GENOME-WIDE MODELING OF TRANSCRIPTION PREINITIATION COMPLEX DISASSEMBLY MECHANISMS USING CHROMATIN IMMUNOPRECIPITATION DATA

A Dissertation in
Integrative Biosciences
by
Eric Samorodnitsky

© 2011 Eric Samorodnitsky

Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

May 2011
The dissertation of Eric Samorodnitsky was reviewed and approved* by the following:

B. Franklin Pugh  
Willaman Chair in Molecular Biology  
Professor of Biochemistry and Molecular Biology  
Dissertation Advisor  
Chair of Committee

István Albert  
Associate Professor of Biochemistry and Molecular Biology

Claude W. dePamphilis  
Professor of Biology

Stephen W. Schaeffer  
Associate Professor of Biology

Peter J. Hudson  
Director, Integrative Biosciences Graduate Program  
Huck Institute of the Life Sciences

*Signatures are on file in the Graduate School
ABSTRACT

Eukaryotic genes are regulated by hundreds of proteins that assemble into a preinitiation complex (PIC), which functions to initiate transcription. PIC mechanisms of assembly and disassembly in vivo are largely unknown. To address this issue, in Chapter 2, I wrote the computational tool, PathCom (short for PATHway COMpatibility). PathCom takes, as input, an assumed PIC assembly pathway and genome-wide occupancy data. Assuming that occupancy data can be treated as binding duration and explicitly defined assembly/disassembly steps, PathCom outputs plausible PIC disassembly pathways. I exemplify this process by modeling ChIP-chip data of the general transcription factors (TBP, TFIIB, TFIIE, TFIIF, TFIIH, and RNA polymerase II), sequence-specific regulators, and chromatin remodelers of budding yeast Saccharomyces cerevisiae. My modeling found that TBP, sequence-specific regulator, and chromatin remodeler occupancy to be transient compared with the other general transcription factors, given the assumptions inherent in the system. Furthermore, I used PathCom to model the disassembly of GTF’s under heat shock conditions and found TBP occupancy to still be transient under heat shock conditions. PathCom can be used to model any assembly/disassembly process, given all species form a complex together. The reliability of the output modeling of PathCom relies on the accuracy of the assumptions inherent in the modeling and on the quality of the input occupancy data.

To improve the quality of ChIP data, ChIP-chip is being replaced by ChIP-seq. This new technology offers higher resolution and lower background for transcription factor binding site identification. Therefore, in Chapter 3, I wrote a four-step algorithm to infer true transcription factor binding sites from ChIP-seq data. This algorithm
accounts for peak representation on both DNA strands, reproducibility of peaks, and the presence of a motif. This algorithm was written for the genomic toolkit Galaxy. Steps in the algorithm, when possible, were written using preexisting tools on Galaxy. When not possible, certain steps were written in Python. This algorithm is freely available at http://dancluster.g2.bx.psu.edu.
TABLE OF CONTENTS

List of Figures ........................................................................................................ vi
Acknowledgements ............................................................................................... vii
Chapter 1- Introduction ......................................................................................... 1
  1.1 The General Transcription Factors ................................................................. 2
  1.2 Sequence Specific Regulators .......................................................................... 6
  1.3 The Chromatin Immunoprecipitation (ChIP) Assay ......................................... 7
  1.4 Modeling Biological Systems with Chemical Kinetics .................................... 12
  1.5 Computer Programs to Analyze Biochemical Kinetics .................................. 14
  1.6 Different Approaches to Analysis of ChIP-chip and ChIP-seq Data ............... 15
    1.6.1 Hierarchical and K-means Clustering ....................................................... 15
    1.6.2 Principal Component Analysis .................................................................... 17
    1.6.3 Unsupervised and Supervised Learning ..................................................... 19
    1.6.4 Transcription Factor Binding Site Discovery .............................................. 20
  2.1 Summary ........................................................................................................... 23
  2.2 Introduction ....................................................................................................... 24
  2.3 Assumptions ..................................................................................................... 27
  2.4 Results ............................................................................................................... 27
    2.4.1 Genome-wide occupancy modeling of two factors at 25°C ...................... 28
    2.4.2 Development of PathCom to Model Three Factor Occupancy at 25°C .... 33
    2.4.3 Four-, Five-, and Six-Factor PIC Assembly at 25°C ............................... 39
    2.4.4 Modeling Sequence-Specific Regulators and Chromatin Remodelers .... 45
    2.4.5 Modeling PIC Disassembly Under Heat Shock Conditions .................. 50
  2.5 Discussion ......................................................................................................... 50
  2.6 Methods ........................................................................................................... 64
Chapter 3- Incorporation of ChIP-seq Analysis Pipeline to Galaxy ....................... 71
  3.1 Summary ........................................................................................................... 71
  3.2 Introduction ....................................................................................................... 71
  3.3 Results ............................................................................................................... 75
    3.3.1 Initial Analysis ............................................................................................ 75
    3.3.2 The First Criterion: Peak Matching .......................................................... 77
    3.3.3 The Second Criterion: Combining Replicates ........................................... 83
    3.3.4 The Third Criterion: Sequence Analysis ................................................... 86
    3.3.5 The Fourth Criterion: Motif Distance ....................................................... 90
    3.3.6 The Fifth Criterion: Read Count ............................................................... 92
  3.4 Discussion ......................................................................................................... 92
Chapter 4- General Significance and Future Directions ....................................... 96
  4.1 Modeling PIC Disassembly using ChIP-chip data ........................................... 96
  4.2 Incorporation of ChIP-seq Analysis Pipeline to Galaxy ................................... 98
Appendix A- Different Association Orders Affect Plausible Disassembly Pathways .......................................................... 100
Appendix B- TBP Footprints on UCSC Genome Browser .................................... 101
Appendix C- Sequence-Specific Factors and Chromatin Remodelers Analyzed .... 102
Appendix D- Different Dissociation Orders When a Sequence-Specific Factor and General Transcription Factors Are Analyzed .......................................................... 103
Appendix E- The Dissociation Mechanisms With Four General Transcription Factors
Appendix F- Principal Component Analysis of Six General Transcription Factors
Appendix G- Resetting the DNA Concentration to 1 does not Affect Compatibility
Bibliography
List of Figures

Figure 1- General schematic of Chromatin Immunoprecipitation (ChIP) .................. 10
Figure 2- An example of a microarray ................................................................. 17
Figure 3- Principal Component Analysis of Sporulation Genes ......................... 18
Figure 4- Spurious peak calls ............................................................................. 21
Figure 5- Two-Factor Model ................................................................................. 32
Figure 6- Three-factor models ............................................................................. 35
Figure 7- Four factor models ................................................................................. 41
Figure 8- Five factor models ................................................................................ 43
Figure 9- Six factor model using the TBP→TFIIB→pol II→TFIIF→TFIIE→TFIIE association order ................................................................. 44
Figure 10- Six factor model using the TBP→TFIIB→TFIIE→pol II→TFIIF→TFIIE association order ................................................................. 45
Figure 11- Sequence specific factor modeling ..................................................... 48
Figure 12- Mechanism count in sequence specific factor analysis .................... 49
Figure 13- This is the same as Figure 11, except for chromatin remodelers ....... 50
Figure 14- Heat shock analysis for TBP and TFIIB ............................................. 52
Figure 15- Heat shock analysis using TBP, TFIIB, and pol II .............................. 54
Figure 16- Heat shock analysis using the association order TBP → TFIIB → pol II → TFIIF → TFIIE association order ................................................................. 56
Figure 17- Heat shock analysis using the association order TBP → TFIIB → pol II → TFIIF → TFIIE association order ................................................................. 57
Figure 18- Heat shock analysis using six factors .................................................. 59
Figure 19- Shown is an example of a Galaxy workflow ....................................... 75
Figure 20- Transcription Factor-DNA complex following sonication ............. 77
Figure 21- GeneTrack screenshot of ChIP-seq of Reb1 ..................................... 78
Figure 22- Peak Matching Scenario .................................................................... 79
Figure 23- Using Galaxy to make a histogram of the frequency distribution of the shifts of all putative peaks pairs for the human transcription factor CTCF .... 80
Figure 24- Second criterion scenario .................................................................. 84
Figure 25- MEME analysis for the motif of Abf1 .............................................. 87
Figure 26- Motif placement relative a possible binding site .................................. 89
Figure 27- Fourth criterion analysis .................................................................... 91

Appendix A- PathCom analysis for different association orders ...................... 100
Appendix B- TBP footprint on UCSC browser .................................................. 101

Appendix A- PathCom analysis for different association orders ...................... 100
Appendix B- TBP footprint on UCSC browser .................................................. 101
Acknowledgements

I would like to thank the members of the Pugh laboratory who have come and gone and the members of my committee for many helpful discussions. In particular, I would like to my advisor, Dr. Frank Pugh, for helping through some very tough times when the modeling part of my Ph.D. work was being developed.

I would also like to thank the following current members and former members of the Pugh laboratory for performing ChIP-chip on certain factors which I analyzed in this thesis: Bryan Venters, Barbara Andersen, and Andrew Sinnamon. István Albert, Zhenhai Zhang, and Robert S. Harris also deserve a thank-you for programming help. I also need to thank Cizhong Jiang and Shinichiro Wachi for providing some computational code. I need to thank Daniel Blankenberg in Anton Nekrutenko’s group for his collaboration in the third chapter of this thesis. Finally, Bryan Venters deserves another acknowledgement for help and guidance in developing the work in the second and third chapter of this thesis and for reviewing this thesis. Also, my committee members: Drs. B. Franklin Pugh, István Albert, Claude dePamphilis, and Stephen Schaeffer deserve an acknowledgement for discussion of my work overall.

I would also like to thank the members of my family for all the emotional and maturation support given to me over the years since September 14th, 1984. They are a wonderful family and without their help, I would not be where I am today.
Chapter 1 - Introduction

Deoxyribonucleic acid (DNA) controls the metabolism for all cells. DNA in every cell is compacted for storage and regulation. In eukaryotic cells specifically, DNA is compacted by protein octamers called nucleosomes. The subunit of the nucleosome is called the histone. Four different variants of the histone (H2A, H2B, H3, and H4), each present in duplicate, together make up the nucleosome core particle (NCP) [1,2]. Additionally, histone H1 is located in linker regions between nucleosome core particles [3]. Because compacted DNA is generally inaccessible, removing impeding nucleosomes is one obstacle proteins must overcome to express genes.

Gene expression is governed by hundreds of proteins. These proteins can be categorized into four groups: sequence-specific regulators, chromatin remodelers, general transcription factors, and elongation factors [4]. The first group, the sequence-specific regulators bind to specific sequences in the promoters of the gene they regulate. Because they bind only certain sequences, each sequence-specific regulator is found only at a subset of genes in the genome [5]. After sequence-specific factors are recruited, the chromatin remodelers are required to modify nucleosomes and reposition them allowing other proteins to obtain access to the gene promoter [6]. Following chromatin remodelers, general transcription factors (GTF’s) are found within the promoter of every active gene. This group consists of the TATA-binding protein (TBP), TFIIB, TFIIE, TFIIF, TFIIH [7]. Finally, elongation factors, like RNA polymerase, synthesize mRNA.

Together, all the proteins involved in eukaryotic gene regulation form a pre-initiation complex (PIC). The GTF steps for PIC assembly, in vitro, are generally
thought to be TBP $\rightarrow$ TFIIB $\rightarrow$ pol II $\rightarrow$ TFIIF $\rightarrow$ TFIIE $\rightarrow$ TFIIH [7]. However, evidence has been shown that TFIIE can assembly into the PIC before pol II and TFIIH [8] and that TFIIE can influence TBP early in PIC assembly [9]. Despite the existent evidence of the order of PIC assembly, the question remains open how much in vitro defined reactions correspond to in vivo reactions.

1.1 The General Transcription Factors

TBP, despite its name, binds TATA or TATA-less genes and is instrumental for recruiting other GTF’s (for a review of TBP, see [10]). These GTF’s, along with TBP, assemble into the pre-initiation Complex (PIC). Once the PIC is assembled, the DNA unfolds and RNA Polymerase II (pol II) transcribes mRNA by reading the template strand of the gene.

Because of its importance in PIC assembly, TBP is considered one of the most important factors in eukaryotic gene regulation and is subject to regulation at various levels [4]. TBP dimerizes with itself to become inactive to prevent overexpression [11,12]. For further regulation, TBP interacts with other subunits called TBP-associated factors (TAF’s) [13].

The additional subunits TBP interacts with, TAF’s vary between two complexes in which TBP is found. TBP is part of the TFIID and SAGA (stands for Spt-Ada-Gcn5-Acetyltransferase) [14] complex, each of which can initiate its own PIC assembly pathway and has its own unique set of TAF’s. TFIID is used by approximately 90% and SAGA 10% of genes in the Saccharomyces genome [15], and occasionally both complexes assemble at select genes [16].
The TFIID-assembly pathway regulates constitutively expressed genes (housekeeping genes), whereas the SAGA-assembly pathway regulates stress-response genes [15]. Negative Cofactor 2 (NC2) and Modifier of Transcription (Mot1) inhibit the SAGA-pathway, but not the TFIID-pathway [15]. Whether TBP is part of the TFIID or SAGA complex, TBP is instrumental for the recruitment of downstream GTF’s.

The crystal structure of TBP, TFIIA, and TFIIIB on DNA has been determined [17,18,19,20]. TFIIB binds downstream of TBP [20] and opposite of TFIIA [17].

TFIIA (reviewed in [21]) is a heterodimer whose two subunits share no sequence homology, but have similar folds: a β-sheet domain and four helices [17]. TFIIA was initially identified as a general transcription factor [21,22]. However, recent evidence refines this notion. Depletion of TFIIA causes only a moderate decrease in a selection of genes [23]. Furthermore, experimental shutdown of Toa2 (one of TFIIA’s subunits) to 1% of its wild-type levels causes only a 2-3 fold drop in transcription [24]. This evidence suggests TFIIA functions as a co-activator, rather than a general transcription factor. This notion is further enforced by evidence that activators and repressors interact with TFIIA (see Table 1 in [21] and references therein). As mentioned before, TFIIA’s interaction with the GTF TFIIIB has been determined [17,18,19,20].

TFIIB is one of the GTF’s to arrive to the promoter following TBP. TFIIB’s (reviewed in [25]) N-terminal domain is located on the outer surface of the protein, while the C-terminal is protected within the core of the protein [26]. The core domain of TFIIB interacts with TBP, while the outer domain interacts with pol II and TFIIF [25,27,28,29]. TFIIB functions in proper transcription start site selection [30]. TFIIB also functions after PIC assembly. After pol II clears the promoter, pol II would clash with the N-
terminal subunit of TFIIB [31], which would result in pause of transcription, however the N-terminal subunit of TFIIB retracts and allows elongation to continue by RNA polymerase [32].

Three kinds of RNA polymerases exist in eukaryotic cells: I, II, and III. RNA polymerase I synthesizes ribosomal RNA (rRNA), which is a structural unit of the ribosome. RNA polymerase II synthesizes messenger RNA (mRNA), which is translated at ribosomes for protein synthesis. Finally, RNA polymerase III synthesizes transfer RNA (tRNA) [33]. These kinds of RNA’s deliver amino acids to a translating ribosome. Transcription elongation is reviewed by [34]. Because this thesis is about regulation of mRNA-coding genes, only RNA polymerase II (pol II) will be considered henceforth.

Pol II has 12 subunits; each of its subunits is numbered Rpb1 through Rpb12 [35]. The core enzyme’s regions are similar in structure and sequence to RNA polymerase I and III, bacterial RNA polymerases, and the archaeal RNA polymerase [33,36,37,38]. The core enzymes of all three eukaryotic polymerases share many common subunits [33]. Subunits Rpb1, Rpb2, and Rpb6 of pol II bind DNA in the active center, while Rpb1, Rpb5, and Rpb9 bind DNA downstream of the active center [39]. The interaction of pol II and Mediator is important for transcriptional activation and is termed the pol II holoenzyme [40,41,42].

RNA polymerase II, as with any polymerase, synthesizes complimentary RNA by reading the template strand of a gene. However, DNA is wound by nucleosomes, which makes the natural state of DNA inhibitory to movement of polymerase [43]. RNA polymerase II is associated with nucleosome loss both in the promoter region and in the genic coding region (it is still unclear if polymerase causes nucleosome loss or
nucleosome loss allows a polymerase to transcribe [34]), and the nucleosomes are re-deposited after polymerase has synthesized mRNA [44,45,46]. Pol II functions in transcription start site selection aided by TFIIE [47].

TFIIE (reviewed in [48]) consists of four subunits, two TFIIE-α subunits and two TFIIE-β subunits [49]. TFIIE-α stimulates the C-terminal domain (CTD) of TFIIH [50]. TFIIE-β is important for both basal and activated transcription and interacts with single-stranded DNA [51]. TFIIE binds pol II and functions with TFIIF in transcription start site selection [47].

TFIIF is a later general transcription factor to arrive. This factor binds pol II [28]. The higher eukaryotic version of this protein is composed of a 30 kDa and a 70 kDa subunit [52]. However, the yeast version of this protein is composed of three subunits, two of which are related to the more derived protein [53]. TFIIF aids pol II with promoter escape and elongation activity [54]. Following TFIIF is TFIIH as the final GTF in PIC assembly.

