared with other models. As shown in Table 8.1, our scores are best among all teams, almost for every metric. For the only exception $C_p$, we got 0.6427, while the Meyer et al. (2013) got 0.6429, which is slightly better than ours. Our results are even better than a model reported in Meyer et al. (2013), which combines $k$-mer based sequence features from the FiRST team and biologically mechanistic features reported in Zeevi et al. (2011).

8.2 The gRFE path of feature elimination on DREAM6 data set

As given in Tables C.1 – C.4, our algorithms first went through Steps A and B to promoter sub-groupwise and groupwise sparsity with gRFE using ridge, and then went through Step C to compare lasso, ridge, and elastic-net results for selecting the best one. Finally a elastic-net model is chosen. The whole model selection process is based on cross-validation mean MSE, following the algorithm described in section 4.4. Comparing the gRFE path of eliminating features on DREAM6 data set with that on cell cycle data set, we figured out they are mostly identical, except “kCCR” is selected and none of subgroups in “fluctuation” group for TFs is selected. We will discuss this difference in section 8.3.
Figure 8.2: Scatter plots of PreTrans’ results on DREAM6 data set.
Figure 8.3: The comparison plot of our PreTrans results. Promoters are sorted by absolute error.
Table 8.2: Fitting models with only “kCCP”, the model performance became worse than the final model (see Table C.4).

<table>
<thead>
<tr>
<th>Features</th>
<th>Lasso</th>
<th>Ridge</th>
<th>Elastic Net</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CV Mean $C_p$: 0.529, with SE: 0.096</td>
<td>CV Mean $C_p$: 0.555, with SE: 0.092</td>
<td>CV Mean $C_p$: 0.540, with SE: 0.098</td>
</tr>
<tr>
<td></td>
<td>2. CV Mean $r^2$: 0.135, with SE: 0.144</td>
<td>2. CV Mean $r^2$: 0.126, with SE: 0.130</td>
<td>2. CV Mean $r^2$: 0.143, with SE: 0.148</td>
</tr>
<tr>
<td></td>
<td>3. CV Mean MSE: 0.216, with SE: 0.036</td>
<td>3. CV Mean MSE: 0.229, with SE: 0.043</td>
<td>3. CV Mean MSE: 0.214, with SE: 0.036</td>
</tr>
</tbody>
</table>

8.3 Interpreting our final model

Our final model selected “amplifier” group only,

$$y = \beta_0 + X^{(a)}\beta^{(a)} + \epsilon,$$

(8.1)

where $X^{(a)} = (kCCR, kCCP)$.

We firstly validated that “kCCR” is helpful for prediction. With “kCCR” excluded, the model performance significantly decreased, as given in Table 8.2. Therefore, “kCCR” is helpful for DREAM6 data set. Surprisingly, “fluctuation” group, in which features reflect TFs’ roles, is not selected.

A possible explanation would be that the “kCCR” group of features are from the Zeevi et al. (2014), which was specifically designed for study the conservation of RP gene promoter sequences across different yeast species. Given that, it is possible some of the $k$-mer sequences are about TFs regulating RP genes, though they are not annotated as TF relevant features in Zeevi et al. (2014).

It is notable that such features do not take into account information outside of the small region upstream of TrSS. Moreover, with cell cycle data, we have shown that these core promoter features themself are not capable of predicting dynamic transcript production rates over time. Given these, we believe the good performance
of models with these features resulted from the specificity of RP genes, as discussed in Zeevi et al. (2014).

8.4 COMPETE is helpful on feature engineering through providing occupancy profile

Lastly, in such synthetic promoter data sets, since the downstream reporter gene is always identical, some may assume the nucleosome occupancy in gene body to be identical for different promoters. Actually, according to Figure 8.4, we observed a relatively wide distribution of the feature values in both training and test sets. This might be resulted from the upstream promoter region difference: since DBFs compete with each other, the DBF binding at a certain location could be influenced by the sequence difference far away. This phenomenon is demonstrated in Figure 8.5, using RPL41B and its first mutant promoter as an example. We conclude that COMPETE is helpful on modeling DBF occupancy profile. At a certain genomic location, COMPETE can infer the change of occupancy profile caused by DNA sequence mutation somewhere else. Consequently, features extracted from COMPETE outputs for a genomic region, even if it is far away from the mutation site, would change. In contrast, sequence features would be identical for the same genomic region, like gene body.
Figure 8.4: The respective histograms of “ENNGBD” feature value in training set and test set.

Figure 8.5: A mutation to nucleosome disfavoring sequence at core promoter region (−83 relative to TrSS) would influence not only nucleosome occupancy at core promoter region, but also that at gene body.
Instructions for when, where and to what level each gene should be expressed are encoded within the genome. However, even decoding transcriptional regulation only remains a major challenge in biology. In this work, we devised a machine learning framework for learning models to predict transcript production rates. We showed our framework could learn highly predictive models in predicting dynamic transcript production rates over cell cycle, and in predicting average transcript production rates during certain phase for synthetic promoter sequences. Moreover, we showed that, by integrating features reflecting knowledges learned in specific settings, our framework could identify generalizable features, and suggest insights about the underlying mechanisms.

9.1 Our contributions

9.1.1 First successful attempt to have a dynamic model of transcript production and transcriptional regulation

As far as we know, our framework is the first successful attempt to have a model of dynamic transcript production and transcriptional regulation: with cell cycle data...
set, we got Pearson correlation coefficient $C_p = 0.751$ and coefficient of determination $r^2 = 0.564$ on test set for predicting dynamic transcript production rate over time; with DREAM6 Gene Promoter Expression Prediction challenge, we outperformed all participant teams, best of all teams, and a combined model of best team and biological mechanistic features, in terms of almost all scoring metrics.

There are two other works have settings similar to our setting of predicting dynamic transcript production rates over time, since they both have changing TF concentrations over space (Segal et al., 2008; He et al., 2010). In *Drosophila* embryonic development (anterior-posterior axis specification), researchers have tried to predict transcript production rates in response to changing TF concentrations over space, using TF relevant features only, like expected number of TF binding instances and expected number of TF binding instances weighted by distance to ORF 5′ boundary (Translation Start Site, TrSS). However, even if they were to add higher-order TF relevant features, like interactions between TFs and distances between TFs (He et al., 2010), they have thus far failed to be highly predictive. The main reason of lacking enough accuracy in these works might be that elements in other regulatory regions, especially core promoter, are assumed constant, and we showed the importance of core promoter for predicting transcript production rates in both cell cycle data set and DREAM6 synthetic promoter set.