TFIIH (reviewed in [55]) is composed of nine different subunits, which are divided into two different sub-complexes: the core complex and the cyclin-activating kinase complex (called CAK) [55]. TFIIF is ring-structured [56,57]. Aside from transcription, TFIIH is also involved in DNA repair and cell cycle control [55]. TFIIH phosphorylates the C-terminal domain of pol II [55], enabling pol II to initiate mRNA production [58], after PIC assembly is complete.

As previously stated, general transcription factors are found at every active gene. These factors are recruited by sequence-specific regulators [59]. Each such regulator, unlike GTF’s, assembles only at a subset of genes in the genome.
1.2 Sequence Specific Regulators

Sequence-specific regulators are the first class of proteins to arrive at active genes. They recruit downstream proteins (chromatin remodelers, GTF’s, etc.) and activate or repress gene expression. The region where sequence-specific regulators bind is called either the Upstream Activator Sequence (UAS) or the Upstream Repressing Sequence (URS), depending on its effect on transcription. Sequence-specific regulators have varying binding motifs (i.e., the DNA sequence the factor binds), number of binding sites, and classes of genes they regulate [4]. Three different example sequence-specific regulators will be used here to exemplify the how the aforementioned characteristics can vary: Reb1, Gal4, and Rap1.

DNA-binding motifs of sequence-specific regulators range from simple to complex. Reb1 binds the simple motif 5’-TTACCG-3’; every base position in the motif is defined by only one base [5,60]. The motifs of other sequence-specific regulators are more complex. Gal4’s binding motif is 5’-CGG(N11)CCG-3’ [5]. This binding motif reflects the dimer crystal structure of Gal4 [61]. Rap1’s binding motif has still higher complexity. Different reports find conflicting results for Rap1’s motif [5,62,63]. These reports all conclude that Rap1’s motif is approximately 10 bp. One example of a reported motif for Rap1 is 5’-CAYCCRTRCA3’ [5]. Binding motifs of the various sequence-specific regulators have varying abundances across the Saccharomyces genome; as a result, sequence specific factors will bind varying quantities of sites and varying classes of genes.
Reb1 controls the expression of ribosomal DNA genes (rDNA) and binds hundreds of sites in the *Saccharomyces* genome [5,60]. Gal4 activates only GAL genes, of which only a few exist; this activation occurs only when galactose is present [64]. Gal4 is predicted to bind ~13 locations in the *Saccharomyces* genome [5]. Rap1 binds to genes with diverse functions (reviewed in [65]). It binds hundreds sites in the genome; it targets promoters, silencers, or even telomeres [65,66]. Rap1 has been found to activate or repress mating-type genes [67].

Thus far, a sub-selection of the kinds of proteins involved in eukaryotic gene regulation has been discussed. To elucidate their genomic locations, the Chromatin Immunoprecipitation Assay (ChIP) is used.

### 1.3 The Chromatin Immunoprecipitation (ChIP) Assay

The most commonly used method for investigating protein-DNA interactions and measuring the occupancy levels of a particular transcription factor bound to DNA is the Chromatin Immunoprecipitation (ChIP) Assay (Figure 1). This technique was initially developed in the 1990’s to map histones and transcription factors and has become an important technique for genomics [68,69,70].

This assay commences with the use of formaldehyde to crosslink a transcription factor (“TF” in Figure 1) to DNA [68]. If the transcription factor of interest interacts indirectly with DNA (i.e. the factor binds another protein that directly interacts with DNA), the protein of interest and its interacting proteins must be crosslinked together as
well [71]. If nucleosome positioning is under investigation, a researcher may start ChIP without crosslinking. This is called native ChIP (or “NChIP”) [70].

Following crosslinking (or NChIP), the cells are lysed and their nuclei are exposed. If studying factor binding, high frequency sound waves are employed to shear the genomic DNA into fragments 200-1000 bp in length. This is called sonication. If studying nucleosome positioning, micrococcal nuclease (MNase) digests all linker DNA between nucleosomes [68].

Subsequent to digestion, the protein of interest is immunoprecipitated with its bound DNA using antibodies. Antibodies specific for the factor may be used, but generating such antibody often delays further experimentation [72]. It is commonly preferred to have a common tag physically attached to all possible factors of interest; hence only one antibody is required for all transcription factors in a set of ChIP experiments. Examples of such tags are the Tandem Affinity Purification tag (TAP-tag) [73] or the Myc-tag [72]. Such tags increase immunoprecipitation efficiency [72]. Following elution, ChIP-chip or ChIP-seq are used to interrogate the eluted DNA for the locations of transcription factor binding sites [4].

ChIP-chip (a type of microarray) was developed in the early 2000’s [74,75]. The immunoprecipitated DNA is labeled with a specific dye-color. In addition, a reference DNA is needed as a “background” dataset and is labeled with a separate color. An example of reference DNA is immunoprecipitation of the transcription factor without its tag [68].

After transcription factor-bound DNA and reference DNA are isolated, in ChIP-chip, both DNA sets are hybridized to oligonucleotide DNA probes. These probes are
synthesized to be specific to certain regions in the genome. Depending on how many probes are synthesized, all or only a portion of the genome may be covered. The resolution of ChIP-chip is limited by the oligonucleotide probe length and genome coverage of the probes. The first generation oligonucleotide probes were large and spanned entire genes [4]. The second generation probes were smaller and were spaced around every 5-bp in yeast and every 40-bp in the larger genomes of fly and human. These second generation probes, cost permitting, span a whole genome with some overlap increasing resolution compared with first generation probes [4,68]. Probes can also be customized against desired regions in the genome [76].
Figure 1- General schematic of Chromatin Immunoprecipitation (ChIP)
Explanation of this figure is in the text. The bottom of the figure shows a screenshot from GeneTrack [77]. Source: [4].
Because each transcription factor-bound DNA and background DNA fragment is incorporated with dye-colors, upon microarray slide exposure to light, the dye color to which each fragment is attached is emitted as shown in Figure 1, using red and green dye. The intensity of colors shows if DNA bound by the transcription factor binds significantly higher than the reference DNA at determined locations in the genome (if either binds at all) [4]. Overall, ChIP-chip is instrumental for understanding protein-DNA interactions. However, ChIP-chip has limitations, which are improved upon in an evolved form of ChIP.

Another method for measuring occupancy is ChIP-sequencing (ChIP-seq). In ChIP-seq, transcription factor-bound DNA is sequenced using Next Generation platforms such as Illumina’s Solexa, Applied Biosystems’ Sequencing by Oligonucleotide Ligation and Detection (SOLiD) sequencers, or the Roche 454 GS20 Sequencer. Once all the DNA fragments are sequenced, they are aligned back to a reference genome (which in the case of this thesis is the yeast genome) [4,68]. Software packages have been designed to align ChIP-seq reads to a genome. Examples of such packages include SHRiMP [78], SOAP [79], and Bowtie [80]. Following alignment, the aligned reads can be visualized on the internet. For this purpose, the UCSC Genome Browser [81] or Ceres [82] are commonly used. GeneTrack is also available to be run on a website or on a local computer [77], a screenshot of which is own at the bottom of Figure 1. This newer version of ChIP is currently becoming the preferred method for genome-wide mapping.

ChIP-seq has certain advantages over ChIP-chip. ChIP-chip is limited by the resolution of the probes used on the array; ChIP-seq reads can theoretically be aligned to any point in the genome. Second generation tiling probes may be used to cover the
genome, but this may prove costly for large genomes [83]. Also, numerous factors affect hybridization between ChIP fragments and DNA probes, and hybridization between imperfectly matched sequences occurs [84]. Furthermore, ChIP-seq shows lower background than ChIP-chip. One report, in particular, showed ChIP-chip and ChIP-seq performed on the same factor. A biologically meaningful peak seen in ChIP-seq was occluded under ChIP-chip [85]. One final advantages of ChIP-seq over ChIP-chip is that ChIP-seq enables alignment to repetitive regions; such regions are harder to map in ChIP-chip [84,86].

Regardless of whether ChIP-chip and ChIP-seq is chosen to map a transcription factor, analysis must be performed to extract meaningful results. These methods of analysis will be discussed later.

1.4 Modeling Biological Systems with Chemical Kinetics

Two methods of modeling biological systems use the principles of chemical kinetics (also called “deterministic”) and stochasticity. Mass action chemical kinetics states the speed of a reaction is directly proportional to the concentrations of all reactants. These principles assume the system is homogenous, diffusive, continuous, and immune to random fluctuations [87]. Cellular processes involve a relatively small number of molecules, which makes the intracellular environment discrete and subject to the probability that of two molecules colliding. For this purpose, stochastic modeling may also be used [88]. When stochastic modeling is employed, reaction speeds have a
probability distribution to account for the likelihood of any two molecules colliding and reacting [89] using the Gillespie algorithm [88,90].

I utilized deterministic chemical kinetics to model PIC assembly in this thesis. Deterministic chemical kinetics is computationally simpler than stochastic calculations [89]. Additionally, the models presented in this thesis are based not just one Saccharomyces cell, but a collection of cells; therefore, an entire group of proteins is being modeled [91].

As stated before, the basic principle of chemical kinetics is that the speed of a reaction is proportional to the concentration of all reactants contributing to the reaction. For example, in the elementary reaction below:

\[ A + B \leftrightarrow C + D \]

In one equation, two reactions are implied: the forward and reverse reaction. The speeds of the reactions at any one point in time are given by the following equations:

\[
\text{forward speed} = \frac{d[C]}{dt} = \frac{d[D]}{dt} = k_f [A][B] - k_r [C][D]
\]

\[
\text{reverse speed} = \frac{d[A]}{dt} = \frac{d[B]}{dt} = -k_f [A][B] + k_r [C][D]
\]

In the equations, \( k_f \) and \( k_r \) are unique rate constants for the forward and reverse reactions, respectively, while \( t \) is time [92].

Almost a century ago, deterministic chemical kinetics was used to model enzymes [93]. Later, more general formulations for enzyme kinetics were derived, formulating the classic Michaelis-Menten equation taught in contemporary biochemistry courses and are used to model cellular kinetics [88,94,95].
1.5 Computer Programs to Analyze Biochemical Kinetics

A variety of kinetic modeling software packages are available in the public domain, only a few of which can be discussed here (a detailed review can be found in [96]). While examining these software packages, the function I required was the Parameter Estimation module. The purpose of the Parameter Estimation function is to, given a set of biochemical kinetic equations and user-inputted parameters (e.g., steady-state concentration of all reacting species), optimize a set of user-chosen parameters (e.g. rate constants for the equations) to satisfy the user-inputted parameters as much as possible. For my purposes, I inputted a set of steady-state concentrations of factors and allowed rate constants to float in PIC assembly/disassembly reactions in Chapter 2 of this thesis. The package that was eventually chosen was COPASI [97].

One chemical kinetics modeling software in the public domain is FluxAnalyzer [98]. The focus of this software is stochastic reactions and to simplify the reaction network as far as possible by looking for redundancies. Cellware [99] performs both chemical kinetics and stochastic simulation, but lacks the Parameter Estimation function. Grid Cellware is an advanced version of Cellware and includes a Parameter Estimation function [100], which is equally compatible for my work in this thesis as COPASI. Dynetica models both deterministic and stochastic chemical kinetics, but does not include a Parameter Estimation module [101].

The software I chose for modeling purposes was COPASI (short for COmplex PAthway SImulator) [97]. As expected from such software, a user inputs chemical equations, rate constants, and initial concentrations, after which a time course and steady-state concentrations are outputted. I chose COPASI, because of the versatility of its
Parameter Estimation function and that I was able to download COPASI and run it locally on my computer.

While the deterministic principles of chemical kinetics were the method of choice in Chapter 2, other methods exist for gleaning information from large-scale datasets. Resolving patterns in massive datasets is the general theory of bioinformatics.

1.6 Different Approaches to Analysis of ChIP-chip and ChIP-seq Data

Because ChIP-chip and ChIP-seq produce large datasets, a challenge is to infer novel concepts from the datasets. The inference of such concepts can be categorized in two classes: the bottom-up and the top-down approach. The bottom-up approach begins with an equation to describe a biochemical process and asks whether the data are consistent with the models. The second approach, top-down, does exactly the reverse: it begins with a large dataset and distills models based on the dataset [102]. This section will briefly review the different methods developed for analyzing large datasets.

1.6.1 Hierarchical and K-means Clustering

For ChIP-chip and expression microarray data, a commonly used top-down approach for seeking similarities in occupancy and expression patterns is clustering. The two main algorithms for clustering are hierarchical and K-means clustering.

Hierarchical clustering seeks to arrange genes into a hierarchy such that similarly behaving genes are placed next to each other [103]. The method starts with single-gene clusters, successively grouping the closest clusters until all gene clusters are connected.
A dendogram is generated to indicate statistical similarity between genes [105]. To calculate similarity between genes, “distance” is computed. Users are given options of how to calculate distance. They include single linkage (where clusters are separated based on nearest genes), complete linkage (where clusters are separated based on farthest genes), average linkage (where clusters are separated based on average distance between all pairs of genes), and centroid (where clusters are separated based on centroid distances) [105,106]. While hierarchical clustering generates a continuum of genes clusters, K-means clustering produces explicit groups of genes.

K-means seeks to divide genes into explicitly-defined groups based on similarity without forming any hierarchical relationship [103,104]. A user-defined “K” number of clusters is chosen and K-centroids are randomly generated. Each gene is putatively assigned the closest centroid. These centroids are next recalculated to reflect the putative gene groupings. The process of reassigning genes starts again and centroids are recalculated until genes no longer change clusters. A drawback of K-means clustering is different starting centroids may generate different clusters, thus it is important to execute the algorithm multiple times in order to determine the reproducibility of any round of clustering [104,107].

Clustering is commonly performed using the software packages Cluster (for actual clustering) and TreeView (for display) [108]. Use of these software packages employing both hierarchical and K-means clustering is exemplified in Figure 2. Each column represents one (out of 48) mutation. These mutations are hierarchically clustered by how similarly gene expression is perturbed. Each row represents a different gene; these genes are K-means clustered into 10 different groups.
Figure 2 - An example of a microarray
Each column shows a different mutation. The mutations are hierarchically clustered by how gene expression is affected. Each row represents one gene. The genes are K-means clustered into 10 unique groups. Source: [16]

1.6.2 Principal Component Analysis

Clustering is only one method for analyzing large scale ChIP-chip data. Such data can further be analyzed using Principal Component Analysis (PCA). In this mathematical procedure, given a microarray dataset of genes × arrays, the dataset is reduced to n-principal components, where “n” represents the number of arrays (columns) in the dataset [109]. The first principal component, of n-principal components, is a
vector that attempts to capture as much variation as possible of the arrays; the second principal component captures as much of the remaining variation as possible, etc. The last principal component finds the least variation of the dataset. These n-principal components, in the n-dimensional space, are linearly independent and can be used to remap each gene [110]. To elaborate PCA, an example will be shown.

The expression patterns of 477 sporulation genes in *Saccharomyces cerevisiae* were found to have seven different temporal patterns [111]. **Figure 3** shows the expression of these genes mapped by two and three principal components [112].

![Figure 3](image)

**Figure 3- Principal Component Analysis of Sporulation Genes**
A, Sporulation genes mapped by the first two principal components. B, Sporulation genes mapped by the first three principal components. Source: [112]

**In Figure 3A**, some overlap between different clusters of genes is shown. Gene expression in 3A is mapped by the first two principal components, which capture 85.9% of the variation. Going further to map gene expression by the first three principal components in **Figure 3B**, 93.2% of the variation is captured and the genes are more clearly separated [112]. Further principal components will separate clusters more clearly.
PCA analysis and clustering are forms of unsupervised learning. Supervised and unsupervised learning are important categories of interpretation of mass datasets.

### 1.6.3 Unsupervised and Supervised Learning

In supervised learning, a set of elements (e.g., genes) must be classified into groups based on a set of rules. This set of rules, however, is not always known when a mass dataset is produced. Therefore, any supervised technique must first deduce the rules of classification the dataset [113]. An example of supervised learning was predicting functional RNA’s in bacteria and archaea [114]. Using neural networks and supports vector machines (see [113]), the authors used the patterns of functional RNA’s in the *E. coli* genome as a training set and applied this training set to identify functional RNA’s in other bacteria and archaea. Supervised approaches have been utilized for other purposes, such as gene prediction [115] and distinguishing coding from non-coding regions of human DNA [116], seeking explicitly defined groups or patterns. Explicit information on classification or patterns are not always readily available and supervised learning assumes the information gleaned from the training dataset is representative of the experimental dataset [117].

When training datasets are not available to yield information on existing patterns, unsupervised learning must be used. This method of learning finds existing patterns, lacking explicit information on pre-defined groups [117]. Unsupervised learning has been used to find genomic signatures among microbes living in acidic environments [118]. In addition, gene identification has been performed disregarding assumed models
of gene behavior [119]. Overall, unsupervised learning is used to generate hypotheses, but supervised integration is used to test these hypotheses using external datasets [120].

1.6.4 Transcription Factor Binding Site Discovery

Unsupervised and supervised learning have important applications in the discovery of protein-DNA interactions, such as nucleosome positioning [121] and transcription factor binding site mapping. The discovery of transcription factor binding sites is the focus of Chapter 3, hence this section reviews current methods for their discovery. Earlier reports using ChIP-seq to identify genomic locations of transcription factors operated on read density solely (i.e., that read density at binding site must be above a defined threshold) [122,123], however read density, by itself, has shortcomings for true binding site discovery [124].