9.1.2 A framework that can test mechanistic hypotheses and identify generalizable features

Moreover, our model demonstrated its capability of identifying generalizable features by interpreting the highly predictive models, and thereby provide support for associated hypothesized mechanisms. Specifically, the features selected by our models provided support for the following biological insights:

1. in section 6.3, our results support the hypothesis that the roles of activators
and repressors TFs played are usually determined by transcription cofactors;

2. in section 6.3, our results support the hypothesis that different TF’s contribution to transcript production might be distinct and non-redundant from each other;

3. in section 6.3.1, our results provide support for the hypothesis that TFs govern the probability of RNAPII recruitment and initiation, possibly through interactions with PIC components and transcription cofactors;

4. in section 6.4.1, our results provide support for the hypothesis that the core promoter amplifies the transcript production probably by influencing PIC formation, RNAPII recruitment, DNA melting, RNAPII searching for and selecting TSS, releasing RNAPII from general transcription factors and thereby initiation;

5. in section 6.5, our results provide support for the hypothesis that there is strong transcriptional synergy between TFs and regulatory elements within the core promoter, which very likely represent respective DNA sequence signals for recruiting general transcription factors and transcription cofactors, and for TSS scanning and selection;

6. in section 6.4.2, our results provide support for the hypothesis that the regulatory elements within core promoter region are more than TATA box and nucleosome free region, suggesting the existence of still unidentified TAF-dependent core promoter elements in yeast *S. cerevisiae*;

7. in section 6.4.2, our results provide support for the hypothesis that nucleosome occupancy profile is helpful for representing −1 and +1 nucleosomes’ regulatory roles on transcription.
Moreover, there are also several hypothesized mechanisms for which our results did not provide support, for which we discussed the possible reasons. In section 6.3.2, we discussed why modeling nucleosomes compete with TFs for DNA binding site is not helpful for predicting transcript production rates. In section 6.3.3, we discussed why stability weight, saturation, and distance decay are not helpful for predictive modeling of transcript production rates in cell cycle. In section 6.6, we discussed why modeling nucleosomes resisting RNAPII elongation does not improve predictive performance of our models.

Also, in section 8.3, we discussed why “kCCR” is helpful for DREAM6 data set, while “fluctuation” group, in which features reflect TFs’ roles, is not selected at all. We believe that such selection resulted from the specificity of RP genes, as discussed in Zeevi et al. (2014).

9.1.3 Presenting the synergistic interactions between TFs and core promoter elements

There has been no work modelling the relationship between TFs and core promoter elements. As far as we know, our model is the first one presenting the synergistic interactions between TFs and core promoter elements with highly predictive model.

Relevant works either have core promoter fixed or assumed as constant (Segal et al., 2008; Raveh-Sadka et al., 2009; He et al., 2010; MacIsaac et al., 2010; Rajkumar et al., 2013), or do not model the changes in TF availabilities over time (Zeevi et al., 2011; Sharon et al., 2012; Raveh-Sadka et al., 2012; Meyer et al., 2013; Zeevi et al., 2014; Lubliner et al., 2015). The closest work probably is Eser et al. (2014). They observed that some cell cycle regulators share similar shape of time course with their respective target genes for TFs Swi4p, Yox1p, and Mbp1p (see Figure 2.5), and that PIC components measurements from Rhee and Pugh (2012) are correlated with time average transcript production rates (see Figure 2.8). Therefore, they proposed that
TFs control the timing of transcript production, while the core promoter governs the magnitude. However, Eser et al. (2014) did not provide a computational model. Also, measurements of PIC components are only available from experiments: we are able to infer them from neither DNA sequence nor occupancy profiles. Lastly, with their work, we were still not sure that whether there is only additive function between TFs and the core promoter, or there are both additive function and synergy, or there is only synergy.

With Eser et al. (2014) data set, our work showed synergistic interactions between TFs and regulatory elements within core promoter region are crucial for cell cycle (see section 6.5). The transcriptional synergy might be general for TFs and core promoter elements, though it is hard to tell which TFs each core promoter element is associated with.

9.2 Improving our framework

9.2.1 Improving the application of our framework in yeast S. cerevisiae

Some improvements may be made to our framework that might boost its predictive accuracy without fundamentally altering its structure or effectiveness.

Understanding elongation, termination, and the hypothesized mechanisms for which our results could not provide support, will probably improve our framework, if we could design and incorporate new features that are promising.

As discussed in chapter 2, we ignored several hypothesized mechanisms, because of various reasons. For those hypothesized mechanisms that we did not model due to missing data, modeling them might be helpful once we have enough data. For example, TF binding site flanks might be helpful for predicting TF binding probability (Levo et al., 2015). We could get better PSSMs for each studied TF by incorporating sequences flanking core TF binding sites. As a result, we could get better results on predicting TF binding probabilities so that we would get better results on predicting
transcript production rates.

Most importantly, in the future we need a further exploration about the regulatory elements within core promoter so that we could relate the sequence features to PIC formation, RNAPII recruitment, PIC scanning for TSS, and transcriptional initiation in an explicit way. Such exploration would require the advancement of experimental techniques and could bring in better features for modeling. Especially, as discussed in section 2.3, it is suggested that there are unidentified TAF-dependent and other cofactor-dependent core promoter motifs (regulatory elements) in yeast.

9.2.2 Improving our framework for its application to metazoans

Our work is all about yeast, because it is the minimal model eukaryote. But this does not mean our framework could not be applied to higher order eukaryotes. We could exploit our framework for any metazoan, as long as we have DBF binding specificities, DBF concentrations, and DNA sequences as inputs. However, we also need to take care the additional mechanisms we will have in metazoans.