Read density does not account for how reads may have mapped. The possibility exists that reads may map heavily to one strand and lightly to the other at a particular genomic location; this likely indicates a spurious peak. This weakness is addressed by ensuring that any bona fide peak call is represented on both the W- and C-strands [125]. Some, but not all, ChIP-seq analysis packages require such representation (e.g.,[124], see Table 1 in [125]). Even requiring that a peak must be represented on both strands, spurious peaks may still occur if all reads map to one or a few chromosomal coordinates on each strand (Figure 4). Therefore, it is important that any true transcription factor binding site is represented on both strands and the peak’s reads are not concentrated strictly at one or a few coordinates [125].
Two possible scenarios representing false peaks are shown. W-strand reads (or peaks) and C-strand reads (or peaks) are colored red and blue, respectively. On the left, false W- and C-peaks are rejected for lack of a partner peak on the opposite strand. On the right, false peaks are rejected, because reads mapped to only one coordinate (or a few coordinates). Source: [125].

Given that peak calls are represented on both strands and their reads are distributed over a narrow interval, it is important that peaks calls be reproducible between replicates. Previous reports managed replicates differently. One report consolidated ChIP-seq reads if peaks from replicates were reproducible [126]. Another report consolidated transcription factor peaks from two replicates, if peaks from separate replicates overlapped by at least one base-pair [127]. Others consolidated peaks calls from the same library, if they were relatively close [128].

Consolidating replicate peak calls and ensuring peak representation on both DNA strands are each pre-processing steps for transcription factor binding site identification (i.e., these steps are performed before binding sites are declared). After binding sites are declared, a recurring downstream step is motif identification [129]. Discovery of motifs is performed by various software packages, most commonly MEME [130,131]. Motif identification has been proposed to instead be a further pre-processing step to filter out spurious peaks, because transcription factors recognize specific motifs [132]. However,
it has been found previously using ChIP-chip data that the human factor E2F may bind sites lacking a recognizable motif [133,134,135].

The following two chapters of this thesis have separate goals. The goal of Chapter 2 is the development of a top-down approach to biological modeling to infer plausible disassembly mechanisms of transcription machinery from ChIP-chip data. In Chapter 3, the top-down approach is utilized again to distill transcription factor binding sites from ChIP-seq data.
Chapter 2-Genome-wide Modeling of Transcription Preinitiation Complex Disassembly

Mechanisms Using ChIP-chip Data


2.1 Summary

Apparent occupancy levels of proteins bound to DNA in vivo can now be routinely measured on a genomic scale. A challenge in relating these occupancy levels to assembly mechanisms that are defined with biochemically isolated components lies in the veracity of assumptions made regarding the in vivo system. Assumptions regarding behavior of molecules in vivo can neither be proven true nor false, and thus is necessarily subjective. Nevertheless, within those confines connecting in vivo protein-DNA interaction observations with defined biochemical mechanisms is an important step towards fully defining and understanding assembly/disassembly mechanisms in vivo. To this end, we have developed a computational program PathCom that models in vivo protein-DNA occupancy data as biochemical mechanisms under the assumption that occupancy levels can be related to binding duration and explicitly defined assembly/disassembly reactions. We exemplify the process with the assembly of the general transcription factors (TBP, TFIIB, TFIIE, TFIIF, TFIIH, and RNA polymerase II) at the genes of the budding yeast Saccharomyces. Within the assumption inherent in the system our modeling suggests that TBP occupancy at promoters is rather transient compared to other general factors, despite the importance of TBP in nucleating assembly of the preinitiation complex. PathCom is suitable for modeling any
assembly/disassembly pathway, given that all the proteins (or species) come together to form a complex.

### 2.2 Introduction

Eukaryotic genes are thought to be regulated by hundreds of proteins that assemble into pre-initiation complexes (PIC’s) at promoters using an ordered pathway. One aspect of the PIC assembly pathway involves the recruitment of the general transcription factors (GTF’s), such as TBP and TFIIB, by sequence-specific activators. TBP and TFIIB then contribute to the recruitment of RNA polymerase II (pol II) and other GTF’s, which eventually start transcription. A fundamental question concerning our understanding of gene regulation is the extent to which each assembly and disassembly step is distinct at every gene in a genome. Is the traditional biochemical view that TBP “‘locks in’” or commits to a promoter, and in a recurring manner nucleates PIC formation valid in vivo? And is the PIC disassembly process in vivo, simply the reverse of the assembly process? Parts of the assembly/disassembly pathway have been rigorously defined in vitro with a few purified proteins and DNA, and this has provided us with our current parsimonious view of PIC regulation [7,136,137,138]. In no case have assembly or disassembly reactions been reconstituted in a way that fully recapitulates the physiological setting (presence of sequence-specific regulators, coactivators, specifically positioned nucleosomes, chromatin regulators, GTFs, etc) at every gene, and so these questions remain open, in regards to the extent to which in vitro defined reactions mimic the in vivo events occurring throughout a genome. The answer to
this question is not readily addressed in vivo, since reactions are not as definable nor as quantifiable as in vitro biochemical reactions with purified components. Nonetheless, assays do exist for measuring relative levels of protein-DNA complex formation in vivo, and so mechanistic inferences will be sought.

The goal here is to evaluate in vivo occupancy data in light of biochemical mechanisms that are intended to reflect the corresponding in vivo reaction. The extent of biological insight is predicated on rather subjective assessments of the assumptions associated with interpretation of in vivo data. Within the context of declared constraints and assumptions, we propose a means to model in vivo protein-DNA occupancy data, so as to better integrate and conceptualize massive genomic datasets. This study is focused on the means of such modeling and the assumptions inherent in the data, using specific examples of PIC assembly.

Currently, perhaps the most widely used assay to measure the occupancy of proteins at genes in vivo is the chromatin immunoprecipitation assay (ChIP). In ChIP, proteins are crosslinked to DNA, the protein is then purified, and the bound DNA identified either through directed PCR or through genome-wide detection platforms (ChIP-chip and ChIP-seq). In this way, for example, the relative occupancy level of TBP, TFIIB, pol II, and many other proteins at every promoter in the genome in a population of cells can be assayed. Recent studies using differential ChIP and photobleaching experiments have provided compelling evidence for a dynamic state of PIC components in living cells [139,140,141]. Therefore, it is now within a conceptual framework to expect factors like RNA polymerase II, TBP, and other GTFs to undergo multiple assembly and disassembly cycles at promoters for each productive transcription
event, rather than the traditional simple view that GTF’s remain locked in place during multiple transcription cycles.

The existence and origins of distinct occupancy levels of PIC components on genes has not been systematically explored, and thus is the impetus for conducting the modeling studies described here. Differential occupancy patterns for the GTFs have been described [8], and may be caused by gene-specific regulators that influence the recruitment or retention of specific general transcription factors (among other proteins), and thus assembly/disassembly mechanisms might differ from gene to gene (or sets of genes). Here, we attempt to utilize ChIP-chip binding information gleaned at every promoter in the yeast genome to either plausibly infer or exclude PIC assembly/disassembly mechanisms. The major limitation in any such approach is that the number of permutations of possible assembly/disassembly mechanisms exceeds the amount of data available to constrain such mechanisms. In other words, occupancy data, alone, is insufficient to uniquely specify an ordered PIC assembly and disassembly pathway. Imposition of additional constraints (or assumptions), such as predefining either the assembly (or disassembly) pathway, might however eliminate certain dissociation (or association) mechanisms as incompatible with the data, and thus serves the purpose of plausibly excluding mechanisms rather than uniquely identifying a mechanism.

Here, we develop a ChIP modeling program, termed PathCom, in the context of a fixed PIC assembly pathway to infer allowable dissociation mechanisms. We validate the simulation using an existing chemical kinetics simulator COPASI [97]. Within the declared constraints, we discern the compatibility of different PIC disassembly
mechanisms at nearly every transcriptionally-active gene in the yeast genome with existing ChIP-chip occupancy data.

2.3 Assumptions

Because biological systems are generally very complicated, we had to make some assumptions in order to simplify the modeling process.

- PIC assembly can be modeled using the principles of deterministic chemical kinetics.
- ChIP-chip data can easily translate into simple chemical kinetic steps.
- When converting the ChIP-chip signal to percent binding (see Methods), probes with the lowest signal were assumed to have 0% binding, while the probes with the highest signal were assumed to have 100% binding. We assumed also that all probes in between scaled linearly with occupancy signal.
- All factors were assumed to assemble and disassemble as separate entities, while in fact, some factors can come on/off linked together.
- We assumed a constant assembly pathway across the whole genome and the assembly pathway is based on *in vitro* data.

2.4 Results
2.4.1 Genome-wide occupancy modeling of two factors at 25°C

The overall goal here is to inter-relate ChIP in vivo occupancy data with biochemical assembly/disassembly mechanisms, in a way that attempts to support or dispute such mechanisms. Such inter-relationships can be complex when one considers that hundreds of proteins are involved in transcriptional regulation. Therefore, we start by modeling only two factors (the GTF’s TBP and TFIIB), and increase complexity by adding more GTFs one at a time up to six factors. While we focus on PIC assembly/disassembly mechanisms on a genomic scale, any number of factors and combination of assembly/disassembly steps in gene regulation may be considered, given that all proteins (or species) come together to form a complex.

TBP (T) binds to DNA (D) to form a protein-DNA (TD) complex, and in the presence of TFIIB (B) form a TDB ternary complex (Figure 5A) [17,18,19]. In the presence of sufficient levels of these proteins, their DNA occupancy level will vary from 0% to 100% as dictated by the context of each promoter. In principle, there are two pathways by which TBP and TFIIB assemble step-wise onto DNA (Figure 5B) [20]: A) TBP binds to DNA, then TFIIB binds; or B) TFIIB binds DNA first, then TBP. Their reversal constitutes two pathways for dissociation.

The constant availability of energy to drive directional processes allows the pairing of any association and dissociation mechanism. Consequently, there are four paths by which an in vivo occupancy level is achieved for a two-component reaction. The availability of only two experimental constraints (TBP and TFIIB occupancy levels on DNA) is insufficient to specify the predominant association and dissociation pathways. In the absence of a necessary additional experimental constraint, we created a hypothetical
constraint for the purposes of modeling, in which we eliminated all but one association pathway. That allowed us to evaluate the two possible dissociation pathways. The reciprocal modeling could also be done, by eliminating all but one dissociation mechanism. Since the purpose of this study is to demonstrate how the modeling works and to discuss its assumptions, caveats, and utility, we illustrate the process using a single association pathway that has good experimental support and model all possible dissociation pathways.

Biochemical [7] and crystallographic [20] evidence shows that TBP binds DNA first, followed by TFIIB, which makes cooperative contacts with both TBP and the DNA (Figure 5A). On this basis, we fixed assembly pathway “A” (Figure 5B), which sufficiently constrains the system so that measured TBP and TFIIB occupancy levels can distinguish between the two dissociation pathways, “1” and “2”. In this context, dissociation pathway “1” allows either TBP or TFIIB occupancy to be greater than the other, but pathway 2 is only plausible if TBP occupancy is greater.

Using published genomic datasets of TBP and TFIIB occupancy [76], we modeled four groups of genes, each having either a high (H) or low (L) experimentally measured level of TBP and TFIIB (Figure 5C). These occupancy levels were reproducible and verified by a second data source (Affymetrix high density tiling arrays) also present in the previous study (not shown) [76]. We chose four subdivisions so as to separately consider different types of occupancy patterns. In principle, each gene could be treated independently. However, grouping of similarly behaving genes had the advantage of creating more robust occupancy values that are based upon hundreds of measurements, rather than just one. Aggregating the data dampened the variability
caused by gene-specific differences in crosslinking efficiency and detection. It also
served to identify predominant occupancy patterns that might reveal underlying themes in
gene regulation. One limitation of such grouping is that it assumes a single underlying
mechanism exists for an individual gene and for an entire group of genes, which may be
unlikely in detail but reasonable for purposes of demonstration.

To compare occupancy levels between proteins, it was necessary to place them on
the same scale. We achieved this by scaling ChIP occupancy values (fold over
background) for each factor from 0% to 100%. Our rationale, assumptions, and method
for doing this are described in the Methods section.

**Figure 5D** shows a cluster-plot of the genes with their TBP and TFIIIB percent
occupancies. Since the “(L, L)” group (**Figure 5D**) had low levels of both factors, TBP
and TFIIIB did not substantially occupy these genes. Consequently, modeling would not
be informative for this group, and thus was not examined further. In addition, the “(H,
L)” group comprised, 1% of all genes, and so it too was not examined further. For the
remaining two groups, TFIIIB occupancy was greater than TBP occupancy. When
assembly pathway A was fixed, in which TFIIIB assembles last, then the observed higher
level of TFIIIB occupancy over TBP can only be accommodated by a situation where
TFIIIB dissociates last. Thus, for both groups ((L, H) and (H, H)), the data reject
dissociation pathway 2 (TFIIIB dissociates first) and accept pathway 1. These outcomes
are illustrated in **Figure 5D**, by the black (incompatible) and green (compatible) squares.
Note that when the alternative assembly pathway B is fixed, both dissociation pathways
were compatible. This simple case illustrated how different starting assumptions
(assembly pathway A vs B) resulted in a different set of compatibility outcomes.
From this analysis, several insights were obtained: 1) Some occupancy levels simply do not distinguish among mechanisms. 2) In contrast to the simplified in vitro derived biochemical mechanism, TFIIB might remain at most promoters after TBP has dissociated (although TFIIB may nevertheless be dynamic). How TFIIB does so is a matter of speculation that the data do not address.

Based upon known TBP/TFIIB/DNA biochemical interactions, the notion that TFIIB might dissociate after TBP would seem untenable. However, the additional complexity that exists in vivo might accommodate such a mechanism if other proteins not explicitly defined in this model retained TFIIB at the promoter, after TBP had dissociated. TFIIB engages pol II at promoters via specific interactions [31,33,142]. Pol II tightly associates with DNA in an “open” promoter complex [143,144], and tends to accumulate at the 5’ ends of genes [76,145,146,147]. If an active mechanism removes TBP, such as through the well-established ATPdependent mechanism of Mot1 [148], then TFIIB might remain on promoter DNA via pol II and in the absence of TBP.
Figure 5 - Two-Factor Model

A. Crystal structure of a TBP-DNA [17,19] and TBP-DNA-TFIIB complex [20]. B. Shown are the two possible association and dissociation mechanisms of TBP (T), TFIIB (B), and DNA (D). C. Shown is a cluster plot using TreeView [108]. Every row is a single gene, with black spots indicating lower occupancy (0%) and red indicating higher occupancy (100%). The cluster plot is split into the four different groups. D. Each row is one of the four possible clusters. The blue bars show the number of genes in each cluster. The pie charts show the median percent occupancy of each factor for each cluster (out of 100%). The red bars show the median transcription frequency (mRNA/hr) [149]. For compatibility testing, both association pathways (A and B) were tried and paired with each dissociation pathway (1 and 2) to see if each association/dissociation pair were compatible with the median occupancies of each cluster. Black rectangles indicate incompatibility and green rectangles indicate compatibility, which are all results the program PathCom. The row for (L, L) is shaded out because of low signal and the row for (H, L) is shaded out because of low membership. Source: [150]
2.4.2 Development of PathCom to Model Three Factor Occupancy at 25°C

Towards our goal of modeling the assemblage of many proteins, we next consider a three-factor assemblage. The interaction of TFIIB with pol II (P) is structurally and biochemically well-defined [20,31]. As in the two-step modeling, based upon biochemical precedent, we constrain the system to the following assembly pathway: TBP → TFIIB → pol II (Figure 6A, black arrows). Since there are three-factors, there are six-possible dissociation pathways. Modeling three factors through six mechanisms for eight groups of genes became conceptually challenging to work through in the intuitive manner described for two factors. However, we determined that the plausibility of any mechanism could be evaluated by two basic rules:

**Rule 1: Does the mechanism make it unconditional that one protein’s occupancy level must be greater than another’s?** For example, in the two-factor mechanism, if TFIIB enters last and leaves first (Figure 6B, left path), then such a mechanism requires that TFIIB occupancy be less than TBP occupancy. On the other hand, if TFIIB leaves last (Figure 6B, right path), then such a mechanism allows both TBP and TFIIB to occupy the DNA independent of each other. Therefore, this mechanism will accommodate any occupancy levels observed for these proteins.