Basically, all the simplicities we took for modelling with yeast *S. cerevisiae* would become the obstacles we need to tackle for metazoans. As discussed in section 1.7, we will need to take care of enhancers and DNA looping, as well as RNA alternative splicing. Given that these areas are extremely active and the community keeps making progress, we believe in the near future we will be successfully improve our framework to apply it to metazoans, by incorporating mechanisms about enhancers and DNA looping, as well as RNA alternative splicing.
Appendix A

$k$-mer count sequence features extracted from the core promoter
Figure A.1: “kmerConservation” subgroup of k-mer features Zeevi et al. (2014) (adapted from Zeevi et al. (2014)). The authors fitted 10 linear models 10 different partitions of the data into training and held-out test sets. A feature is included in the list shown in Part (A), only if it was selected in 8 out of the 10 fitted linear models. This list of features include k-mer existence and counts, features of hits of PSSMs of known RP regulators, and a feature of the predicted intrinsic nucleosome occupancy. Then based on the features’ locations on the promoter and their effect on expression, the authors grouped 25 out of 38 robust features, and provided a schematic representation of the groups of features, as shown in Part (B). Please note we excluded features 8, 14, 15, 22, 29, 30, 31, 32, 33, 37, and 38 in our work. These eleven features are about TF binding, TBP binding, and nucleosome occupancy. We assume that COMPETE is better on this.
Figure A.2: “kmerCorePromoter” subgroup of $k$-mer features reported in Lubliner et al. (2015) (adapted from Lubliner et al. (2015)). The authors fitted 10 linear models 10 different partitions of the data to training and held-out test sets. A feature is included in the list shown in Part (A) only if it was selected in 6 out of the 10 fitted linear models. The authors then manually classified them according to the associated window’s location in core promoter, and its base content, as shown in Part (B).
Appendix B
Supplement for cell cycle study

Table B.1: The *Pearson* correlation coefficient between our “ENTCP” and the features in TATA element class of “kCCP” subgroup. They are sorted in ascending order.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Pearson Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1_ATAAA_counts_m118_2_m61_Conensus_TATA</td>
<td>0.238</td>
</tr>
<tr>
<td>2_TATA_counts_m118_2_m91_Conensus_TATA</td>
<td>0.250</td>
</tr>
<tr>
<td>3_TATAA_counts_m118_2_m61_Conensus_TATA</td>
<td>0.352</td>
</tr>
</tbody>
</table>
Table B.2: The Pearson correlation coefficient between our “PONOCP” and the features in the initiation class of “kCCP” subgroup.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>44. TATGT_1mismatch_counts_m30_2_m1_Initiation_T-rich</td>
<td>-0.091</td>
</tr>
<tr>
<td>4. ACGAA_1mismatch_counts_m30_2_m1_Initiation_A-rich</td>
<td>0.000</td>
</tr>
<tr>
<td>43. TGTTT_1mismatch_counts_m30_2_m1_Initiation_T-rich</td>
<td>0.009</td>
</tr>
<tr>
<td>29. CAAGA_1mismatch_counts_m45_2_m16_Initiation_A-rich</td>
<td>0.030</td>
</tr>
<tr>
<td>25. GAAAG_1mismatch_counts_m30_2_m1_Initiation_A-rich</td>
<td>0.097</td>
</tr>
<tr>
<td>10. CAAAG_1mismatch_counts_m45_2_m16_Initiation_A-rich</td>
<td>0.100</td>
</tr>
</tbody>
</table>

Table B.3: The Pearson correlation coefficient between our “MONOCP” and the features in the TATA element, PIC, and scanning classes of “kCCP” subgroup.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. TATAA_counts_m118_2_m61_Consensus_TATA</td>
<td>-0.272</td>
</tr>
<tr>
<td>2. TATA_counts_m118_2_m91_Consensus_TATA</td>
<td>-0.240</td>
</tr>
<tr>
<td>9. ATACA_1mismatch_counts_m105_2_m76_PIC_A/T-rich</td>
<td>-0.162</td>
</tr>
<tr>
<td>1. ATAAA_counts_m118_2_m61_Consensus_TATA</td>
<td>-0.154</td>
</tr>
<tr>
<td>16. CAGAT_1mismatch_counts_m75_2_m46_PIC_A/T-rich</td>
<td>-0.074</td>
</tr>
<tr>
<td>21. TACAA_1mismatch_counts_m90_2_m61_PIC_A/T-rich</td>
<td>-0.073</td>
</tr>
<tr>
<td>24. TCTGA_1mismatch_counts_m90_2_m31_Scanning_T/C-rich</td>
<td>-0.068</td>
</tr>
<tr>
<td>14. TACTT_1mismatch_counts_m90_2_m61_Scanning_T/C-rich</td>
<td>-0.039</td>
</tr>
<tr>
<td>8. TATCT_1mismatch_counts_m105_2_m76_PIC_A/T-rich</td>
<td>-0.033</td>
</tr>
<tr>
<td>22. TTCA_1mismatch_counts_m75_2_m46_Scanning_T/C-rich</td>
<td>-0.018</td>
</tr>
<tr>
<td>5. TTCTT_1mismatch_counts_m90_2_m31_Scanning_T/C-rich</td>
<td>-0.015</td>
</tr>
<tr>
<td>12. TCTT_1mismatch_counts_m90_2_m31_Scanning_T/C-rich</td>
<td>-0.013</td>
</tr>
<tr>
<td>13. TCTG_1mismatch_counts_m90_2_m31_Scanning_T/C-rich</td>
<td>0.003</td>
</tr>
<tr>
<td>15. TCGTT_1mismatch_counts_m90_2_m31_Scanning_T/C-rich</td>
<td>0.003</td>
</tr>
<tr>
<td>18. ATCTA_1mismatch_counts_m45_2_m16_Scanning_T/C-rich</td>
<td>0.013</td>
</tr>
<tr>
<td>17. TTGTGT_1mismatch_counts_m60_2_m31_Scanning_T/C-rich</td>
<td>0.015</td>
</tr>
<tr>
<td>19. TCT_counts_m90_2_m31_Scanning_T/C-rich</td>
<td>0.028</td>
</tr>
<tr>
<td>11. TCTCA_1mismatch_counts_m90_2_m31_Scanning_T/C-rich</td>
<td>0.040</td>
</tr>
<tr>
<td>26. TCTAT_1mismatch_counts_m45_2_m16_Scanning_T/C-rich</td>
<td>0.041</td>
</tr>
<tr>
<td>27. TTCA_1mismatch_counts_m118_2_m1_Scanning_T/C-rich</td>
<td>0.075</td>
</tr>
</tbody>
</table>
Table B.4: The *Pearson* correlation coefficient between our “PONOCP” and the features in “kCCP” subgroup, which could not be assigned into any of TATA element, PIC, scanning, and initiation classes.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Mismatch Count</th>
<th>Core Promoter Type</th>
<th>C-rich Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGGCG</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>-0.133</td>
</tr>
<tr>
<td>TAGGC</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>-0.121</td>
</tr>
<tr>
<td>CGGCCG</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>-0.098</td>
</tr>
<tr>
<td>AGCCGC</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>-0.080</td>
</tr>
<tr>
<td>ACTAG</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>-0.060</td>
</tr>
<tr>
<td>ATAG</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>-0.060</td>
</tr>
<tr>
<td>GAGTA</td>
<td>1 mismatch</td>
<td>Core Promoter Null</td>
<td>-0.040</td>
</tr>
<tr>
<td>ACAGG</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>-0.021</td>
</tr>
<tr>
<td>CGTG</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>-0.012</td>
</tr>
<tr>
<td>TCGGC</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>-0.010</td>
</tr>
<tr>
<td>GGGCA</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>-0.004</td>
</tr>
<tr>
<td>CACCG</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>0.001</td>
</tr>
<tr>
<td>TCTAG</td>
<td>1 mismatch</td>
<td>Core Promoter Null</td>
<td>0.001</td>
</tr>
<tr>
<td>CGCCG</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>0.002</td>
</tr>
<tr>
<td>GGCCG</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>0.002</td>
</tr>
<tr>
<td>GGGCT</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>0.002</td>
</tr>
<tr>
<td>GGGC</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>0.008</td>
</tr>
<tr>
<td>GCGT</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>0.019</td>
</tr>
<tr>
<td>CCGG</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>0.025</td>
</tr>
<tr>
<td>AGCGT</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>0.028</td>
</tr>
<tr>
<td>AGCTA</td>
<td>1 mismatch</td>
<td>Core Promoter Null</td>
<td>0.031</td>
</tr>
<tr>
<td>CCGTG</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>0.033</td>
</tr>
<tr>
<td>ATGG</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>0.039</td>
</tr>
<tr>
<td>GAGCG</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>0.050</td>
</tr>
<tr>
<td>CGTG</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>0.055</td>
</tr>
<tr>
<td>CATGG</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>0.065</td>
</tr>
<tr>
<td>TGAC</td>
<td>1 mismatch</td>
<td>Core Promoter Null</td>
<td>0.066</td>
</tr>
<tr>
<td>CTA</td>
<td>1 mismatch</td>
<td>Core Promoter Null</td>
<td>0.070</td>
</tr>
<tr>
<td>CGGGC</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>0.071</td>
</tr>
<tr>
<td>CAACC</td>
<td>1 mismatch</td>
<td>Core Promoter Null</td>
<td>0.082</td>
</tr>
<tr>
<td>GGTGG</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>0.097</td>
</tr>
<tr>
<td>GGCT</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>0.098</td>
</tr>
<tr>
<td>CAA</td>
<td>1 mismatch</td>
<td>Core Promoter Null</td>
<td>0.099</td>
</tr>
<tr>
<td>CATGG</td>
<td>1 mismatch</td>
<td>Core Promoter G</td>
<td>0.099</td>
</tr>
<tr>
<td>G</td>
<td>C content</td>
<td>Core Promoter G</td>
<td>0.160</td>
</tr>
</tbody>
</table>
Table B.5: The Pearson correlation coefficient between our “MONOCP” and the features in “kCCP” subgroup, which could not be assigned into any of TATA element, PIC, scanning, and initiation classes.

<table>
<thead>
<tr>
<th>Feature</th>
<th>CorePromoter</th>
<th>C-rich</th>
<th>C-rich</th>
<th>C-rich</th>
<th>C-rich</th>
<th>C-rich</th>
<th>C-rich</th>
<th>C-rich</th>
<th>C-rich</th>
</tr>
</thead>
<tbody>
<tr>
<td>30_CGTG_1mismatch_counts</td>
<td>-0.162</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>59_GGGCT_1mismatch_counts</td>
<td>-0.119</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38_GCTGG_1mismatch_counts</td>
<td>-0.106</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65_G_C_content</td>
<td>-0.102</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28_AGCTA_1mismatch_counts</td>
<td>-0.076</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52_CGTGG_1mismatch_counts</td>
<td>-0.074</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45_TAGGC_1mismatch_counts</td>
<td>-0.057</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34_CGCCG_1mismatch_counts</td>
<td>-0.055</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55_ACTAG_1mismatch_counts</td>
<td>-0.054</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42_ATGG_1mismatch_counts</td>
<td>-0.050</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>61_CCGG_1mismatch_counts</td>
<td>-0.046</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>57_ATAG_1mismatch_counts</td>
<td>-0.042</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>62_GCGT_1mismatch_counts</td>
<td>-0.034</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>49_GGCT_1mismatch_counts</td>
<td>-0.031</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40_CAACCG_1mismatch_counts</td>
<td>-0.031</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32_CGAC_1mismatch_counts</td>
<td>-0.026</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>63_GGCCG_1mismatch_counts</td>
<td>-0.019</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23_GAGTA_1mismatch_counts</td>
<td>-0.018</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33_CATGG_1mismatch_counts</td>
<td>-0.016</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>58_CATGG_1mismatch_counts</td>
<td>-0.011</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36_GGACG_1mismatch_counts</td>
<td>-0.005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>46_ACACGG_1mismatch_counts</td>
<td>-0.005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>48_TCGCC_1mismatch_counts</td>
<td>0.003</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37_CCCTG_1mismatch_counts</td>
<td>0.003</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31_TCTAG_1mismatch_counts</td>
<td>0.005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50_ACGTG_1mismatch_counts</td>
<td>0.007</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>51_GGCGC_1mismatch_counts</td>
<td>0.007</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39_CTA_1mismatch_counts</td>
<td>0.013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6_AGCGAG_1mismatch_counts</td>
<td>0.015</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>54_GGGG_1mismatch_counts</td>
<td>0.016</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64_GCGGGC_1mismatch_counts</td>
<td>0.025</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41_CGCCCG_1mismatch_counts</td>
<td>0.025</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>53_TGAC_1mismatch_counts</td>
<td>0.028</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7_CAACG_1mismatch_counts</td>
<td>0.031</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35_CGCCG_1mismatch_counts</td>
<td>0.054</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>47_CGGCGG_1mismatch_counts</td>
<td>0.055</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20_CAA_1mismatch_counts</td>
<td>0.061</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>56_CGCC_1mismatch_counts</td>
<td>0.112</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table B.6: Iterations of our algorithm on Eser et al. (2014) cell cycle data set (Step A, Part 1). Please see section 4.4 for details of our algorithm.