**Rule 2: Does the occupancy of one protein, other than the first or last to assemble, have an occupancy level greater than the summed occupancy of any previously-associating protein and subsequently-associating protein?** If so, does the mechanism
give the possibility that the protein’s occupancy is greater than the combined
occupancies of these two other proteins? This rule is applicable towards modeling of
more than two-factors. When this condition is met, then the proteins at some point must
occupy the DNA without the other two proteins, and thus must be the last of three to
dissociate (but not necessarily the last to dissociate overall if the mechanism has more
than three proteins. When iterated over all factors in a mechanism this rule determines
the allowable orders of dissociation. For example, consider a fixed assembly order with
TBP first, then TFIIB, then pol II (Figure 6C): If TFIIB occupancy is greater than the
sum of TBP and pol II occupancy, then only those dissociation mechanisms that have
TFIIB dissociate last are compatible. If this condition is not true, then any dissociation
mechanism can be accommodated by this rule, including the ones having TFIIB
dissociate last (but some might be disallowed in the context of rule 1).
Figure 6- Three-factor models

A. At the center is the full TBP-TFIIB-pol II-DNA complex. Each gray arrow represents the dissociation of a particular factor. T is TBP, B is TFIIB, P is pol II, and D is DNA. Choosing a particular path from the center to a vertex, where D is unbound, one can see the six different pathways. The black arrows show the assumed association pathway. B, The first rule of compatibility is pictured. Because of the assumed TBP \(\rightarrow\) TFIIB pathway, the pathway on the left requires TBP occupancy to be greater than TFIIB, while the pathway on the left has no such restraint (other than TBP and TFIIB occupancies are greater than 0%). C, The second rule of compatibility is illustrated. Only two of the dissociation pathways are shown. Given the assumed assembly pathway, TBP \(\rightarrow\) TFIIB \(\rightarrow\) pol II, if TFIIB occupancy is greater than or equal to the combined occupancy of TBP and pol II, only the two dissociation pathways shown will work. Note, however, that
while these pathways permit this condition, they don’t require it. D, Using PathCom to do compatibility testing for the eight possible clusters. The blue bars show the number of genes in each cluster and the pie charts indicate the median percent occupancies. The green and black rectangles indicate compatibility and incompatibility, respectively, for each cluster with each dissociation pathway (number above the rectangles correspond to the dissociation pathway in part A of this figure). Note we tried excluding the tRNA genes for the normalization, because tRNA binds TBP very highly, but not the other GTF’s. E, The red bars on the left show the median transcription frequencies for each cluster. To the right is a compatibility chart (similar to PathCom) using COPASI [97]. The numbers on the chart are log_{10} of the E-values. The histogram below that shows the log_{10} E-values plotted. One can see the distribution is bimodal: the group of log_{10} E-values on the left represents compatible log_{10} E-values, while the group on the right represents incompatible ones. Source: [150]

These two rules, together, determine which dissociation mechanisms will be compatible with the data, given an assumed assembly pathway. Note that depending on the actual percent occupancies, these rules will have varying effectiveness in narrowing down the dissociation mechanisms. If the rank order of the observed occupancy is the same as the order of association, then any dissociation mechanism will work.

We transformed these queries into a program termed PathCom (short for Pathway Compatibility), which was used to generate the compatibility chart in Figure 6D (green = compatible, black = incompatible). This software is available as supplementary material in the original publication with a version for Windows users and a version for Mac users. Using the rationale from the two-step model, we generated eight groups of genes corresponding to either high or low occupancy of each of the factors (Figure 6D).

We sought to validate the approach taken by PathCom, to ensure that it reflected enzymological concepts for which this modeling attempts to emulate. Our validation employed COPASI, a freely program that simulates biochemical kinetics [97]. Reaction mechanisms and concentrations (the latter equivalent to occupancy levels described here)
represent their input parameters. For each mechanism and each group of genes, COPASI iteratively “guesses and checks” in an attempt to find a set of rate constants that delivers the observed occupancy levels for TBP, TFIIB, and pol II. It then reports a goodness-of-fit by measuring the square difference between the observed and the optimized occupancies, reporting this as an E-value (see Methods).

To maximize the parameter search space and avoid local minima, COPASI imposes some randomness in moving through the decision making process. Since the system is under-constrained and randomness is involved, each repeated modeling run converges on a different solution for each mechanism (ie., many different combinations of rate constant values can produce the observed occupancy levels, if a solution can be found). The values of the underlying rate constants generated by the Parameter Estimator in COPASI are not meaningful; rather the resulting E-value provides a quantitative measure of the suitability of a mechanism to fit the data. Re-running COPASI on the same dataset returns essentially the same E-value (not shown). Thus COPASI provides a robust means of evaluating alternative mechanisms and validating PathCom.

**Figure 6D** shows the compatibility findings of all eight possible clusters using three factors against the six possible dissociation mechanisms using PathCom. **Figure 6E** shows the corresponding log_{10} E-value assessments using COPASI. In all cases, the COPASI-reported E-values matched the Boolean decisions made by PathCom (compare **Figure 6D** and **6E**). log_{10} E-values generated by COPASI were bimodal (**Figure 6E**, bottom bar graph), providing a demarcation between compatible and incompatible outcomes. Thus, the simplified Boolean process associated with PathCom was validated by a kinetic mechanism simulator (COPASI).
Importantly, the analysis indicates that given a fixed association mechanism, there are limited number of dissociation mechanisms (green squares in Figure 6D) that can account for the observed occupancy data. Fixing different association pathways generates different compatibility patterns (Appendix A). In Figure 6D, clusters of genes that had very few members (e.g. (H, L, L) and (H, L, H)) or had low occupancy of all tested factors (e.g. (L, L, L)) may not be particularly robust, and thus less reliably interpreted. For the remaining clusters, one to two mechanisms were found to be compatible. A common theme was that TBP dissociated first, then pol II, and then TFIIB, which was consistent with the conclusions drawn from the two-factor assembly analysis described above.

In principle, dissociation of pol II may proceed via removal into the bulk nucleoplasm and/or translocation down the DNA upon transcription, where ChIP occupancy would not be detected by microarray probes at the 59 ends of genes. Consistent with the latter possibility, high transcription frequencies are observed at the (H, H, L) set, which has high TBP and TFIIB occupancy but relatively low occupancy of pol II (Figure 6C). These genes are also enriched with pol II in the body of the gene (not shown).

The suggestion that TFIIB dissociates after both TBP and pol II dissociation is consistent with some reports in the literature [24], and suggests that perhaps other factors retain TFIIB at promoters in the absence of TBP and pol II. TFIIB and TFIIF are known to interact with each other [25], and potentially with activators [24,26,27,28].

We further examined the plausibility that TBP might not be fully bound at ‘‘high’’ occupancy promoters by looking at experimentally determined ‘‘digital
footprints” of TBP bound at those promoters having the highest TBP occupancy (Appendix B) [151]. Indeed, in all cases, no TBP footprint was detected over the TATA box, which is consistent with the notion that TBP does not fully occupy even its most highly occupied sites.

Groups of genes that had very few members (e.g., (H, L, L) and (H, L, H)), or had very low occupancy of all tested factors (e.g. (L, L, L)) are expected to have higher variation, and thus less reliably interpreted. Therefore, these groups were not examined further. For the remaining groups, one to two mechanisms were found to be compatible. A common theme was that TBP dissociated first, then pol II, and then TFIIB, which was consistent with the conclusions drawn from the two-factor assembly analysis described above.

2.4.3 Four-, Five-, and Six-Factor PIC Assembly at 25°C

As more factors were added to the modeling, and genes grouped according to low or high occupancy levels of each protein, the number of possible groups grew exponentially ($2^n$, where $n$ is the number of modeled proteins). Consequently, membership in each group diminished, some to negligible levels. Those with negligible membership did not represent predominant patterns and may have arisen by chance as a consequence of noisy occupancy levels. Therefore, we combined groups of genes that lacked a viable membership level (see Methods for membership criteria).

Using the in vitro model for PIC assembly, we next added TFIIH (H) to the mechanism: TBP → TFIIB → pol II → TFIIH. This mechanism is applicable even if pol II and TFIIH were entering together. As shown in Figure 7A, the groups with the highest
membership of genes included those with low TBP occupancy levels, and either low or high levels of the other GTFs (indicated by asterisks for gene groups that had at least two high occupancy GTFs). A group having high levels of all GTFs predominated among those groups having high TBP occupancy, denoted (H, H, H, H). In the context of the modeled assembly pathway, these results suggest that TBP is removed from most measured genes before the other GTFs, except in cases where PIC assembly is maximal. The latter could be interpreted to reflect continuous reloading of TBP, which has recently been shown to be fairly dynamic [140,141]. Our modeling studies with PathCom suggest that the most plausible mechanisms for gene groups with abundant membership and at least two high abundance GTFs include early TBP dissociation (Figure 7B). However for one abundant gene set (L, H, L, H), the data are also compatible with an early dissociation of pol II followed by TBP (or simultaneous with it) (Figure 7B, dissociation mechanisms 13 and 14).

In the four-factor mechanism, groups having a relatively large gene membership typically were limited to being compatible with only one or two of the 24 theoretically possible dissociation mechanisms (Figure 7A, compatibility chart). Thus, the modeling of more factors increased the number of potential mechanisms in a factorial relationship (n!) with the number (n) of proteins being modeled. However, the number of plausible mechanisms remained largely fixed at one to two, with a few exceptions.
Figure 7- Four factor models
A. Results of using PathCom on the four-factor model. B. The predominant dissociation pathways found for the clusters with asterisks next to them in part A of this figure. The black arrows show the association of the PIC and the gray arrows shows the various dissociation pathways. Source: [150]

We next added TFIIF (F) (Figure 8) and TFIIE (E) (Figures 9 and 10). While evidence suggests that TFIIF fits into the following fixed assembly pathway (including simultaneous recruitment with pol II) [137]: TBP → TFIIB → pol II → TFIIF → TFIIH [7,137] the literature reports seeming conflicting evidence for TFIIE entry [7,8,9], and thus we chose to pursue to two alternative assembly mechanisms: TBP → TFIIB → pol II
TFIIF $\rightarrow$ TFIIE $\rightarrow$ TFIIH (Figure 9) and one where TFIIE enters prior to pol II (Figure 10). We focused on the few clusters that had the most members and had multiple factors with high occupancy (indicated by asterisks). These included clusters with 687, 580, and 252 members (Figs. 8, 9, and 10). The membership for these particular clusters remained unchanged as more factors were included in the modeling because they failed to generate new gene groups that had sufficient membership to avoid consolidation. Thus, the occupancy data and the associated mechanisms displayed robust consistency as multiple GTF’s were added on, which is consistent with them working together in a PIC.

The occupancy levels in the five-factor modeling were compatible with mechanisms that had TBP and pol II dissociate early and TFIIB and TFIIF dissociating late (Figure 8B). Interestingly, groups with few genes tended to have a larger number of compatible mechanisms (more green boxes in Figure 8A). While the significance of this is unclear, it might reflect a cellular design that avoids ambiguity in the PIC disassembly pathway. That is multiple, alternative dissociation pathways may be problematic to control.
Figure 8- Five factor models
See Figure 6 for a description. The transcription frequency bars are omitted for clarity purposes. If one wishes to see the transcription frequency, see the original source of this figure. Source: [150]

In modeling six factors (Figures 9 and 10), the predominant compatible disassembly pathways for the two alternative assembly pathways retained the dissociation of TBP and pol II as early steps in all mechanisms. Whether we define TFIIE assembly as early (upper panel) or late (lower panel), the occupancy data supported the following two
predominant dissociation mechanisms: \(P \rightarrow T \rightarrow H \rightarrow B \rightarrow (E,F)\) and \(T \rightarrow P \rightarrow (E,F,H) \rightarrow B\), although when \(E\) associated early, the following pathway was also acceptable: \(T \rightarrow P \rightarrow (F,H) \rightarrow (E,B)\). Spot checks of our results using COPASI confirmed our findings (not shown).

**Figure 9** - Six factor model using the TBP \(\rightarrow\) TFIIB \(\rightarrow\) pol II \(\rightarrow\) TFIIF \(\rightarrow\) TFIIE \(\rightarrow\) TFIIH association order

Source: [150]
2.4.4 Modeling Sequence-Specific Regulators and Chromatin Remodelers

Having modeled PIC assembly and disassembly of GTF’s using PathCom, we were interested in applying PathCom to model the complexes recruiting the GTF’s: sequence-specific regulators and chromatin remodelers. Modeling of these factors, aside from reporting the disassembly mechanisms, allows us to understand if sequence-specific
regulators and chromatin remodelers, in general, leave the promoter early or late during gene expression (as compared with the GTF’s). These factors have shown lower fold-enrichment relative to the background signal than have GTF’s at active genes [8]. In this section, we analyzed previously published [76] and unpublished ChIP-chip data for sequence-specific regulators and chromatin remodelers using PathCom.

The ChIP-chip slides used by our laboratory utilize three probes for each mRNA-coding gene: a UAS probe, a TSS probe, and a genic probe at calculated distances from each gene’s ORF start (see Methods section of this thesis and the Methods section of [76]). When PathCom was applied to GTF’s, occupancy signals at TSS probes were analyzed. However, sequence-specific regulators and chromatin remodelers bind to the UAS, rather than near the TSS; however, we analyzed each sequence-specific regulator’s (henceforth referred to as “SSR”) and chromatin remodeler’s (henceforth referred to as “CR”) signals at both the UAS and TSS probe.

With each sequence-specific regulators or chromatin remodeler, we included TBP and pol II into its model. Additionally, to keep calculations simple, we took the median percent occupancies of TFIIB, TFIIE, TFIIF, and TFIIH for each gene (this group of factors will be referred to as “G” henceforth), because their disassembly pathway was not easily elucidated from the previous section. Because it is established [4] that GTF’s are recruited by sequence-specific regulators and chromatin remodelers, we assumed the following PIC assembly pathway: (SSR or CR) $\rightarrow$ TBP $\rightarrow$ pol II $\rightarrow$ G. Fifteen SSF’s and five CR’s were analyzed at both the TSS and UAS probes (GTF’s at the TSS probe in both scenarios), attempting to learn when an SSR or CR dissociates from the PIC, as
compared with all the GTF’s. See Appendix C for a list of SSR’s and CR’s analyzed in this thesis.

The modeling results of all 20 factors at both probes cannot be shown, given space limitations. Instead, we show the modeling results of only one factor, which is a representative of all the other factors. As done when analyzing GTF’s, all genes were classified into the H-L categories depending on whether a gene had greater than or less than 10% binding for each factor. However, to increase the likelihood each SSR or CR was present at the promoter of analyzed genes, we looked at genes whose SSR or CR binding was classified as “H.” Additionally, for simplicity, consolidation was not performed in any modeling of SSR’s or CR’s.

The shown example is the analysis of the SSR Ash1 at the UAS probe (note that an analysis for Ash1 at the TSS probe was also done). GTF signals were strictly at the TSS probe. The assumed order of association order was Ash1 → TBP → pol II → G. The results of the analysis are shown in Figure 11 for only the genes whose Ash1 percent-binding was greater than or equal to 10%.
Figure 11- Sequence specific factor modeling

Shown is the dissociation modeling analysis of Ash1 at the UAS (the GTF’s signals are at the TSS) using PathCom. The predominant dissociation mechanisms are highlighted at the top of the figure. The disassembly sequences corresponding to each mechanism number are shown in Appendix D.

Not all clusters in Figure 11 were reliably interpreted, for similar reasons as in the previous section. Some clusters had few members; also, some clusters had too few of their GTF’s above 10% binding to have reasonable signal to make a judgment call for the proper dissociations. Therefore, we decided to only interpret only clusters that had at least two of factors TBP, pol II, and G classified as “H” and the cluster had to have at least ten genes.

Given our method for deciding clusters to interpret, and the general occupancy patterns of genes, only three clusters from the analysis of all SSR’s or CR’s (UAS or TSS) were interpreted: cluster (H, L, H, H), cluster (H, H, L, H), and cluster (H, H, H, H).
These were the only clusters that yielded useful information about PIC disassembly accounting for an SSR or CR.

What is interesting about these clusters is that six disassembly mechanisms (out of 24 possible) dominated: mechanisms 1, 3, 7, 9, 13, and 15 (see Appendix D for which mechanism numbers correspond to which dissociation patterns). All of these mechanisms have in common is that they all have the G-proteins (i.e., the median of TFIIB, TFIIF, TFIIE, and TFIIH) dissociate last, while TBP, pol II, and the SSR or CR dissociated earlier.

The patterns of G-proteins dissociating last was prevalent among all SSR and CR models. For each SSR or CR analyzed, the frequency that each dissociation sequence appeared was tallied in the three mentioned clusters above and plotted in Figure 12 for SSR’s and Figure 13 for CR’s. Mechanisms 1, 3, 7, 9, 13, and 15 were predominant both when the SSR or CR was analyzed at the UAS or the TSS.

![Figure 12- Mechanism count in sequence specific factor analysis](image)
PathCom analysis was done for each sequence-specific regulator and the GTF’s in a four-step mechanism. Each mechanism number corresponds to a different dissociation order. This graph shows the number of times each dissociation order appeared from all the factors. This analysis was done separately for the SSR at the TSS and the UAS.

![Mechanism Count](image)

**Figure 13** - This is the same as Figure 11, except for chromatin remodelers.

As previously mentioned, all the predominant mechanisms had in common TFIIB, TFIIF, TFIIE, and TFIIH dissociating last. We can conclude from this that sequence-specific regulators and chromatin remodelers dissociate from the PIC early, around the same time that TBP and pol II do, and before TFIIB, TFIIF, TFIIE, and TFIIH.

PathCom analysis, thus far, has focused on the assembly and disassembly of transcription machinery under unstressed condition. However, it is known that cellular machinery is reprogrammed when the cell is undergoing stress [8,152,153].

### 2.4.5 Modeling PIC Disassembly Under Heat Shock Conditions
As a further addition to our modeling, we wanted to see if genes change PIC disassembly pathways under heat shock conditions. In our laboratory, heat shock exposure is performed in the following procedure: yeast strains are grown at 25°C and exposed to heat shock conditions, 37°C, for 15 minutes. After heat shock (and non-heat shock control), the experimentalist performs ChIP-chip for the factor of interest [8,153]. We can, then, compare the occupancy levels of a particular factor between non-heat shock and heat shock to determine changes in factor binding.