<table>
<thead>
<tr>
<th></th>
<th>Best Model</th>
<th>Current Model</th>
<th>Decision</th>
</tr>
</thead>
</table>
| A1 | N/A        | 1. CV Mean $C_p$: 0.781, with SE: 0.020  
          2. CV Mean $r^2$: 0.595, with SE: 0.039  
          3. CV Mean MSE: 0.062, with SE: 0.006 | Remove two subgroups from “amplifier”:  
          1. “OPCP”: 0.12509(-95.70%) 225;  
          2. “OPNT”: 0.11929(-95.90%) 225;  
          Feature number: 7349 → 6899 |
| A2 | 1. CV Mean $C_p$: 0.778, with SE: 0.021  
          2. CV Mean $r^2$: 0.591, with SE: 0.040  
          3. CV Mean MSE: 0.063, with SE: 0.006 | Remove a subgroup from “amplifier”:  
          1. “kCCR”: 1.23065(-60.36%) 2025;  
          Feature number: 6899 → 4874 |
| A3 | 1. CV Mean $C_p$: 0.778, with SE: 0.021  
          2. CV Mean $r^2$: 0.591, with SE: 0.040  
          3. CV Mean MSE: 0.063, with SE: 0.006 | Remove a subgroup from “resistance”:  
          1. “ENNGB”: 1.22796(-10.70%) 1625;  
          Feature number: 4874 → 3249 |
Table B.7: Iterations of our algorithm on Eser et al. (2014) cell cycle data set (Continued: Step A-B). At the end of Step A5, our algorithm transited to Step B.

<table>
<thead>
<tr>
<th></th>
<th>Best Model</th>
<th>Current Model</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A4</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1. CV Mean $C_p$: 0.770, with SE: 0.021</td>
<td>1. CV Mean $C_p$: 0.753, with SE: 0.022</td>
<td>Remove a subgroup from “fluctuation”:</td>
</tr>
<tr>
<td></td>
<td>2. CV Mean $r^2$: 0.583, with SE: 0.038</td>
<td>2. CV Mean $r^2$: 0.560, with SE: 0.037</td>
<td>1. “POTF”: 1.66679(-1.09%) 1560;</td>
</tr>
<tr>
<td></td>
<td>3. CV Mean MSE: 0.064, with SE: 0.005</td>
<td>3. CV Mean MSE: 0.067, with SE: 0.005</td>
<td>Feature number: 3249 → 1689</td>
</tr>
<tr>
<td><strong>A5</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1. CV Mean $C_p$: 0.753, with SE: 0.022</td>
<td>1. CV Mean $C_p$: 0.760, with SE: 0.019</td>
<td>Stop A. Go to B. Remove these two groups:</td>
</tr>
<tr>
<td></td>
<td>2. CV Mean $r^2$: 0.560, with SE: 0.037</td>
<td>2. CV Mean $r^2$: 0.550, with SE: 0.041</td>
<td>1. “fluctuation × resistance”: 0.02555(-99.11%) 12; 0.0021295(-52.16%) 12</td>
</tr>
<tr>
<td></td>
<td>3. CV Mean MSE: 0.067, with SE: 0.005</td>
<td>3. CV Mean MSE: 0.069, with SE: 0.006</td>
<td>2. “resistance”: 0.00026(-99.99%) 1; 0.0002582(-94.20%) 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Feature number: 1689 → 1676</td>
</tr>
<tr>
<td><strong>B1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1. CV Mean $C_p$: 0.760, with SE: 0.019</td>
<td>1. CV Mean $C_p$: 0.760, with SE: 0.019</td>
<td>Remove a group:</td>
</tr>
<tr>
<td></td>
<td>2. CV Mean $r^2$: 0.550, with SE: 0.041</td>
<td>2. CV Mean $r^2$: 0.550, with SE: 0.042</td>
<td>1. “amplifier × fluctuation × resistance”: 2.69408(-5.78%) 768; 0.0035079(-21.19%) 768</td>
</tr>
<tr>
<td></td>
<td>3. CV Mean MSE: 0.069, with SE: 0.006</td>
<td>3. CV Mean MSE: 0.069, with SE: 0.006</td>
<td>Feature number: 1676 → 908</td>
</tr>
</tbody>
</table>
Table B.8: Iterations of our algorithm on Eser et al. (2014) cell cycle data set (Continued: Step B).

<table>
<thead>
<tr>
<th></th>
<th>Best Model</th>
<th>Current Model</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>B2</td>
<td>1. CV Mean $C_p$: 0.760, with SE: 0.019</td>
<td>1. CV Mean $C_p$: 0.756, with SE: 0.021</td>
<td>Remove a group:</td>
</tr>
<tr>
<td></td>
<td>2. CV Mean $r^2$: 0.550, with SE: 0.042</td>
<td>2. CV Mean $r^2$: 0.552, with SE: 0.041</td>
<td>1. “fluctuation”: 0.03931(-99.08%) 12; 0.0032755(-41.32%) 12</td>
</tr>
<tr>
<td></td>
<td>3. CV Mean MSE: 0.069, with SE: 0.006</td>
<td>3. CV Mean MSE: 0.069, with SE: 0.006</td>
<td>Feature number: 908 → 896</td>
</tr>
<tr>
<td>B3</td>
<td>1. CV Mean $C_p$: 0.756, with SE: 0.021</td>
<td>1. CV Mean $C_p$: 0.755, with SE: 0.021</td>
<td>Remove a group:</td>
</tr>
<tr>
<td></td>
<td>2. CV Mean $r^2$: 0.552, with SE: 0.041</td>
<td>2. CV Mean $r^2$: 0.551, with SE: 0.041</td>
<td>1. “amplifier × resistance”: 0.32381(-92.46%) 64; 0.0050595(-9.50%) 64</td>
</tr>
<tr>
<td></td>
<td>3. CV Mean MSE: 0.069, with SE: 0.006</td>
<td>3. CV Mean MSE: 0.069, with SE: 0.006</td>
<td>Feature number: 896 → 832</td>
</tr>
<tr>
<td>B4</td>
<td>1. CV Mean $C_p$: 0.755, with SE: 0.021</td>
<td>1. CV Mean $C_p$: 0.751, with SE: 0.021</td>
<td>Remove a group:</td>
</tr>
<tr>
<td></td>
<td>2. CV Mean $r^2$: 0.551, with SE: 0.041</td>
<td>2. CV Mean $r^2$: 0.546, with SE: 0.041</td>
<td>1. “amplifier × fluctuation”: 4.42001(0.00%) 768; 0.0057552(-14.74%) 768</td>
</tr>
<tr>
<td></td>
<td>3. CV Mean MSE: 0.069, with SE: 0.006</td>
<td>3. CV Mean MSE: 0.069, with SE: 0.006</td>
<td>Feature number: 832 → 64</td>
</tr>
</tbody>
</table>
Table B.9: Iterations of our algorithm on Eser et al. (2014) cell cycle data set (Continued: Step B-C).