The analysis in this section is similar to previous analyses at 25°C. The primary distinction lies in the Boolean classification of genes. If a gene had <10% binding for a factor at both 25°C and 37°C, the factor at that gene was given the label “LL.” If the factor had <10% binding at 25°C, but ≥10% binding at 37°C, the factor at that gene was given “LH.” If binding was ≥10% at 25°C, but <10% at 37°C, the factor was given “HL.” Finally, if the factor had ≥10% binding at both 25°C and 37°C, it was given “HH.” As in previous analyses, all genes with identical factor labels were grouped together and the median percent binding of each cluster was calculated. Unlike the previous analyses where we merely sought the prevalent dissociation pathways given a single association pathway, in this analysis, we sought to understand if clusters changed dissociation pathways under stress conditions. This analysis also tells us if new dissociation pathways arise under stress conditions that were normally not found at 25°C, given differences in cell behavior under heat shock from 25°C.

Unlike the 25°C analysis, each factor can be classified four possible ways (LL, LH, HL, HH). Because four classification possibilities exist, with each factor added to the model, the number of theoretically possible clusters equals $4^n$ (rather than $2^n$ in the
original analysis), where “n” is the number of factors. For two factors, 16 clusters are possible; for six factors, 4096 clusters are possible. Therefore, to keep the number of analyzed clusters low, two requirements were developed in order for a cluster to be analyzed. The first requirement was the cluster must have at least 50 genes. Additionally, at least one factor had to be classified as “LH” or “HL,” because we wanted to make sure at least one factor was changing occupancy classifications.

Figure 14 shows the heat shock analysis when applied to the two-factor model: TBP and TFIIB using the assumed constant association pathway TBP $\rightarrow$ TFIIB.

![Figure 14- Heat shock analysis for TBP and TFIIB.](image-url)
On the left is shown the cluster label. The figure follows formats of previous figures. The occupancies for each cluster at each temperature are shown in the pie charts. As before, Mech 1 is the dissociation pathway TBP $\rightarrow$ TFIIB and Mech 2 is TFIIB $\rightarrow$ TBP. For a pictorial representation of the associations/dissociations, see Figure 5.

Figure 14 shows that only the dissociation pathway TBP $\rightarrow$ TFIIB appears as compatible with all the occupancy data, both at 25°C and 37°C. This is consistent with the results of the two-factor analysis only at 25°C that TBP $\rightarrow$ TFIIB is predominant dissociation pathway. As mentioned before, TFIIB might remain at the promoter by its interactions with pol II [31,33,142].

Like the 25°C analysis, we next added to pol II to our analysis using the constant association pathway TBP $\rightarrow$ TFIIB $\rightarrow$ pol II. The results of the analysis are shown in Figure 15.
Cluster/Temp

<table>
<thead>
<tr>
<th>Temperature</th>
<th>TBP</th>
<th>TFIIB</th>
<th>pol II</th>
</tr>
</thead>
<tbody>
<tr>
<td>25°C</td>
<td>(LL, LL, LH)</td>
<td>(LL, LL, LH)</td>
<td></td>
</tr>
<tr>
<td>37°C</td>
<td>(LL, LL, LH)</td>
<td>(LL, LL, LH)</td>
<td></td>
</tr>
<tr>
<td>(LL, LL, HL)</td>
<td>(LL, LL, HL)</td>
<td>(LL, LL, HL)</td>
<td></td>
</tr>
<tr>
<td>(LL, LL, HH)</td>
<td>(LL, LL, HH)</td>
<td>(LL, LL, HH)</td>
<td></td>
</tr>
<tr>
<td>(LL, HL, LL)</td>
<td>(LL, HL, LL)</td>
<td>(LL, HL, LL)</td>
<td></td>
</tr>
<tr>
<td>(LL, HL, HL)</td>
<td>(LL, HL, HL)</td>
<td>(LL, HL, HL)</td>
<td></td>
</tr>
<tr>
<td>(LL, HL, HH)</td>
<td>(LL, HL, HH)</td>
<td>(LL, HL, HH)</td>
<td></td>
</tr>
<tr>
<td>(LL, HH, LL)</td>
<td>(LL, HH, LL)</td>
<td>(LL, HH, LL)</td>
<td></td>
</tr>
<tr>
<td>(LL, HH, HL)</td>
<td>(LL, HH, HL)</td>
<td>(LL, HH, HL)</td>
<td></td>
</tr>
<tr>
<td>(LL, HH, HH)</td>
<td>(LL, HH, HH)</td>
<td>(LL, HH, HH)</td>
<td></td>
</tr>
<tr>
<td>(HL, HH, LL)</td>
<td>(HL, HH, LL)</td>
<td>(HL, HH, LL)</td>
<td></td>
</tr>
<tr>
<td>(HL, HH, HL)</td>
<td>(HL, HH, HL)</td>
<td>(HL, HH, HL)</td>
<td></td>
</tr>
<tr>
<td>(HL, HH, HH)</td>
<td>(HL, HH, HH)</td>
<td>(HL, HH, HH)</td>
<td></td>
</tr>
</tbody>
</table>

Figure 15- Heat shock analysis using TBP, TFIIB, and pol II.
See previous figures for caption. The assumed association pathway was TBP $\rightarrow$ TFIIB $\rightarrow$ pol II. See Figure 6A to see which mechanism numbers correspond to which dissociation pathways.

Except for one case, every dissociation mechanism is Figure 15 was either T$\rightarrow$B$\rightarrow$P, T $\rightarrow$P$\rightarrow$B, or P$\rightarrow$T$\rightarrow$B. This concurs to some extent with the corresponding analysis using only 25°C data: mechanisms 2 and 5, corresponding to the dissociation pathways T$\rightarrow$P $\rightarrow$B and P$\rightarrow$T$\rightarrow$B respectively, were the most prevalent. This still suggests that TFIIB can remain at the promoter after the dissociation of TBP and pol II. Besides the
congruence of the dissociation mechanisms with previous analysis, it is important to note that Figures 14 and 15 were organized for a specific purpose: to show if clusters are changing dissociation mechanisms between room temperature and heat shock conditions. Figure 15 (and 14) show that clusters, in general, are not changing their dissociation mechanism(s). TFIIB might remain in the promoter, in the absence of TBP and pol II, by its interactions with activators and downstream factors [30,32,34,43,44].

As before, we next added TFIIH for our four-factor analysis (Figure 16, the list of dissociation mechanisms is shown in Appendix E). The four-factor heat shock analysis confirms the analysis at 25°C. Mechanisms 3, 4, 13, and 14 were prevalent; mechanisms 3 was $T \rightarrow P \rightarrow B \rightarrow H$, 4 was $T \rightarrow P \rightarrow H \rightarrow B$, 13 was $P \rightarrow T \rightarrow B \rightarrow H$, and 14 was $P \rightarrow T \rightarrow H \rightarrow B$. Each mechanism had TBP and pol II being first or second to dissociate and TFIIB and TFIIH being third or fourth to dissociate. This is in agreement with earlier analyses that TBP and pol II leave the promoter early, while TFIIB and TFIIH leave the promoter late. Importantly, we saw that most clusters were compatible with only one or two dissociation mechanisms.

To find out whether clusters were changing dissociation mechanisms under heat shock, we decided to analyze in Figure 16 only cluster (LL, HH, LH, HH), cluster (LL, HH, LH, HH), cluster (LL, HH, HL, HH), cluster (LH, HH, LL, HH), cluster (LH, HH, LH, HH), and cluster (LH, HH, HH, HH), because these clusters under both temperatures had at least half of their factors with over 10% binding and had at least 50 genes.
Of these clusters, three of them changed mechanism compatibility, while two others did not. These clusters were sometimes compatible with mechanism 3 (T→P→B→H) at
25°C and 37°C, but always compatible with mechanism 4 (T→P→H→B) at both temperatures. Overall, we saw some change in the dissociation mechanisms under heat shock, but the change in dissociation mechanisms did not reveal any mechanisms that did not appear in the PathCom analysis done only at 25°C.

Like the 25°C models, we next added TFIIF to our model, using the association pathway TBP → TFIIB → pol II → TFIIF → TFIIH (Figure 17).

Cluster/Temp

<table>
<thead>
<tr>
<th>Cluster/Temp</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>(LL, LL, LL, LL, LH)</td>
<td>25°C</td>
</tr>
<tr>
<td>(LL, LL, LL, LL, LH)</td>
<td>37°C</td>
</tr>
<tr>
<td>(LL, LL, LL, LL, HL)</td>
<td>25°C</td>
</tr>
<tr>
<td>(LL, LL, LL, LL, HL)</td>
<td>37°C</td>
</tr>
<tr>
<td>(LL, LL, LL, LH, LL)</td>
<td>25°C</td>
</tr>
<tr>
<td>(LL, LL, LL, LH, LL)</td>
<td>37°C</td>
</tr>
<tr>
<td>(LL, LL, HL, LL, LL)</td>
<td>25°C</td>
</tr>
<tr>
<td>(LL, LL, HL, LL, LL)</td>
<td>37°C</td>
</tr>
<tr>
<td>(LL, HL, LL, HL, HL)</td>
<td>25°C</td>
</tr>
<tr>
<td>(LL, HL, LL, HL, HL)</td>
<td>37°C</td>
</tr>
<tr>
<td>(LL, HL, HL, HL, HL)</td>
<td>25°C</td>
</tr>
<tr>
<td>(LL, HL, HL, HL, HL)</td>
<td>37°C</td>
</tr>
<tr>
<td>(LH, HH, LL, HH, HH)</td>
<td>25°C</td>
</tr>
<tr>
<td>(LH, HH, LL, HH, HH)</td>
<td>37°C</td>
</tr>
<tr>
<td>(LH, HH, LH, HH, HH)</td>
<td>25°C</td>
</tr>
<tr>
<td>(LH, HH, LH, HH, HH)</td>
<td>37°C</td>
</tr>
<tr>
<td>(LH, HH, HH, HH, HH)</td>
<td>25°C</td>
</tr>
<tr>
<td>(LH, HH, HH, HH, HH)</td>
<td>37°C</td>
</tr>
</tbody>
</table>

Figure 17- Heat shock analysis using the association order TBP → TFIIB → pol II → TFIIF → TFIIH. See Figure 14 for a description.
Observing the clusters in Figure 17, we decided clusters (LH, HH, LL, HH, HH), (LH, HH, LH, HH, HH), and (LH, HH, HH, HH, HH) were most robust for interpretation, because, as before, more than half of their factors had at least 10% binding under both 25°C and 37°C. The dissociation mechanisms under both conditions for all three clusters were of the form (TBP, pol II) \( \rightarrow \) (TFIIB, TFIIF, TFIIH). This is somewhat in agreement with the corresponding analysis done for the 25°C previously. However, in the 25°C models, the most prevalent mechanisms mostly saw TBP dissociate first and pol II second, while the dissociation order of TFIIB, TFIIF, and TFIIH was less discerned. In the 37°C analysis, however, pol II dissociating first was more prevalent than in the corresponding 25°C analysis.

Lastly, we added TFIIE to our heat shock model. Also, as done earlier, we tried two different association pathways (Figure 18, see corresponding analysis earlier for 25°C).
Figure 18- Heat shock analysis using six factors. The top panel uses the association order TBP → TFIIB → pol II → TFIIF → TFIIE → TFIIH. The bottom panel uses the association order with TFIIE between TFIIB and pol II.

Observing the clusters in Figure 18, we deemed clusters (LH, HH, LL, HH, HH, HH), (LH, HH, LH, HH, HH, HH), and (LH, HH, HH, HH, HH, HH) (they are the same label in both panels of Figure 18; the bottom three clusters in both panels) to be most reliable to interpretation, because at least half of their factors had over 10% binding under both conditions. As before, regardless of which association order is chosen, we still found that
TBP and pol II dissociate first in either order, followed by the other GTF’s whose dissociation order is less discerned. We also saw that these clusters do not have congruent compatibilities of dissociation mechanisms between 25°C and 37°C; therefore clusters are changing dissociation mechanisms between 25°C and 37°C.

**2.5 Discussion**

Genome-wide occupancy data for the many hundreds of proteins involved in gene regulation is now accumulating. One major challenge has been to inter-relate such occupancy data and conceptualize it in light of models about how these proteins function together. Such models, as in the case of the assembly of the transcription machinery at promoters, are derived from biochemical experiments conducted on isolated components of the transcription machinery. The extent to which inferred biochemical mechanisms reflect in vivo processes is not known. We are not aware of any means of modeling genome-wide occupancy data to determine whether it is compatible with biochemical mechanisms. To this end, we developed the software tool PathCom. PathCom is generic in that it will determine whether any number of user-defined mechanisms is compatible with measured occupancy data of any number of relevant proteins. We applied PathCom to transcription complex assembly/disassembly, which has been extensively defined biochemically and for which genome-wide ChIP-chip occupancy data is available for. Biological insight gleaned from the modeling is subject to the veracity of the assumptions regarding what in vivo ChIP occupancy data actually measures, and the quality of the data being modeled.
Eukaryotic protein coding genes utilize a common set of general transcription factors to assemble RNA polymerase II at promoters. A long-standing question that biochemistry has attempted to explain is the order of assembly of the transcription machinery and what happens to individual components during multiple transcription cycles. As far as the general transcription machinery is concerned, in vitro ordered assembly starts with TBP followed by TFIIB, then pol II and TFIIF, and then TFIIE and TFIIH [7,137]. In vivo ChIP occupancy data alone cannot discern whether such an assembly pathway is correct at any or all genes, and thus is a premise of the modeling example employed here. In the context of such a fixed assembly pathway, we explored different occupancy patterns of the general transcription machinery observed across the yeast genome, and interpret such occupancy patterns to potentially reflect alternative dissociation mechanisms. Should alternative association mechanisms be considered, then alternative dissociation mechanisms are likely.

In regards to the genome-wide distribution of the GTF’s, for the analyses at 25°C, we did not see a random partitioning of genes into high vs low occupancy states for each factor. Principal component analysis (PCA) indicates the presence of a single major component (not shown), and several minor ones. This would be consistent with the strong tendency of the GTF’s to work together. What is interesting about the PCA is that TFIIB, pol II, TFIIF, and TFIIH were the main drivers in the first principal component, despite pol II having relatively low occupancy at the promoter region. TBP contributed the least to the principal components (Appendix F). In addition, we determined whether genes with <10% occupancy or ≥10% occupancy had a tendency toward having TATA versus TATA-less promoters, using data from [154]. We found that approximately 20% of genes
with $<10\%$ or $\geq 10\%$ occupancy levels were TATA-containing genes. Therefore, we did not see a clear bias toward TATA or TATA-less for “L”-occupancy or “H”-occupancy genes. Also we took the very highest TBP binding genes (at least 50% binding) and they also had 20% TATA-box genes. It does not seem likely that factor percent binding shows any association with the percent of genes that have TATA-boxes or sequence-effects in general.

When clustering all GTF’s and pol II, three high occupancy states stood out as having a large membership. These included genes with high levels of 1) all GTF’s, 2) all GTF’s except TBP, and 3) all GTF’s except TBP and pol II. The group having high levels of all GTF’s was by far the most highly transcribed, which is not surprising. This group included the ribosomal protein genes. However, for the major groups, low levels of TBP were more closely linked to low levels of transcription than the occupancy level of any of the other factors including pol II. This confirms on a genomic scale the earlier notion established on a few genes that TBP recruitment or retention is rate-limiting in transcription [155]. However, since pol II and the other GTF’s are commonly found at high levels at many promoters even when TBP levels are low, it also seems likely that steps after TBP recruitment will be rate-limiting at certain genes. Otherwise, a rapid initiation and elongation phase would be expected to result in low pol II occupancy at all promoters.

While the number of dissociation mechanisms scale factorially ($n!$) with the number ($n$) of proteins involved, we did not see an equal distribution of genes into each type of mechanism, and we did not see a corresponding increase in the number of compatible dissociation mechanisms. Instead, the number of compatible mechanisms
remained rather fixed at one to two, for a given association mechanism. The general pattern observed for most genes, was that if TBP, TFIIB, pol II, and the other GTFs assembled in the listed order, then the dissociation order was generally: TBP, then pol II, then the other GTFs, with the latter being less resolved.

Having observed the dissociation order of the GTF’s, we next wanted to see when sequence-specific factors and chromatin remodelers disassemble from the PIC relative to the GTF’s. We found that sequence-specific regulators and the chromatin remodelers dissociated around the same time as TBP and pol II and before the other GTF’s. We observed these patterns regardless of whether we used the signal from TSS probe or UAS probe for the sequence-specific regulator or chromatin remodeler. This is consistent with published results, where the occupancy levels of most sequence-specific regulators and chromatin remodelers were lower than those of TFIIB, TFIIE, TFIIF, and TFIIH at active (rather than quiescent) genes than relative to non-promoter regions [8]. These results could mean that sequence-specific regulators and chromatin remodelers leave the promoter once the GTF’s are recruited or that these complexes are as transient in promoter binding as TBP and pol II.

After demonstration of PathCom’s modeling at 25°C for GTF’s, sequence-specific regulators, and chromatin remodelers, we wanted to see if a stress response, like heat shock, might change PIC disassembly pathways from 25°C and reveal PIC disassembly pathways that did not manifest previously at 25°C. As before, the most robust clusters were compatible with no more than a handful of mechanisms. Like in the 25°C analysis, we found that TBP and pol II still left the PIC early, while TFIIB, TFIIF, TFIIE, and TFIIH left the PIC late. However, despite these similarities, we found
clusters generally had different PIC disassembly pathways at 25°C and 37°C for the higher order models. This analysis may provide evidence that, biochemically, no dissociation mechanisms, other than the ones found at in the 25°C analysis, manifested. This analysis may further indicate that genes up- or downregulated by heat shock change their PIC dissociation pathway to adjust to the elevated temperature.

2.6 Methods

Background Normalization

Low-density tiling microarray occupancy data for the following factors at 25°C were obtained from ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) using the accession numbers E-MEXP-1676 and E-MEXP-1677: TBP, TFIIB (Sua7 subunit), pol II (Rpo21 subunit), TFIIF (Tfg1 subunit), TFIIE (Tfa1), TFIIH (Ssl1), Cin5, Ifh1, Rap1, Reb1, Rfx1, Swi4, Xbp1, Yap6, INO80 (Ino80 subunit), RSC (Rsc9 subunit), SWI-SNF (Swi3 subunit), which was published in [76]. All raw factor occupancy data at 37°C and any 25°C data not previously mentioned was obtained from our lab’s databases. The following proteins were analyzed whose ChIP-chip data was not published previously: Ash1, Bas1, Cha4, Rph1, Sfp1, Zap1, and Zms1. High-density tiling microarray was obtained from supplementary material in [76]. The 25°C YPD media occupancy data for the TSS probe (in the low-density data) was used for modeling, representing occupancy values near the TSS for 5743 genes. The probe for the TSS was designed for regions between 30 bp and 90 bp upstream from the start of the actual ORF. However, all raw data (from 20,000 probes) was processed as follows: First, a background dataset was
calculated. Each BY4741 background dataset was normalized to the median value of the entire dataset. Then replicates were combined by computing the median value for each probe. Second, each factor ChIP dataset was divided by the background BY4741 dataset, on a probe-by-probe basis, then divided by the median value for all probe signals located in T-T regions (‘‘tail-to-tail’’ intergenic regions between convergently transcribed genes, which are expected to be devoid of bound factors). The resulting occupancy levels represent fold over background, centered so that the ratio in nonpromoter regions equals 1. For further information on the experimental design, see [76].

**Scaling Datasets from 0% to 100% Binding**

This scaling was necessary to compare occupancy levels across different factor datasets. In principle such scaling eliminates differences in crosslinking efficiencies and ChIP yields between factors. Fold-over-background values equal to or less than 1 represent background and thus were re-coded as 0% occupancy. Several limitations of the ChIP assay precluded accurate assessment of 100% occupancy. First, ChIP hybridization signals generally correlated with actual occupancy levels, but were not tightly linked (see below), and so the maximum detected fold enrichment over background could not simply be set to 100%, inasmuch as the variance might be quite substantial. Second, ChIP assays do not measure absolute binding, and so even if the variance were eliminated, we could not be certain that the maximum detected level of binding represented 100% occupancy. Nonetheless, if all factors are held to the same standard, and data from groups of similarly behaving genes are aggregated, then approximations can be made. Therefore, we coded any value above the 99th percentile
rank (top 200 probes) as 100% (setting the 100% mark to the upper 98th percentile gave essentially the same results). All remaining data were scaled between 0 and 100% occupancy by subtracting background (1.0) from all data, and dividing through by the value at the 99th percentile rank.

**Assumption of Linearity of Occupancy Levels**

It is generally assumed that ChIP signals scale linearly with actual occupancy level. However, it is possible that a factor bound to one type of DNA sequence may crosslink more readily than when bound to a different sequence. To test the effect of underlying DNA sequence on crosslinking efficiency, we examined the distribution of TBP occupancy levels at each of the eight TATA box subtypes [154], TATA(A/T)A(A/T)(A/G). A chi-square test demonstrated that TBP occupancy levels were independent of DNA sequence (p-value = 0.48) (not shown). Next we tested TFIIB, which binds both TBP and DNA, and also found it to be independent of the sequence of the TATA box (p-value = 0.76) (not shown). Nevertheless, to minimize the influence of crosslinking efficiency on measured occupancy levels, similarly behaving genes were grouped, and their median occupancy level was used in the modeling. In addition, we focused on those groups having high gene membership, which should further alleviate fluctuations associated with individual genes.

**Grouping of Genes into Low and High Occupancy Levels**

To increase the robustness of the occupancy values, as well as focus the modeling on predominant patterns, we grouped genes in accordance with their occupancy level for
each factor. For only the 25°C analysis, genes (probes) having a GTF occupancy below 10% were parsed into low (L) occupancy groups. All others were parsed into high (H) occupancy groups, resulting in $2^n$ theoretically possible groups, where “n” is the number of GTF’s being modeled. Parsing the data at a 15% cutoff, or into three groups (low, medium, high using the 10% and 20% for the low-medium and medium-high cutoffs, respectively) did not substantially alter the outcomes, and its main conclusions. When sequence-specific regulators and chromatin remodelers were being modeled, the median of the percent binding of TFIIB, TFIIF, TFIIE, and TFIIH was calculated for each gene; the four factors were modeled as one group for simplicity.

Unlike modeling at 25°C, when modeling PIC disassembly under heat shock, if a factor at a gene had <10% occupancy at both 25°C and 37°C, that factor at that gene was given “LL”. If the occupancy was <10% at 25°C, but ≥10% at 37°C, “LH” was given. If the occupancy was ≥10% at 25°C, but <10% at 37°C, “HL” was given. Lastly, if a factor at a gene at ≥10% occupancy under both temperatures, “HH” was given.

Groups having low membership do not represent predominant patterns and so were consolidated as follows (only done for GTF analysis at 25°C): Groups having ≥100 genes were exempt from consolidation because they have substantial membership, and groups having <10 genes were required to be consolidated for lack of viable membership. Otherwise, if the membership of an existing group was split by more than a 4:1 ratio when an additional factor was added to the model (e.g. from 2-factor models to 3-factor models), then the two resulting clusters were consolidated (i.e., not split; note that the label of the consolidated clusters was assigned the label of the larger cluster). Because of consolidation, the number of actual clusters is less than $2^n$. Note that consolidation was
not performed when we were analyzing the two- and three-factor models in order to make the modeling explanations more clear. Also consolidation was not performed when sequence-specific regulators and chromatin remodelers were analyzed and for heat shock analysis.

**PathCom**

PathCom requires the user to enter occupancies of proteins in a tab-delimited text file followed by the name of the cluster line by line. In a header, before the occupancies are entered, users enter one-letter codes to denote protein identities (of the user’s choice) followed by a number to indicate the order in which the proteins assemble. Below each protein in the header, the user enters the percent occupancies calculated along with the name of each cluster (or gene). After execution, the program then reads each cluster’s occupancies on each line. Given the fixed order of association of proteins specified by the user in the header, the program generates all possible dissociation sequences. Note that if the user changes the association order, the pool of dissociation reactions will remain the same, but the numbering of each dissociation reaction will be different, because PathCom uses the specific association to generate the dissociation sequences. The program processes each dissociation sequence, pairing it with the fixed association sequence, and given the rules of compatibility (discussed in the paper), computes whether the input protein occupancies are compatible with the mechanism (association + dissociation) it is testing. PathCom processes all possible dissociation sequences for all groups entered. PathCom writes the results to a tab-delimited text file. In this file, the
horizontal axis is labeled with every mechanism identification number and the vertical axis is labeled with every cluster name. Also, PathCom writes a file that matches each dissociation sequence with its dissociation sequence identification number. Every time a set of occupancies and a mechanism are compatible, the program reports ‘‘-1’’, and when they are not, the program reports ‘‘0.’’ Results can be clustered through Cluster then visualized graphically in Treeview [108].

COPASI

COPASI conducts chemical kinetic and stochastic simulations [97], and is freely available for download at www.copasi.org. Reactions were set to be irreversible for simplicity. Initial input protein and DNA concentrations were set to be equal, having an arbitrary value of 10 (setting the DNA concentration to 1 gave the exact same results in terms of compatibility, not shown). Since the observed occupancy levels for a factor represent the sum of all intermediate species having that factor, it was necessary to employ the Parameter Estimation function to optimize this sum, using the free protein concentration equal to \((1 – \text{Occ}/100)\times10\), where ‘‘Occ’’ is the measured percent occupancy level, and had a practical lower limit of 0.1% (this formula is only valid when all species concentrations were set to 10). The Parameter Estimator may converge on a local minimum, which may not represent the optimal solution. Running the estimator multiple times alleviated the local minimum, since it employs a random search component. COPASI creates an objective value \((E)\) used to measure goodness of fit between simulated and measured values:
\[ E = \sum_i w_i (x_i - y_i)^2 \]

where ‘‘i’’ represents each of the protein factors involved in the modeling, ‘‘w’’ is the weight that is given to a particular protein in the optimization procedure, which is calculated automatically by COPASI, ‘‘x’’ is the measured occupancy, and ‘‘y’’ is the simulated occupancy. Since COPASI aims to minimize this sum of squares, lower E values (more negative log\(_{10}\) E) reflect better congruence between modeled and measured data. Since each modeling run has a manual component and becomes computationally draining with a large number of factors, it became impractical to run COPASI to fully generate the compatibility charts for four or more factors. Nonetheless, we employed COPASI to spot check these charts (not shown), and found 100% agreement with PathCom.
Chapter 3- Incorporation of a new ChIP-seq Analysis Pipeline to Galaxy

3.1 Summary

Chromatin Immunoprecipitation (ChIP) followed by massive parallel sequencing (ChIP-seq) is becoming the generally preferred way to measure occupancy of a transcription factor on DNA. After ChIP-seq reads are mapped to the genome, an analysis must be performed to call the true transcription factor binding sites. In this report, I have developed a new method to distill the true transcription factor binding sites. This method first filters out peaks on each strand that cannot be paired with a peak on the opposite strand, because a true binding site must be represented on both strands. Next, the analysis removes peaks that are not reproducible between replicate ChIP-seq experiments. Then, binding sites that do not have an associate cognate motif are removed. Lastly, a precise base within the motif is calculated to be the representative coordinate of the whole motif. The algorithm I developed was assembled using tools in the genomic toolkit, Galaxy, and coding using Python. The algorithm has been uploaded to a separate version of Galaxy (http://dancluster.g2.bx.psu.edu) and is ready to be used by any researcher doing ChIP-seq.

3.2 Introduction

To investigate gene regulation, researchers map the location, in the genome, where transcription factors bind. Mapping of transcription factors is performed using the Chromatin Immunoprecipitation (ChIP) assay [69,70]. When a transcription factor binds
to DNA, formadyde can be used to secure its position. Next the cells are disrupted and the DNA is sheared into fragments between 200-1000 bp. Following fragmentation, an antibody against the specific factor (or against a synthetic tag attached to the protein) is used to immunoprecipitate DNA crosslinked to the protein of interest. After DNA is isolated, two methods exist for further experimentation: ChIP-chip and ChIP-seq [4,68].

In ChIP-chip [74,75], DNA segments are incorporated with a fluorescent label and hybridized to oligonucleotide probes, which are designed against specific regions in the genome. If fluorescent DNA fragments bind the probes, the fluorescent DNA fragments emit their color upon exposure to light. The intensity of fluorescent colors yields data on the occupancy of factors in the region of that probe [4]. The length and genomic coverage of the probes determine how much information a ChIP experiment yields. If a binding site is not in the vicinity of a probe region, then the binding site will be overlooked, which is a disadvantage of ChIP-chip and was resolved by changing how ChIP fragments are assessed [84].

To improve the way ChIP material is handled, instead of hybridizing to probes, a new method was developed where the ChIP material is passed through massive parallel sequencing [123,156,157,158,159], known as ChIP-sequencing (abbreviated as ChIP-seq). Following sequencing, the DNA fragments (now known as “reads” or “tags”) are aligned to any chromosomal coordinate in the genome of the organism. Because reads are aligned along a continuum in the genome, it has advantages over ChIP-chip (in addition to improved signal-to-noise ratio), which is restricted to genomic regions that are arrayed. The alignment along the genomic continuum generates regions of enrichment,
which may be used to derive peak calls [77]. These peaks calls give explicit genomic coordinates of putative factor binding sites.

Despite the improvements of ChIP-seq over ChIP-chip, spurious peaks may still occur, and bioinformatics analysis is required to identify true binding locations. Initial applications ChIP-seq used reads density in genomic windows to distinguish signal from noise [122,123]. Read density, alone, fails to account for the distribution of reads between strands (i.e., reads may have mapped largely to the W-strand, largely to the C-strand, or evenly between both). Therefore, ChIP-seq analyses must ensure any binding site is represented by peaks on both strands [124]. Peak representation on each strand alone, though, carries other weaknesses.

Even if represented by peaks on both strands, a true binding site must also be reproducible between ChIP-seq experiments. Previously, ChIP-seq reads have been pooled to further increase signal [126]. In another report, peaks from different (biological) replicates have been merged, if overlap between peaks is found [127]. Additionally, peaks from the same (technical) replicate have been merged, if the peaks were within vicinity of each other [128,160].

Given reproducible binding site identifications, previous reports used what they deemed true binding sites to discover novel motifs for the transcription factor under investigation [129]. This methodology uses motifs as a post-processing step, meaning that deemed true binding sites are used to identify motifs. Recently, it has been proposed to change this methodology. Instead, it has been suggested to use motifs as a criterion for true binding site identification, given each transcription factor binds to family of
sequences [132]. As a result, putative binding sites are screened for motifs before being declared a true binding site [160].

ChIP-seq software packages may neglect some of the aforementioned criteria of true ChIP-seq peaks (see [125] for a list of available packages). I sought to create a protocol for binding site identification that employs the characteristics of true binding sites mentioned above. This method involves four steps. First, it matched peaks on opposite strands uniquely; that is, for every peak on one strand, the algorithm attempts to find a partner peak for it on the opposite strand. Second, the protocol compares binding sites between replicate peak calls and removes peaks that are not reproducible between replicates. Third, all peaks are matched with a cognate motif; any peak that is not within vicinity of a motif is removed. Lastly, the protocol ensures that the motif paired with a peak is, at least partially, within the boundaries of the peak. This step further calculates a coordinate within the motif to serve as a reference coordinate for the motif.

The ChIP-seq pipeline was constructed using modules in Galaxy [161] and scripts using Python programming. In the genomic toolkit Galaxy, users select specific modules to perform specific tasks. However, a user can use the output of one module to pass as input into another module. Therefore, multiple functions can be set to work in series or parallel to perform a specific analysis. This is known as a “workflow” in Galaxy (Figure 19). Whenever possible, the steps of the protocol were constructed using workflows. When it was not possible to execute a step using only Galaxy functions, I used Python programming.
Figure 19- Shown is an example of a Galaxy workflow.
Shown is an example of functions being linked together in Galaxy to create a workflow.

3.3 Results

3.3.1 Initial Analysis

Before the four-step analysis is performed, raw sequencing data must be converted to a format to initiate this four-step analysis. Therefore, a module was created (using strictly Galaxy tools) to take raw sequencing reads and align them to the genome using the available software package Bowtie [80], incorporated onto Galaxy. The output
of this (or any) alignment program is a General Feature Format (GFF) file, indicating chromosomal coordinates to which each DNA read aligned (if the read was aligned at all). This aligned reads are passed to GeneTrack (incorporated onto Galaxy) [77] to make peak calls on both DNA strands. These peak calls are used to initiate the four-step analysis.

While peak calls may be generated starting from raw sequencing data using the previously mentioned module, users may wish to align their sequencing data employing software packages other than Bowtie and/or make peak calls using an algorithm other than GeneTrack. I, therefore, developed a Galaxy workflow to input BED files, if users wish to align their ChIP-seq reads with an alternate software package (e.g. SOAP [79] or SHRiMP [78]). Alternately, to accommodate users making peak calls using algorithms other than GeneTrack, I created an additional workflow to input general peak calls. Therefore, whether or not a user chooses to use Bowtie and/or GeneTrack, a module exists to accommodate the user’s preferences.

The possibility exists that for users performing ChIP-seq on an organism whose genome is large (like human), uploading to Galaxy a large GFF file may be time-consuming. To address this, I devised a workflow to input read index files and peak call files (in the format of GeneTrack). These files are smaller than the GFF file for a large genome and are easier to upload to Galaxy.

Before moving onto the four-step analysis, it is important to note various methods of indexing reads after alignment to the genome. A ChIP-seq fragment may be 200-1000 bp long and only the first 25-35 bp are sequenced. A commonly preferred method is to use the alignment position of the 5’-end and estimated length of the read to determine the
binding site location [162]. Another algorithm utilizes the 5’-ends of the reads on both DNA strands as the position of the read [124]. For reasons that will become clear later, I chose the latter method.

Overall, the previously mentioned workflows generate peak calls used to initiate the four-step analysis described in the following sections. They differ in the various starting points to generate peak calls, but all end at the same point.

### 3.3.2 The First Criterion: Peak Matching

After peak calls are generated, the first step in the algorithm to decide the true transcription factor binding sites is, for every peak on the W- and C-strand, find a partner peak on C- and W-strand, respectively. If a peak on either strand cannot be partnered, the particular peak is discarded.