<table>
<thead>
<tr>
<th>Decision</th>
<th>1. CV Mean $C_p$: 0.751, with SE: 0.021</th>
<th>1. CV Mean $C_p$: 0.388, with SE: 0.020</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2. CV Mean $r^2$: 0.546, with SE: 0.041</td>
<td>2. CV Mean $r^2$: 0.148, with SE: 0.017</td>
</tr>
<tr>
<td></td>
<td>3. CV Mean MSE: 0.069, with SE: 0.006</td>
<td>3. CV Mean MSE: 0.131, with SE: 0.005</td>
</tr>
<tr>
<td></td>
<td>Add the group back:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1. “amplifier $\times$ fluctuation”: 4.42001(0.00%) 768; 0.0057552(-14.74%) 768</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Remove the other group:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1. “amplifier”: 0.43201(-90.23%) 64; 0.0067501(0.00%) 64</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Feature number: 64 → 832 → 768</td>
<td></td>
</tr>
</tbody>
</table>

Table B.10: Iterations of our algorithm on Eser et al. (2014) cell cycle data set (Continued: Step C).

<table>
<thead>
<tr>
<th>Decision</th>
<th>1. CV Mean $C_p$: 0.733, with SE: 0.021</th>
<th>1. CV Mean $C_p$: 0.740, with SE: 0.021</th>
<th>1. CV Mean $C_p$: 0.737, with SE: 0.023</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2. CV Mean $r^2$: 0.521, with SE: 0.038</td>
<td>2. CV Mean $r^2$: 0.527, with SE: 0.042</td>
<td>2. CV Mean $r^2$: 0.530, with SE: 0.041</td>
</tr>
<tr>
<td></td>
<td>3. CV Mean MSE: 0.073, with SE: 0.005</td>
<td>3. CV Mean MSE: 0.072, with SE: 0.005</td>
<td>3. CV Mean MSE: 0.072, with SE: 0.005</td>
</tr>
<tr>
<td></td>
<td>Choose Elastic Net</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

161
Appendix C

Supplement for DREAM6 study
Figure C.1: Detailed view of the types and locations of mutant RP promoters in DREAM6 (adapted from Zeevi et al. (2011)). The same figure was used in Meyer et al. (2013) as well. Please note DREAM6 has 4 promoters missing information in this figure: RPL3_Mut5, RPL4A_Mut4, RPL4A_Mut5, and RPL4A_Mut6, all of which are about random mutations, according to the Supplementary Table of Meyer et al. (2013).
Figure C.2: Detailed List of the types, locations, and mutated sequences of mutant RP promoters in DREAM6 (adapted from Zeevi et al. (2011)). Please note DREAM6 still has those 4 promoters missing information in this list. Also, the location is relative to translation start site (TrSS). And the NDS means nucleosome disfavoring sequence. A mutation type means that it is a mutation on that type of sequence. For example, a NDS mutation, is a mutation to a sequence which is nucleosome disfavoring before mutation.
Figure C.3: Types of mutations and their effects on RP gene expression (adapted from Zeevi et al. (2011)). For each promoter, shown is activity of the natural promoter and a promoter in which a mutation was executed. Error bars represent two standard errors computed from 24 replicates. The magnitude of the effect of the mutation on promoter activity is indicated above the activity bars of each promoter pair, where two stars mark promoter activity differences that are statistically significant. Here promoter activity is the mean transcript production rates over population exponential growth phase. A) random mutation (18 bp sequence changes). B) Mutations of Fhl1p sites (2bp changes that preserved the G/C content). C) Mutations of Sfp1p sites (2bp changes that preserved the G/C content). D) Mutations of TATA box (2-3bp changes that preserved the G/C content). E) Mutations of A/T-rich nucleosome disfavoring sequences (replacing 16 A/T base pairs with G/C base pairs in a region of 31 base pairs within the promoter that had the lowest predicted nucleosome occupancy (Kaplan et al., 2009)). F) Summary and comparison of the effect of the mutations according to the mutation type.
Figure C.4: Only names of the two best-performing teams are given. Our team was the 3rd one. Cp (see Equation 7.1) indicates the Pearson correlation coefficient, $X^2$ is a score based on the $\chi^2$ metric (see Equation 7.2), Sp is the Spearman’s rank correlation coefficient (see Equation 7.3), and $R^2$ the score based on the rank$^2$ metric (see Equation 7.4). For Cp and Sp, the higher the score is, the better the team’s predictions are. For $X^2$ and $R^2$, the lower the better (This Figure is adapted from Meyer et al. (2013)).