In Figure 20, assuming the top strand in the figure aligns to the W-strand and the bottom strand in the figure aligns to the C-strand, the reference position of each strand is assumed to be the chromosomal coordinate of each 5’-end. Therefore, the 5’-position of the top strand will be upstream of the 5’-position of the bottom strand (Figure 21) The downstream distance of the C-strand read (or peak) from the W-strand read (or peak) is known as shift [124].

![Figure 20 - Transcription Factor-DNA complex following sonication.](image)

After sonication, a transcription factor is bound to DNA 200-1000 bp in length. After sequencing, each strand’s reference position will be its 5’-end.
Figure 21 - GeneTrack screenshot of ChIP-seq of Reb1.
The red peak on the C-strand is downstream of its partner peak in blue on the W-strand. This finite distance between the two peaks is known as shift.

Given the C-peak’s downstream distance from the W-peak, a possible estimate of the true transcription factor binding site is the midpoint between the two peaks [124]. The estimate of a factor’s binding site depends on correctly matching a W-peak with a C-peak. Because peaks may occur along a continuum of DNA, the shifts between C-peaks and W-peaks will be variable, including the inexplicable occurrence of a C-peak positioned upstream of a W-peak. Therefore, the following section describes a two-workflow algorithm I wrote for proper peak-matching using strictly Galaxy tools.

The first workflow I wrote takes the peaks calls on each strand and creates “putative” (ie., initial) peak pairs employing the following logic: for every peak on the W-strand, the algorithm locates C-peaks that overlap the W-peak by at least one bp (of which up to two may occur) and the closest non-overlapping C-peak upstream and closest non-overlapping C-peak downstream of the W-peak (regardless of distance to the W-peak) (Figure 22).
Figure 22- Peak Matching Scenario.
In this hypothetical scenario, the peak on the W-strand, labeled “W,” can be potentially paired with four C-strand peaks: C₁, C₂, C₃, and C₄. C₁ is the closest non-overlapping peak upstream of W, C₄ is the closest non-overlapping peak downstream of W, C₂ and C₃ are overlapping peaks with W. Note this is a “worst-case scenario”; hence not all peaks will have four potential partners.

As shown in Figure 22, a peak on the W-strand may find up to four potential partners on the opposite strand. All putative peak pairs are outputted by the first workflow. The shifts, which are allowed be positive or negative, are also calculated. These shifts are important for the second workflow in the peak-matching steps.

The second workflow in the peak-matching step takes advantage of the distribution of putative peak shifts. The shifts may be unimodally, biomodally, trimodally, etc. distributed. Additionally, the shifts may be randomly distributed. Therefore, the user must use the histogram-producing tool in Galaxy (or another program of the user’s choice) to plot the shifts and decide what distribution is being shown.
Before plotting the shifts, if the dataset is large, the possibility exists that the initial peak-pairings largely comprise noise from low read-count peak-pairs. Therefore, I wrote a workflow to calculate the median read count of the putative peak pairs and display only the highest read-count peak pairs. After removing low read-count peak-pairs, plotting the shifts gives clearer view of the modality of the distribution of shifts (Figure 23).

![Figure 23](image)

**Figure 23**- Using Galaxy to make a histogram of the frequency distribution of the shifts of all putative peaks pairs for the human transcription factor CTCF. A. Shown is the distribution of shifts in the -200 bp to 200 bp range without filtering out low read count peak-pairs. B, Shown is the distribution of shifts after filtering out low read count peak-pairs.

For the human transcription factor CTCF, shown in **Figure 23**, the distribution of shifts is bimodal, with the modes slightly greater than zero and approximately 50 bp, respectively.
While CTCF’s shift distribution was bimodal, other distributions for shift are possible. The shifts may be unimodal, trimodal, etc., or uniform. For each distribution, the user must choose a corresponding workflow to finalize peak-pair matches. For this purpose, I wrote three workflows: one for the unimodal distribution, another for the multimodal distribution, and another for the uniform distribution.

Overall, the purpose of these workflows is to achieve the following: that a peak on one strand is paired with a peak on the opposite strand (up to a user-defined distance), and to ensure all peaks are paired at most once; the differences between these workflows is the algorithm of final peak-pairing.

If the user decides that the distribution of shifts is unimodal, the corresponding workflow for the unimodal distribution should be run. The workflow first requires the user to enter the mode and how far upstream/downstream a partner C-peak may be from a W-peak. If a W-strand peak is paired with more than one C-strand peak, the W-strand peak in conflict will be finalized with the partner such that the distance between the C-peak and W-peak (accounting for whether the C-peak is upstream/downstream of the W-peak) is closest to the mode.

For example, if a peak on the W-strand is paired with a C-strand peak that is 10 bp downstream (shift = +10 bp) and another peak 15 bp upstream (shift = -15 bp), while the mode is +27 bp, the algorithm will choose the former peak, because +10 bp is closer to +27 bp than is -15 bp.

The distribution may also be multimodal, as previously exemplified. After the modes are identified, the next step is to run the workflow corresponding to multimodal distributions. Like in the workflow for unimodal distributions, the user enters the
locations of the modes and how far upstream/downstream a C-peak may be from a W-peak.

When the workflow for multimodal distributions is executed, if a peak is paired multiple times, the workflow chooses the partner peak to either mode. In the case that a peak is partnered with two peaks that are equidistant from both modes, the workflow selects the partner peak that is closest to the peak in conflict in absolute distance. This workflow should be utilized if the distribution of shifts is bimodal or trimodal. If the distribution has more than three modes, the user must execute the next workflow instead.

The third option for settling multiple peak partners is to run a workflow assuming the distribution of peak shifts is uniform. In this workflow, the algorithm selects, for each peak (be it a W-peak or a C-peak), the closest peak in absolute distance to it, regardless of whether the partner is upstream or downstream. If a peak is precisely in the middle between an upstream and a downstream peak, the algorithm will choose the partner peak such that the C-peak is downstream of the W-peak.

A user can rerun the steps mentioned above for each replicate ChIP-seq experiment to get final peak pairs for each replicate. The output of the previously mentioned workflows displays a W-strand peak matched with a C-strand peak, along with the shift, and the midpoint between the peak pair. This midpoint is calculated as the average of the midpoints of both peaks. This midpoint, henceforth, will be the reference point of the peak pair. The matched peak output files are used in the next step in ChIP-seq analysis.
3.3.3 The Second Criterion: Combining Replicates

In ChIP-seq analysis, a true binding site should be reproducible across multiple replicates. Therefore, a Python code was written for this section of the analysis to determine the location of peaks across all replicate matched-peaks files; if peaks from different replicates are within vicinity of each other, the program consolidates the peaks and they will be considered one binding site, whose start and end is the median of the start and end of the replicate peaks, respectively. Other algorithms for consolidation have been different for specific cases [127,128,160], but the algorithm presented in this section takes into account general scenarios.

This replicate-matching tool requires user-defined settings. The user should input all matched-peaks files from the previous criterion (each file as a separate replicate). Next, the user specifies the maximum distance replicate peaks are allowed to be for consolidation into one group. As guidance, a recommended distance is half the peak binding site size. Additionally, the minimum number of peaks across the replicates that constitutes a true binding site should be specified (i.e., how many replicate peaks form a binding site).

To initiate replicate consolidation, the program reads in all the peak information from all matched-peaks files. Afterward, the peaks are compiled into a map on each chromosome. Then, the program, for each chromosome, scans from 5’ to 3’ the map of all the peaks. The program searches for an additional peak in the vicinity of a peak from another replicate. Figure 24 illustrates various cases of peak mapping and consolidation.
Figure 24- Second criterion scenario
Shown are various scenarios of combining replicates. A DNA strand is shown and each colored bar represents a peak; each replicate is represented by one symbol. A, The three bars (peaks) on the left represent peaks within the user-defined distance of each other and may be consolidated. A lone peak is shown on the right. B, Shown are five different peaks all within the user-defined radius of each other. This scenario highlights the possibility of two peaks from the same replicate file (the squares) that conflict for being grouped with other peaks (the diamond- and star-peak). Described in the text is how the program settles the various cases of peak grouping conflicts.

**Figure 24A** displays one case of peak grouping. The program scans a chromosome 5’ to 3’. It then finds the peak symbolized by the circle; then, it finds the peak symbolized by the square and determines it is within the user-defined radius of peak grouping (ie., the distance between the two peaks is less than the binding site size). Next, the program finds the diamond-peak and also determines it is within the user-defined radius of the square-peak and groups the diamond-peak along with the circle- and square-peaks. The diamond-peak may be outside the range of the user-defined radius of the circle-peak, but must be within radius of the closest neighbor (the square-peak) in order to be grouped together. Next, the program finds that the sequential peak, the star-peak, is not within radius of the diamond-peak. Hence, the program will consider the circle-, square-, and diamond-peaks one group. According to **Figure 24A**, there is no other peak
near the star-peak, therefore the star-peak remains as a single replicate. The scenario in **Figure 24A** highlights a simple case.

**Figure 24B** considers an alternate scenario. In the figure, five peaks from four replicate files are displayed. Each peak is within the user-defined radius of its nearest neighbor. Any group of peaks must have at most one peak from each replicate. In this group of five, two square-peaks are in conflict in being grouped with the diamond- and star-peak. The program assigns the diamond- and the star-peaks to the square-peak having the higher read count. In **Figure 24B**, if the square-peak on the left has higher read count, the first four peaks on the left form a group and the square-peak on the right is left unconsolidated. Alternatively, if the square-peak on the right has higher read count, the circle-peak is grouped with the square-peak on the left, and the diamond- and star-peaks are grouped with the square-peak on the right.

After replicate peaks are grouped, the Replicates program produces two output files. The first displays which peaks from which replicates were grouped together. It lists the chromosome, base position, read count, and shift each peak has, along with the standard deviation of the read distribution of each peak. The second output file, used in the third criterion, displays the chromosome, median start and end coordinates, median shift, median read count, the number of replicates in each group, and the combined standard deviations of the replicate peaks. Note that when the second output file is generated, an extra motif search distance is used (specified by the user when the Replicates program is executed) to generate sequence data for the motif analysis in the next section (motif-generating software packages require a distances larger than a typical transcription factor binding site). Therefore, the reported binding site in the second
output file shows the (peak start - motif search distance) and (peak end + motif search distance).

Lastly, a normalized read count for each peak from each replicate is calculated. Normalization is performed to account for differences in sequencing depth between each replicate. Therefore, each peak’s read count (the sum of its W-strand and C-strand read counts) was normalized as follows:

\[
\text{Norm}_\text{Val} = (\text{Read}_\text{Count}) \times \frac{\text{Median}_\text{Read}_\text{Count}_\text{of}_\text{all}_\text{Peak}_\text{Pairs}}{\text{Median}_\text{Read}_\text{Count}_\text{of}_\text{Peaks}_\text{In}_\text{Replicate}_\text{File}}
\]

This formula normalizes each peak to the median read count of all peaks in all files and the median read count of all peaks in the respective peak’s replicate file.

### 3.3.4 The Third Criterion: Sequence Analysis

Following execution of the first two criteria, only peaks that are represented on both strands and are reproducible between replicates remain. The following level of filtering involves sequence analysis.

Transcription factors bind cognate motifs (except general transcription factors, to be explained later); therefore a true binding site must have a recognizable motif. Therefore, this step in the overall ChIP-seq analysis matches each binding site with a cognate motif. If a nearby motif cannot be found, the peak does not pass this step. For this protocol, the Galaxy team implemented the motif-determining software package MEME [131] and the motif-locating software package FIMO [130] onto Galaxy. I created a workflow to interpret FIMO results and match motif locations to those of the transcription factor binding sites (produced from the second criterion). MEME and
FIMO are part of a group of motif-related software packages called the MEME Suite and can all be found at http://meme.nbcr.net/meme4_3_0/intro.html [130].

To initiate this analysis, the user must utilize the second output file from the second criterion (the file with the consolidated peaks) and find sequences near each peak. These sequences are uploaded into MEME, and a motif is generated. If the number of characters in the extracted sequences exceeds MEME’s limit, the user can utilize MEME by uploading only the peaks with the highest read count (and continue the rest of the analysis using all the peaks, given the motif generated by the high read count peaks).

To exemplify the flow of this motif-finding criterion, I use the sequence-specific factor Abf1. After the first two steps in the ChIP-seq analysis were run, I employed MEME on Galaxy to produce Abf1’s motif (consistent with previous findings [5,62,163]), shown in Figure 25.

![Figure 25- MEME analysis for the motif of Abf1.](image)

After a motif is generated, as the example displayed in Figure 25, the following step executes FIMO in Galaxy using the motif from MEME as input. The purpose of FIMO is to display genomic locations of the chosen motif.
FIMO accepts two user inputs for motif mapping. A user may display motif locations in the vicinity of binding sites or the user may display motif locations in the whole genome. The simpler choice is if the user inputs the former for motif mapping. However, as mentioned earlier, a user might have sequences for only a sample of the replicates. Because only a portion of the sequences might be entered into MEME, I designed this third criterion analysis to take, as FIMO input, the locations throughout the whole genome of the motif.

Taking as input the FIMO-generated motif locations and the transcription factor binding sites, I developed a workflow to measure the distance between the motif locations and the transcription factor binding sites. The workflow looks for instances where the motif interval and the binding location intervals (including the extra distance from the second criterion) overlap. If there is at least 1-bp overlap, the motif and binding site are paired. After the motifs are paired with their proper binding sites, three scenarios of motif placement relative to the factor binding site are possible (Figure 26).
Figure 26- Motif placement relative to a possible binding site.
Shown are three possible scenarios of a motif (green bar) location relative to a putative binding site (enclosed within red lines). A, The motif is fully within the binding site. B, The motif is partially within the binding site. C, The motif is outside the binding site, but might be within the extra distance utilized for sequence analysis.

As shown in Figure 26, three possible situations are possible of motif placement relative to a binding site: fully within the binding site, partially within the binding site, or fully
outside the binding site, but within the extra distance used for the generation of a motif. Note that the latter scenario will be filtered out in Criterion 4 (to be explained later).

Despite the importance of ensuring any binding site be paired with a motif, it is possible that the user may perform ChIP-seq on a factor that lacks an overt motif. An example of this kind of factor is a general transcription factor. Therefore, I wrote a workflow in Galaxy that allows a user to upload TSS coordinates of all genes; the workflow uses the replicates output from the second criterion and filters out peaks based on a user-defined distance to the nearest TSS. If a user uses TSS coordinates for the third criterion, the next criterion may be neglected.

### 3.3.5 The Fourth Criterion: Motif Distance

After motifs have been paired with candidate binding sites, the next step is to calculate the distance from the binding site center to the motif. This involves choosing a specific base inside the motif to be the Motif Reference Coordinate (MRC), meaning the specific base serves as the “representative” coordinate of the whole motif. The MRC is defined as the lowest average distance from the coordinate to the binding location midpoint. Therefore, I wrote a Python script to calculate the optimal base within the motif, such that the average distance from the motif to the binding site the center is the least (Figure 27).
Figure 27- Fourth criterion analysis. Two examples are highlighted of motif placement relative to a binding site. The red lines indicate the binding site interval of the factor. The blue line indicates the center of the binding site. A discussion of the analysis for these cases is in the text. The motif shown in the figure is the Reb1 motif. The scenarios in A and B differ where the motif appears relative to the binding site center.

First, the program ensures the motif is, at least partially, within the confines of the binding site. If the motif is fully outside of it, the binding site is filtered out. Then, the program examines each base position within the motif and directly calculates the distance to the binding site center. For example, in Figure 27, the program starts with the 5’-end base of the motif: the first T. In Figure 27A, the distance from that T to the motif center
is 5 bp; in Figure 27B, the distance from the 5’-end T to the binding site center is 10 bp. The program performs this calculation for this T (or mismatch of it) for every binding location in the genome that has passed through the first three criteria and will calculate the average distance from this T to the binding site center.

Then, the program repeats this cycle using the next base, the second T; then the next base, the A; etc. The program takes the base position within the motif that yields the lowest average distance to the binding site center. The particular base that is chosen will become the MRC and the program will calculate the absolute distance from the binding site center to the MRC, which is not necessarily the center of the motif.

3.3.6 The Fifth Criterion: Read Count

The purpose of the fifth and final criterion is to remove any final peaks that have a low read count. In this part of the analysis, the median read count of each peak group is plotted in a histogram, the mode of the read counts is calculated, and any peak with a read count less than 50% of the mode is removed. Calculating the mode is not easily automatable, because a user must judge how wide to make the binning in the histogram before deciding what the mode read count is. Further analysis may include models to automate the proper binning of the read counts.

3.4 Discussion

The current methods for detecting transcription factor binding are through the use of ChIP-chip and ChIP-seq. ChIP-seq is currently overtaking ChIP-chip, because of its
higher resolution, great coverage, and cost effectiveness. However, false positives still occur in ChIP-seq [164].

Because of the possibility of false positives, experimenters must perform analysis to distill the true transcription factor binding sites from ChIP-seq data. Previous methods for separating true peaks from spurious peaks have utilized read density, peak representation on both DNA strands, reproducibility between ChIP-seq replicates, and motif pairing. However, I am not aware of any previously existing ChIP-seq analysis software package or algorithm that accounts for all of the aforementioned characteristics of true peaks. Therefore, in this report, I created an algorithm to glean the true ChIP-seq binding sites using available tools in Galaxy and writing Python scripts [161]. This algorithm is now available on a separate version of Galaxy, not the main version, at http://dancluster.g2.bx.psu.edu for other research groups to use.

In each step in the algorithm, I introduced flexibility to account for different scenarios users might be challenged with. Users may perform their initial analysis (before starting the five criteria for peak filtering) using their method of choice: following sequencing, users can either initiate the workflows I developed, by aligning their ChIP-seq data to the genome using Bowtie [80] and making peak calls using GeneTrack [77]. User may also use another alignment package, like SHRiMP [78] or SOAP [79] before starting their analysis on Galaxy. They may make peaks calls using a package other than GeneTrack. Therefore, any combination of software packages for alignment and peak calling can be handled by the pipeline developed.

Following alignments of ChIP-seq reads and peak calling, the algorithm is highly selective for true binding site peaks. For peak pairing, the algorithm is designed, for each
peak on the W- or C-strand, to select the proper peak on the C- or W-strand, respectively. Therefore, any true binding site is represented on both DNA strands. Following pairing peaks, the algorithm merges peaks from replicate ChIP-seq experiments. For each peak in each replicate, the algorithm considers the locations of nearby peaks from other replicates and decides final peak groupings. Therefore, all binding sites will be reproducible. After reproducibility is checked, a cognate motif for each peak is developed. For this purpose, the previously developed software packages MEME and FIMO are utilized to generate a motif and finds all occurrence of it in the genome, respectively [130]. Finally, the algorithm calculates the proper base position inside each binding site’s cognate motif to serve as a Motif Reference Coordinate (MRC).

Despite the utility and flexibility of the algorithm, the main drawback to the algorithm is that some steps require quantitative information as user input, which may be difficult to estimate objectively. For example, the user sets $\sigma$ (standard deviation) and D (exclusion zone) in GeneTrack at the beginning of the analysis. While peak calls are largely invariant of these parameters, they are not absolutely independent. Extreme values of $\sigma$ and D will merge independent peak calls. In addition, noisy or closely spaced peaks will be more sensitive to the values of $\sigma$ and D. Computational optimization of these parameters in an automated and statistically grounded fashion represents areas of future improvements.

Users must estimate how far upstream or downstream partner C-peaks may be from partner W-peaks when executing the first criterion, in addition to deciding what the distribution of shifts is. In the second criterion, the user is required to input how many replicates are needed to corroborate a true binding site. Additionally in the second
criterion, how close replicate peaks need to be in order to be merged is required input. A further enhancement to this report would be objective modeling and calculations for these subjective input parameters.

Even though this algorithm requires some quantitative user input, which is not easily objectively calculated, this newly developed algorithm is important for a few reasons. It integrates several previously proposed criteria for the identification of true transcription factor binding sites: peak matching on both strands, replicates, and motif matching. As far as I am aware, this is the first study to account for all three criteria. Also, this algorithm is freely available on an alternate version of the genomic toolkit, Galaxy; this website is commonly used for analysis of mass datasets. Therefore, the algorithm does not require downloading and installation and may be run remotely.
Chapter 4- General Significance and Future Directions

4.1 Modeling PIC Disassembly using ChIP-chip data

PIC assembly in vitro has been defined using isolated components [7,138]. A question that has puzzled biochemists is how much in vitro data on PIC assembly represent PIC assembly in vivo. Additionally, little is known about PIC disassembly in vitro or in vivo. Overall, we attempted to forge a connection between in vitro data on PIC assembly and in vivo data on transcription factor occupancy.

My work in Chapter 2 with the writing of the Python program PathCom, by tying together in vitro PIC assembly data and in vivo occupancy data, has elucidated some clues to the PIC disassembly process. PathCom requires some assumptions about the behavior of molecules in vivo. Therefore, the program is not intended to be declarative, but rather a means of extracting patterns from ChIP data. The extent to which a user confides in the results of PathCom relies on the extent the user accepts the assumptions PathCom requires.

To narrow down plausible PIC disassembly mechanisms, two rules are used based on which proteins co-occupy the promoter together. The first rule stipulates that protein A’s occupancy cannot be greater than that of another protein, protein B, that always co-occupies the promoter with protein A. The second rule says, given that protein B associates after protein A and before protein C, and protein B’s occupancy is greater than the combined occupancies of A and C, then B must be the last of the three to disassemble.
Using these rules of compatibility, we initially exemplified this modeling process using the assembly of general transcription factors. We found that at 25°C, TBP and pol II leave the PIC early, while TFIIB, TFIIF, TFIIE, and TFIIH leave the PIC late. This is in agreement with a previous result that found TBP to be dynamic [140]. TBP was found to have varying dynamics between pol I, II, and III genes, with pol II being the highest [141], which is in agreement somewhat with our results. Under stress conditions, we found similar results for PIC disassembly as at 25°C. Additionally, we observed that clusters with higher membership and higher median occupancy were compatible with a few mechanisms, despite a factorially increasing number of disassembly mechanisms every time an additional factor is added to the model. This may reflect a cellular design to maintain PIC disassembly along a defined pathway so as to achieve robust regulation.

We further modeled when sequence-specific regulators and chromatin remodelers dissociate from the PIC compared to the general transcription factors. We did PathCom analysis for fifteen sequence-specific factors and five chromatin remodelers. We found that the sequence-specific regulators and chromatin remodelers dissociate from the PIC early, along with TBP and pol II. This reflects a previous result that shows GTF’s having a higher fold-enrichment over background than sequence-specific regulators and chromatin remodelers [8].

The sequence-specific regulators, chromatin remodelers, and general transcription factors analyzed by PathCom are just a small portion of the proteins involved in eukaryotic gene regulation. Because of limits of computational power, we could only analyze six factors at a time using PathCom.
Therefore, future analysis with greater computational could include a higher number of transcription factors than six (six was the highest number of factors analyzed in this thesis). Further analyses might also include PathCom analysis of sequence-specific factors and chromatin remodelers that bind to the same group of genes. Also, researchers might try experimental studies to verify (or refute) the modeling predictions proposed in Chapter 2 of this thesis. Furthermore, a more complete model of gene regulation might include negative regulators like NC2 and Mot1.

4.2 Incorporation of ChIP-seq Analysis Pipeline to Galaxy

The advent of ChIP-seq has revolutionized the field of genomics [158]. Utilizing ChIP-seq, an experimenter possesses the ability to map a transcription factor (or a nucleosome) to any chromosomal coordinate in the genome. Following sequencing of ChIP fragments and mapping back to the genome, analysis of the mapped regions must be performed to separate the true binding site from the spurious peaks.

Separation of true from false binding sites requires the utilization of characteristics of true binding sites. Previous analyses and software packages used read density (above background), peak representation on both DNA strands, and reproducibility of peaks [123,124,126,127,129]. Later, the use of motifs for as another method to filter out peaks was employed [132]. In this thesis, I took these characteristics of true binding sites and, with the use of Galaxy [161] and Python programming, automated them into a ChIP-seq pipeline, which is available to other research groups for ChIP-seq analysis. Overall, this algorithm should help reduce the False Discovery Rate (FDR) of ChIP-seq peaks.
This algorithm has future applications. With a reduced FDR, a possible future prospect is to improve on the work performed in [76] and elucidate higher resolution and lower false discovery of the promoter organization of transcription machinery. Furthermore, if other important characteristics of true binding sites are discovered, these new features must be updated into the pipeline. This algorithm could be modified for the mapping of nucleosomes as well, in addition to transcription factors.
Appendix A- PathCom analysis for different association orders
Shown are different association pathways on the three-factor model occupancy data using TBP, TFIIB, and pol II. Each different association pathway will cause different dissociation pathways to become compatible with the observed occupancy data.
Appendix B- TBP Footprints on UCSC Genome Browser

All genomic coordinates are using the *S. cerevisiae* Oct. 2003 genome

Appendix B- TBP footprint on UCSC browser

Shown are screenshots retrieved from the UCSC Browser of “digital footprints” of different factors [151] at the most highly TBP-occupied genes with a TATA-box. The digital footprints over the TATA-box show little difference from that of the surrounding areas.
Appendix C- Sequence-Specific Factors and Chromatin Remodelers Analyzed

This appendix shows a list of the 15 different sequence-specific factors and chromatin remodelers analyzed using PathCom.

<table>
<thead>
<tr>
<th>Sequence-specific factors</th>
<th>Chromatin remodelers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ash1</td>
<td>Isw1</td>
</tr>
<tr>
<td>Bas1</td>
<td>Isw2</td>
</tr>
<tr>
<td>Cha4</td>
<td>RSC (Rsc9)</td>
</tr>
<tr>
<td>Cin5</td>
<td>SWI-SNF (Swi3)</td>
</tr>
<tr>
<td>Ifh1</td>
<td>INO80</td>
</tr>
<tr>
<td>Rap1</td>
<td></td>
</tr>
<tr>
<td>Reb1</td>
<td></td>
</tr>
<tr>
<td>Rfx1</td>
<td></td>
</tr>
<tr>
<td>Rph1</td>
<td></td>
</tr>
<tr>
<td>Sfp1</td>
<td></td>
</tr>
<tr>
<td>Swi4</td>
<td></td>
</tr>
<tr>
<td>Xbp1</td>
<td></td>
</tr>
<tr>
<td>Yap6</td>
<td></td>
</tr>
<tr>
<td>Zap1</td>
<td></td>
</tr>
<tr>
<td>Zms1</td>
<td></td>
</tr>
</tbody>
</table>
### Appendix D - Different Dissociation Orders When a Sequence-Specific Factor and General Transcription Factors Are Analyzed

<table>
<thead>
<tr>
<th>Mechanism Number</th>
<th>Dissociation Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>STPG</td>
</tr>
<tr>
<td>2</td>
<td>STGP</td>
</tr>
<tr>
<td>3</td>
<td>SPTG</td>
</tr>
<tr>
<td>4</td>
<td>SPGT</td>
</tr>
<tr>
<td>5</td>
<td>SGTP</td>
</tr>
<tr>
<td>6</td>
<td>SGPT</td>
</tr>
<tr>
<td>7</td>
<td>TSPG</td>
</tr>
<tr>
<td>8</td>
<td>TSGP</td>
</tr>
<tr>
<td>9</td>
<td>TPSG</td>
</tr>
<tr>
<td>10</td>
<td>TPSS</td>
</tr>
<tr>
<td>11</td>
<td>TGSP</td>
</tr>
<tr>
<td>12</td>
<td>TGPS</td>
</tr>
<tr>
<td>13</td>
<td>PSTG</td>
</tr>
<tr>
<td>14</td>
<td>PSGT</td>
</tr>
<tr>
<td>15</td>
<td>PTSG</td>
</tr>
<tr>
<td>16</td>
<td>PTGS</td>
</tr>
<tr>
<td>17</td>
<td>PGST</td>
</tr>
<tr>
<td>18</td>
<td>PGTS</td>
</tr>
<tr>
<td>19</td>
<td>GSTP</td>
</tr>
<tr>
<td>20</td>
<td>GSPT</td>
</tr>
<tr>
<td>21</td>
<td>GTSP</td>
</tr>
<tr>
<td>22</td>
<td>GTPS</td>
</tr>
<tr>
<td>23</td>
<td>GPST</td>
</tr>
<tr>
<td>24</td>
<td>GPTS</td>
</tr>
</tbody>
</table>

Shown are the 24 different dissociation pathways when a sequence-specific factor, TBP, pol II, and the TFII-factors (median of TFIIB, TFIIF, TFIIE, and TFIH occupancy) were analyzed. In this list “S” stands for sequence-specific factor, “T” stands for TBP, “P” stands for pol II, and “G” stands for TFII. The order of association is S → T → P → G.
## Appendix E - The Dissociation Mechanisms With Four General Transcription Factors

<table>
<thead>
<tr>
<th>Mechanism Number</th>
<th>Dissociation Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TBPH</td>
</tr>
<tr>
<td>2</td>
<td>TBHP</td>
</tr>
<tr>
<td>3</td>
<td>TPBH</td>
</tr>
<tr>
<td>4</td>
<td>TPHB</td>
</tr>
<tr>
<td>5</td>
<td>THBP</td>
</tr>
<tr>
<td>6</td>
<td>THPB</td>
</tr>
<tr>
<td>7</td>
<td>BTPH</td>
</tr>
<tr>
<td>8</td>
<td>BTHP</td>
</tr>
<tr>
<td>9</td>
<td>BPTH</td>
</tr>
<tr>
<td>10</td>
<td>BPHT</td>
</tr>
<tr>
<td>11</td>
<td>BHTP</td>
</tr>
<tr>
<td>12</td>
<td>BHPT</td>
</tr>
<tr>
<td>13</td>
<td>PTBH</td>
</tr>
<tr>
<td>14</td>
<td>PTHB</td>
</tr>
<tr>
<td>15</td>
<td>PBTH</td>
</tr>
<tr>
<td>16</td>
<td>PBHT</td>
</tr>
<tr>
<td>17</td>
<td>PHTB</td>
</tr>
<tr>
<td>18</td>
<td>PHBT</td>
</tr>
<tr>
<td>19</td>
<td>HTPB</td>
</tr>
<tr>
<td>20</td>
<td>HTPB</td>
</tr>
<tr>
<td>21</td>
<td>HBTP</td>
</tr>
<tr>
<td>22</td>
<td>HBPT</td>
</tr>
<tr>
<td>23</td>
<td>HPTB</td>
</tr>
<tr>
<td>24</td>
<td>HPBT</td>
</tr>
</tbody>
</table>

Shown are the 24 different dissociation pathways using the general transcription factors TBP, TFIIB, pol II, and TFIIH. Here, “T” stands for TBP, “B” stands for TFIIB, “P” stands for pol II, and “H” stands for TFIIH. The order of association is T → B → P → H.
Appendix F- Principal Component Analysis of Six General Transcription Factors

In this appendix is a Principal Component Analysis (PCA) of the six GTF’s using Cluster [108].

<table>
<thead>
<tr>
<th>EIGVALUE</th>
<th>NAME</th>
<th>GWEIGHT</th>
<th>TBP</th>
<th>TFIIB</th>
<th>TFIIE</th>
<th>Pol</th>
<th>TFIIF</th>
<th>TFIIH</th>
</tr>
</thead>
<tbody>
<tr>
<td>59.9595108</td>
<td>59.95951</td>
<td>1</td>
<td>-0.11106</td>
<td>-0.46208</td>
<td>-0.28596</td>
<td>-0.43615</td>
<td>-0.54131</td>
<td>-0.45731</td>
</tr>
<tr>
<td>27.69308472</td>
<td>27.69308</td>
<td>1</td>
<td>0.041293</td>
<td>-0.07782</td>
<td>0.528392</td>
<td>-0.41172</td>
<td>0.537163</td>
<td>-0.50496</td>
</tr>
<tr>
<td>24.04491425</td>
<td>24.04491</td>
<td>1</td>
<td>-0.0226</td>
<td>-0.00442</td>
<td>0.790099</td>
<td>0.027717</td>
<td>-0.58445</td>
<td>0.18127</td>
</tr>
<tr>
<td>19.90500259</td>
<td>19.905</td>
<td>1</td>
<td>0.056696</td>
<td>0.877649</td>
<td>-0.10431</td>
<td>-0.28399</td>
<td>-0.24752</td>
<td>-0.27152</td>
</tr>
<tr>
<td>12.8575449</td>
<td>12.85754</td>
<td>1</td>
<td>-0.09982</td>
<td>-0.0091</td>
<td>-0.05136</td>
<td>-0.74211</td>
<td>0.101784</td>
<td>0.652859</td>
</tr>
<tr>
<td>12.21739578</td>
<td>12.2174</td>
<td>1</td>
<td>0.986039</td>
<td>-0.10027</td>
<td>-0.03542</td>
<td>-0.09004</td>
<td>-0.07233</td>
<td>0.055496</td>
</tr>
</tbody>
</table>
Appendix G - Resetting the DNA Concentration to 1 does not Affect Compatibility

Association Order = T → B → P

<table>
<thead>
<tr>
<th>Initial Concentrations</th>
<th>Initial Concentrations</th>
</tr>
</thead>
<tbody>
<tr>
<td>T = 10</td>
<td>T = 10</td>
</tr>
<tr>
<td>B = 10</td>
<td>B = 10</td>
</tr>
<tr>
<td>P = 10</td>
<td>P = 10</td>
</tr>
<tr>
<td>D = 10</td>
<td>D = 1</td>
</tr>
</tbody>
</table>

PathCom

COPASI
Bibliography


VITA
Eric Samorodnitsky

Education:

Ph.D in Integrative Biosciences (Bioinformatics and Genomics option),
Spring 2011
The Pennsylvania State University, University Park, PA
Dissertation Title: Genome-wide modeling of transcription preinitiation complex
disassembly mechanisms using ChIP-chip data
Dissertation Advisor: B. Franklin Pugh

B.S. in Biological and Environmental Engineering, Spring 2006
Cornell University, Ithaca, NY

Publications:

Samorodnitsky E, Pugh BF (2010) Genome-wide modeling of transcription
preinitiation complex disassembly mechanisms using ChIP-chip data. PLoS
Comput Biol 6(4):e1000733

Honors and Awards:

Dean’s List, Spring 2006