<table>
<thead>
<tr>
<th>Rank</th>
<th></th>
<th>Cp</th>
<th>$X^2$</th>
<th>Sp</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FiRST</td>
<td>1</td>
<td>0.6475</td>
<td>52.6197</td>
<td>0.6469</td>
<td>35.852</td>
</tr>
<tr>
<td>c4lab</td>
<td>2</td>
<td>0.5386</td>
<td>30.8202</td>
<td>0.4938</td>
<td>37.7716</td>
</tr>
<tr>
<td>Team263</td>
<td>3</td>
<td>0.5184</td>
<td>29.7988</td>
<td>0.4436</td>
<td>37.1604</td>
</tr>
<tr>
<td>Team164</td>
<td>4</td>
<td>0.4925</td>
<td>35.1053</td>
<td>0.4837</td>
<td>40.5142</td>
</tr>
<tr>
<td>Team259</td>
<td>5</td>
<td>0.4959</td>
<td>31.3914</td>
<td>0.456</td>
<td>48.8889</td>
</tr>
<tr>
<td>Team140</td>
<td>6</td>
<td>0.534</td>
<td>61.2089</td>
<td>0.5887</td>
<td>47.7112</td>
</tr>
<tr>
<td>Team250</td>
<td>7</td>
<td>0.4743</td>
<td>40.7905</td>
<td>0.5262</td>
<td>50.29</td>
</tr>
<tr>
<td>Team84</td>
<td>8</td>
<td>0.5253</td>
<td>65.7449</td>
<td>0.4732</td>
<td>37.8666</td>
</tr>
<tr>
<td>Team17</td>
<td>9</td>
<td>0.5211</td>
<td>69.0551</td>
<td>0.477</td>
<td>41.6918</td>
</tr>
<tr>
<td>Team154</td>
<td>10</td>
<td>0.4457</td>
<td>55.4508</td>
<td>0.4901</td>
<td>40.4362</td>
</tr>
<tr>
<td>Team21</td>
<td>11</td>
<td>0.4197</td>
<td>57.103</td>
<td>0.5661</td>
<td>49.2351</td>
</tr>
<tr>
<td>Team76</td>
<td>12</td>
<td>0.469</td>
<td>50.4026</td>
<td>0.4598</td>
<td>42.6173</td>
</tr>
<tr>
<td>Team61</td>
<td>13</td>
<td>0.5327</td>
<td>55.4348</td>
<td>0.4514</td>
<td>46.5912</td>
</tr>
<tr>
<td>Team187</td>
<td>14</td>
<td>0.5467</td>
<td>61.9649</td>
<td>0.5064</td>
<td>55.8043</td>
</tr>
<tr>
<td>Team257</td>
<td>15</td>
<td>0.502</td>
<td>47.6773</td>
<td>0.4547</td>
<td>50.6775</td>
</tr>
<tr>
<td>Team253</td>
<td>16</td>
<td>0.4478</td>
<td>91.0982</td>
<td>0.4386</td>
<td>50.7745</td>
</tr>
<tr>
<td>Team264</td>
<td>17</td>
<td>0.3278</td>
<td>44.0867</td>
<td>0.2259</td>
<td>77.3447</td>
</tr>
<tr>
<td>Team245</td>
<td>18</td>
<td>0.3031</td>
<td>46.4831</td>
<td>0.2973</td>
<td>60.9865</td>
</tr>
<tr>
<td>Team265</td>
<td>19</td>
<td>0.3932</td>
<td>50.2789</td>
<td>0.3729</td>
<td>81.9649</td>
</tr>
<tr>
<td>Team176</td>
<td>20</td>
<td>0.2658</td>
<td>47.6607</td>
<td>0.2385</td>
<td>72.0592</td>
</tr>
<tr>
<td>Team138</td>
<td>21</td>
<td>0.1279</td>
<td>88.8242</td>
<td>0.0928</td>
<td>106.762</td>
</tr>
</tbody>
</table>
Table C.1: Iterations of our algorithm on DREAM6 data set (Step A, Part 1). Please see Section 4.4 for details of our algorithm.

<table>
<thead>
<tr>
<th>A1</th>
<th>Best Model</th>
<th>Current Model</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N/A</td>
<td></td>
<td>Remove two subgroups from “amplifier”:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1. CV Mean $C_p$: 0.738, with SE: 0.054</td>
<td>1. “OPCP”: 0.00520(-95.68%) 63;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. CV Mean $r^2$: 0.344, with SE: 0.126</td>
<td>2. “OPNT”: 0.00463(-96.15%) 63;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. CV Mean MSE: 0.160, with SE: 0.030</td>
<td>Feature number: 2057 → 1931</td>
</tr>
<tr>
<td>A2</td>
<td></td>
<td>1. CV Mean $C_p$: 0.738, with SE: 0.054</td>
<td>Remove a subgroup from “amplifier”:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. CV Mean $r^2$: 0.344, with SE: 0.126</td>
<td>1. “kCCR”: 0.07823(-37.16%) 567;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. CV Mean MSE: 0.160, with SE: 0.030</td>
<td>Feature number: 1931 → 1364</td>
</tr>
<tr>
<td>A3</td>
<td></td>
<td>1. CV Mean $C_p$: 0.735, with SE: 0.054</td>
<td>Add the subgroup back into “amplifier”:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. CV Mean $r^2$: 0.345, with SE: 0.123</td>
<td>1. “kCCR”: 0.07823(-37.16%) 567;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. CV Mean MSE: 0.158, with SE: 0.029</td>
<td>Stop subgroup elimination for “amplifier”. And remove a subgroup from “fluctuation” (according to weights in step A2):</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1. “ENTF”: 0.07703(-4.90%) 828;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Feature number: 1364 → 1931 → 1103</td>
</tr>
</tbody>
</table>
Table C.2: Iterations of our algorithm on DREAM6 data set (Continued: Step A-B). At A5, our algorithm transited to Step B.

<table>
<thead>
<tr>
<th></th>
<th>Best Model</th>
<th>Current Model</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>A4</td>
<td>1. CV Mean $C_p$: 0.735, with SE: 0.054</td>
<td>1. CV Mean $C_p$: 0.743, with SE: 0.051</td>
<td>Remove a subgroup from “resistance”: 1. “ENNGB”: 0.05502(-5.49%) 368; Feature number: 1103 $\rightarrow$ 735</td>
</tr>
<tr>
<td></td>
<td>2. CV Mean $r^2$: 0.345, with SE: 0.123</td>
<td>2. CV Mean $r^2$: 0.359, with SE: 0.106</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. CV Mean MSE: 0.158, with SE: 0.029</td>
<td>3. CV Mean MSE: 0.159, with SE: 0.029</td>
<td></td>
</tr>
</tbody>
</table>

| A5    | 1. CV Mean $C_p$: 0.743, with SE: 0.051 | 1. CV Mean $C_p$: 0.736, with SE: 0.053 | Stop Step A. Go to Step B. Remove these two groups: 1. “fluctuation”: 0.00027(-99.44%) 3; 0.0000915(-73.97%) 3 2. “fluctuation $\times$ resistance”: 0.00036(-99.26%) 3; 0.0001210(-65.55%) 3 Feature number: 735 $\rightarrow$ 729 |
|       | 2. CV Mean $r^2$: 0.359, with SE: 0.106 | 2. CV Mean $r^2$: 0.348, with SE: 0.092 |
|       | 3. CV Mean MSE: 0.159, with SE: 0.029 | 3. CV Mean MSE: 0.168, with SE: 0.030 |

| B1    | 1. CV Mean $C_p$: 0.736, with SE: 0.053 | 1. CV Mean $C_p$: 0.736, with SE: 0.053 | Remove a group: 1. “amplifier $\times$ fluctuation $\times$ resistance”: 0.04900(0.00%) 273; 0.0001795(-49.13%) 273 Feature number: 729 $\rightarrow$ 456 |
|       | 2. CV Mean $r^2$: 0.348, with SE: 0.092 | 2. CV Mean $r^2$: 0.349, with SE: 0.091 |
|       | 3. CV Mean MSE: 0.168, with SE: 0.030 | 3. CV Mean MSE: 0.168, with SE: 0.030 |
Table C.3: Iterations of our algorithm on DREAM6 data set (Continued: Step B).

<table>
<thead>
<tr>
<th></th>
<th>Best Model</th>
<th>Current Model</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>B2</td>
<td>1. CV Mean $C_p$: 0.736, with SE: 0.053</td>
<td>1. CV Mean $C_p$: 0.752, with SE: 0.045</td>
<td>Remove a group:</td>
</tr>
<tr>
<td></td>
<td>2. CV Mean $r^2$: 0.349, with SE: 0.091</td>
<td>2. CV Mean $r^2$: 0.343, with SE: 0.153</td>
<td>1. “amplifier × resistance”: 0.06253(-53.51%) 91; 0.0006871(-37.35%) 91</td>
</tr>
<tr>
<td></td>
<td>3. CV Mean MSE: 0.168, with SE: 0.030</td>
<td>3. CV Mean MSE: 0.143, with SE: 0.024</td>
<td>Feature number: 456 → 365</td>
</tr>
<tr>
<td>B3</td>
<td>1. CV Mean $C_p$: 0.752, with SE: 0.045</td>
<td>1. CV Mean $C_p$: 0.742, with SE: 0.049</td>
<td>Remove a group:</td>
</tr>
<tr>
<td></td>
<td>2. CV Mean $r^2$: 0.343, with SE: 0.153</td>
<td>2. CV Mean $r^2$: 0.338, with SE: 0.147</td>
<td>1. “resistance”: 0.00172(-98.95%) 1; 0.0017157(0.00%) 1</td>
</tr>
<tr>
<td></td>
<td>3. CV Mean MSE: 0.143, with SE: 0.024</td>
<td>3. CV Mean MSE: 0.150, with SE: 0.027</td>
<td>Feature number: 365 → 364</td>
</tr>
<tr>
<td>B4</td>
<td>1. CV Mean $C_p$: 0.742, with SE: 0.049</td>
<td>1. CV Mean $C_p$: 0.736, with SE: 0.050</td>
<td>Remove a group:</td>
</tr>
<tr>
<td></td>
<td>2. CV Mean $r^2$: 0.338, with SE: 0.147</td>
<td>2. CV Mean $r^2$: 0.330, with SE: 0.147</td>
<td>1. “amplifier × fluctuation”: 0.16500(0.00%) 273; 0.0006044(-34.68%) 273</td>
</tr>
<tr>
<td></td>
<td>3. CV Mean MSE: 0.150, with SE: 0.027</td>
<td>3. CV Mean MSE: 0.152, with SE: 0.027</td>
<td>Feature number: 364 → 91</td>
</tr>
<tr>
<td>B5</td>
<td>1. CV Mean $C_p$: 0.736, with SE: 0.050</td>
<td>1. CV Mean $C_p$: 0.712, with SE: 0.063</td>
<td>Stop B. Go to C.</td>
</tr>
<tr>
<td></td>
<td>2. CV Mean $r^2$: 0.330, with SE: 0.147</td>
<td>2. CV Mean $r^2$: 0.293, with SE: 0.109</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. CV Mean MSE: 0.152, with SE: 0.027</td>
<td>3. CV Mean MSE: 0.179, with SE: 0.025</td>
<td></td>
</tr>
</tbody>
</table>
Table C.4: Iterations of our algorithm on DREAM6 data set (Continued: Step C).

<table>
<thead>
<tr>
<th>Lasso</th>
<th>Ridge</th>
<th>Elastic Net</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. CV Mean $C_p$: 0.656, with SE: 0.071</td>
<td>1. CV Mean $C_p$: 0.712, with SE: 0.063</td>
<td>1. CV Mean $C_p$: 0.710, with SE: 0.059</td>
<td>Choose Elastic Net</td>
</tr>
<tr>
<td>2. CV Mean $r^2$: 0.054, with SE: 0.192</td>
<td>2. CV Mean $r^2$: 0.293, with SE: 0.109</td>
<td>2. CV Mean $r^2$: 0.259, with SE: 0.132</td>
<td></td>
</tr>
<tr>
<td>3. CV Mean MSE: 0.209, with SE: 0.023</td>
<td>3. CV Mean MSE: 0.179, with SE: 0.025</td>
<td>3. CV Mean MSE: 0.177, with SE: 0.023</td>
<td></td>
</tr>
</tbody>
</table>
Bibliography


171


175


Biography

Yezhou Huang was born on April 26th, 1986 in Xiangyin County, Hunan Province, China. He earned his Bachelor’s degree in Electrical Engineering and Automation from Xi’an Jiaotong University, China, in July 2007. He earned his Master’s degree in Automation from the Tsinghua University, China, in July 2010. He earned his M.S. degree in Computer Science en route to Ph.D. from Duke University in September 2012.

Publications:

- Huang, Y. and Hartemink, A. “Predicting Transcript Production Rates with Sparse Linear Models”, In Preparation.


- Huang, Y. and Li, S. “Detection of characteristic sub pathway network for angiogenesis based on the comprehensive pathway network”, BMC Bioinformatics (2010), Selected from The Eighth Asia Pacific Bioinformatics Conference.

Awards:

- 2015: Graduate School Summer Research Fellowship
• 2014: Graduate School Summer Research Fellowship

• 2013: IBM Ph.D. Fellowship Nomination

• 2011: DREAM6 Gene Expression Prediction Challenge All Stars (Team Blue1, with Dr. Fantine Mordelet and Daphne Ezer, under the supervision of Prof. Alexander J. Hartemink)